Genomics and proteomics

✓ **Genomics**: Systemic study of the entire genome of an organism
  - To sequence the entire genome and to physically map the genome arrangement (assign exact position of the genes /non-coding regions in genome)
  - Before 1990s, the sequencing and study at a single gene level: Laborious and time-consuming task
  - Development of high-throughput sequencing technologies and highly automated hardware system
    - Faster (in excess of 1 kb/h), cheaper, and more accurate
    - Sequencing a human whole genome: ~ $1,000
cf) Whole eukaryotic genome: ~ $100 by Complete Genomics Inc.

✓ **Human genome project**
  - Launched in 1990 and Completed in 2003: ~ 3.2 giga bases (Gb),
    1,000 times larger than a typical bacterial genome
  - Less than 1/3 of the genome: transcribed into mRNA: Only 5 % of RNA encodes polypeptides
    - Total number of polypeptide-encoding genes: ~ 30,000
    - Total number of proteins in human body: ~ 50,000
  - Human body: to generate **2 million different types of proteins**, coded by only 20,000-25,000 genes
✓ Genome sequences of more than 2,000 organisms

- Genomes of various animals and plants: mouse, rat, sheep, pig, monkey, dog, chicken, wheat, barley, Arabidopsis

- Bacteria: Compact genome architecture: strong correlation between genome size and number of functional genes in a genome and those genes are structured into operons

Log-log plot of the total number of annotated proteins in genomes submitted to GenBank as a function of genome size.
Based on data from NCBI genome reports
Earth Bio-genome Project: BioGenomics 2017 conference

- To understand biodiversity and evolutionary process by deciphering the genomes of every species, starting with 1.5 million eukaryotes
- First step: Sequences of a member of each eukaryotic family: ~ 9,000
- Cost: $ 4.8 billion owing to decreasing costs and improving sequencing technology

A head start

The Earth BioGenome Project could coordinate the efforts below and others that are already sequencing broad swaths of the planet’s life.

<table>
<thead>
<tr>
<th>PROJECT</th>
<th>YEAR STARTED</th>
<th>SEQUENCING GOAL</th>
<th>NUMBER SEQUENCED</th>
</tr>
</thead>
<tbody>
<tr>
<td>G10K</td>
<td>2009</td>
<td>9478 vertebrate genera</td>
<td>100</td>
</tr>
<tr>
<td>i5K</td>
<td>2011</td>
<td>5000 arthropods</td>
<td>30</td>
</tr>
<tr>
<td>GIGA</td>
<td>2013</td>
<td>7000 marine invertebrates</td>
<td>60</td>
</tr>
<tr>
<td>GAGA</td>
<td>2016</td>
<td>All 300 ant genera</td>
<td>25</td>
</tr>
<tr>
<td>B10K</td>
<td>2016</td>
<td>All 10,500 bird species</td>
<td>300</td>
</tr>
<tr>
<td>AOCC</td>
<td>2013</td>
<td>101 African food crops</td>
<td>22</td>
</tr>
</tbody>
</table>

G10K: Genome 10K
Significance of genome data in Biotechnology/Biology/Medical sciences

- **Provide full sequence information of every protein:**
  - Identification of undiscovered proteins: Understanding their functions
  - Discovery of new drug targets
    → Current drugs on the market target one of at most 500 proteins:
      - Major targets are proteins (e.g., Kinases)

- **Sequence data of many human pathogens**
  - (e.g., Helicobacter pylori, Mycobacterium tuberculosis, Vibrio cholerae)
  → Provide drug targets against pathogens (e.g., gene products essential for pathogen viability or infectivity)
  → Offer some clues in underlying mechanism of diseases
  → Development of a new drug

✓ New methods/tools in Biotechnology, Biology, and Medical sciences
• The ability to interrogate the human genome has altered our approach to studying basic biology and complex diseases, and development of therapies.

• The emergence of genome-wide analysis tools has opened the door to investigating the function of each gene, genomic biomarker discovery, validation, and pharmacogenomics.

• **Medical/clinical fields:**
  - Actively studying genomic approaches to understand diseases, and learn how these can be translated into medical and clinical settings.

  ➔ **Translational research**
Issues

- Biological functions of between one-third and half of sequenced gene products remain unknown.

- Assessment of biological functions of the sequenced genes
  - Crucial to understanding the relationship between genotype and phenotype as well as direct identification of drug targets.

- Shift in the focus of genome research
  - Elucidation of biological function of genes.
• In the narrow sense: Biological function/activity of the isolated gene product

• In broader meaning:
  - Where in the cell the gene product acts, and what other cellular elements it interacts with → **Interactome**
  - How such interactions contribute to the overall physiology of the organism → **Systems Biology**

✓ General definition of functional genomics:

• Determining the function of proteins deduced from genome sequence is a central goal in the post genome era
• Elucidating the biological function of gene products
Assignment of function of gene products (Proteins)

- Biochemical (molecular) function
- Assignment based on sequence homology
- Based on structure
- Based on ligand-binding specificity
- Based on cellular process
- Based on biological process
- Based on proteomics or high-throughput functional genomics
Conventional approaches

• Clone and express a gene to produce the protein encoded by the gene

• Try to purify the protein to homogeneity
  - Size, charge, hydrophobicity, oligomeric state, glycosylation

• Develop an assay for its function

• Identify the activity / function
  - Determination of 3-D structure

✓ Time-consuming and laborious for huge numbers of genes
Assignment of function to the sequenced gene products

✓ Comparison of sequence/structure data in a high-through manner

✓ Sequence homology study

• Computer-based sequence comparison between a gene of unknown function and genes whose functions (or gene product function) have been assigned
• High homology: high similarity in function
• Assigning a putative function to 40 - 60% of all new gene sequences
Multiple sequence alignment (MSA)

- Sequence alignment of many biological sequences: DNA, protein, RNA
  - The input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor.
  - Analysis of sequence homology and phylogenetic relationship to assess the sequences' shared evolutionary origins.
  - Visual depictions of the alignment to illustrate mutation events such as point mutations (single amino acid or nucleotide changes), and insertion or deletion mutations (indels or gaps).

- Use of MSA
  - To assess sequence conservation of nucleotides, protein domains, tertiary and secondary structures, and even individual amino acids
  - To create a phylogenetic tree
  - To identify functionally important sites, such as binding sites, active sites, or sites corresponding to other key functions

- Program: Clustal W, MAFFT, T-COFFEE
✓ **Phylogenetic profiling**

- Evolution: “descent with modification”
- Phylogenetic trees: To illustrate the evolutionary history of various taxa
- Study of evolutionary relationships among various biological species or other entities based on similarities and differences in their physical and/or genetic characteristics

- A phylogenetic tree or evolutionary tree: a diagram showing the evolutionary interrelations of a group of organisms or genes or proteins derived from a common ancestral form based on similarities and differences in their physical and/or genetic characteristics

- [Https://www.youtube.com/watch?v=iyAOkzdO3vw](https://www.youtube.com/watch?v=iyAOkzdO3vw)

- Phylogenetic trees composed with a nontrivial number of input sequences are constructed using computational phylogenetics methods
  - Calculation of genetic distance from multiple sequence alignments

- **Closely related species should be expected to have very similar sets of genes**
A tree for rRNA genes, showing the three life domains Bacteria, Archaea, and Eucaryota

Each node with descendants represents the inferred most recent common ancestor of the descendants, and the edge lengths may be interpreted as time estimates.
• Proteins that function in the same cellular context have similar phylogenetic profiles: During evolution, all such functionally linked proteins tend to be either preserved or eliminated in a new species:

→ Proteins with similar profiles are likely to belong to a common group of functionally linked proteins.

• Establishing a pattern of presence or absence of a particular gene coding for a protein of unknown function across a range of different organisms whose genomes have been sequences:

• Discovery of previously unknown enzymes in metabolic pathways, transcription factors that bind to conserved regulatory sites, and explanations for roles of certain mutations in human disease and plant specific gene functions
✓ Rosetta Stone Approach

• Hypothesis: Some pairs of interacting proteins are encoded by two genes in some genome or by fused genes in other genomes

• Two separate polypeptides (X and Y) found in one organism may occur in a different organism as a single fused protein (XY)

• Function of the unknown gene in one organism can be deduced from the function of “fused genes” in different organism
Gyrase: Relieves strain while double-stranded DNA is being unwound by helicase

- Type II topoisomerase (heterodimer): catalyzes the introduction of negative supercoils in DNA in the presence of ATP.

- Gyrase (bacterial topoisomerase II): heterotetramer made up of 2 gyrA (97 kDa) subunits and 2 gyrB (90 kDa) subunits.
✓ Knock-out animal study

• Generation and study of mice in which a specific gene has been deleted
• Phenotype observation

✓ Structural genomics approach
• Resolution of 3-D structure of proteins
Pathway maps

- Linked set of biochemical reactions catalyzed by enzymes
  - Questions:
    - Is the extrapolation between species valid?
    - Have orthologs been identified accurately?

- **Homologs**: A gene related to a second gene by descent from a common ancestral DNA sequence regardless of their functions.

- **Orthologues**: Genes in different species that evolved from a common ancestral gene by speciation, retaining the same function in the course of evolution.
  - Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes.

- **Paralogs**: Genes related by duplication within a genome. Paralogs evolve new functions during a course of evolution, even if these are related to the original one.
View the other chart: Metabolic Pathways
DNA microarray technology : DNA chip

• Sequence data provide a map and possibility of assigning the putative functions of the genes in genome based on sequence comparisons

• Information regarding which genes are expressed and functionally active at any given circumstance and time

• **DNA microarray data :**
  - Provide clues as to the biological function of the corresponding genes
    → Starting point
  
  - Offer an approach to search for disease biomarkers and drug targets

ex) If a particular mRNA is only produced by a cancer cell compared to a normal cell, the mRNA (or its polypeptide product) may be a **target for basic research for cancer**, a good target for a **new anti-cancer drug, biomarker for diagnosis**.
Microarrays: Tool for gene expression profiling

- **Microarray (Gene chip):**
  - Comparison of mRNA expression levels between a sample (cancer cell) and a reference (normal cell), and SNPs in a **high-throughput manner**

- **cDNA chip**: mRNA expression profiling
  - Two-color format: ratio-metric analysis

- **Oligo chip (~ 50 mers):**
  - mRNA expression levels and profiling
  - SNPs (Single Nucleotide Polymorphisms)
  - Single color format: chip-to-chip variation
### Difference between cDNA and oligo chips

<table>
<thead>
<tr>
<th></th>
<th>cDNA chip</th>
<th>Oligo chip</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Probe</strong></td>
<td>PCR products or cDNA of the genes of interest</td>
<td>Oligonucleotides matching positions within the gene of interest</td>
</tr>
<tr>
<td><strong>Fabrication</strong></td>
<td>Spotted on the surface of the glass slide or nylon membrane</td>
<td>Oligonucleotides synthesized in situ on the surface of silicon membrane</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ex) photolithography</td>
</tr>
<tr>
<td></td>
<td>1 gene $\rightarrow$ one spot</td>
<td>1 gene $\rightarrow$ many probes</td>
</tr>
<tr>
<td></td>
<td>Two fluorescent dyes</td>
<td>Single fluorescent dye:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Match and mismatch</td>
</tr>
</tbody>
</table>
cDNA of known sequence are deposited on the surface of the glass or silicon in a grid-like array - 10,000 full-length cDNA in cm$^2$
Oligonucleotide chip

Multiple oligonucleotides synthesis using light directed combinatorial chemistry

Basic Oligo Synthesis

Initialize

Design

Photolithography

Chemical Synthesis Cycle

Remove Protecting Group (Deprotection)

Add Next Nucleotide

Protection Group (DMT)

DMT = 4,4'-dimethoxytrityl

Synthesis of oligonucleotides of known sequence (probes) on the surface of the glass or silicon in a grid-like array
- 250,000 different short oligonucleotide probes in cm²
General procedure

- mRNA is isolated from a matched sample of interest and a reference
- mRNA is typically converted to cDNA, labeled with fluorescence dyes (Cy3 and Cy5) or radioactivity
- Hybridization with the complementary probes on cDNA chip
- Analysis and comparison of expression levels of mRNAs between a sample and a reference
  → mRNA expression profiling
DNA Microarray Methodology - Flash Animation

www.bio.davidson.edu/Courses/genomics/chip/chip.html
mRNA expression profiling using cDNA microarray

- cDNA clones
- PCR amplification/Purification
- Robotic printing
- sample
- reference
- Hybridize targets to microarray
- Reverse transcriptase
- label with fluor dyes
- mRNA
- exitation
- Laser1
- Laser2
- emission
- Computer analysis

Cy3: ex 550 nm / em 570
Cy5: ex 649 nm / em 670

Green: up-regulated in a sample
Red: up-regulated in a reference
Yellow: equally expressed
Questions to be addressed using microarrays

- Normal versus disease tissue (cancer, autism)
- Wild-type versus mutant cells
- Cultured cells with or without drug
- Physiological states (hibernation, cell polarity formation)
Organisms represented on microarrays

- **Metazoans**: human, mouse, rat, worm, insect
- **Fungi**: yeast
- **Plants**: *Arabidopsis*
- **Other organisms**: e.g. bacteria, viruses

Metazoans: Any animal that undergoes development from an embryo stage with three tissue layers, namely the ectoderm, mesoderm, and endoderm.
Commercially available DNA chip

GeneChip®
Human Genome
U133 Plus 2.0

GeneChip®
Mouse Genome
430 2.0 Array

P/N: 520019
Lot #: 4010008
Exp. Date: 05/07/06
For Research Use Only

P/N: 520029
Lot #: 4015112
Exp. Date: 11/24/06
For Research Use Only
Overall procedure

Sample acquisition

Data acquisition

Data analysis

Data confirmation

Biological insight

Experimental design

Purify RNA, label

Hybridize, wash, image

Data storage
Stage 1: Experimental design

- Biological samples: technical and biological replicates
- mRNA extraction, conversion, labeling, hybridization
- Arrangement of array elements on a surface

Stage 2: RNA and probe preparation

- Confirm purity by running agarose gel
- Measure the absorbance at 260 and 280 nm and calculate the ratio to confirm purity and quantity
- Synthesis of cDNA and labeling using reverse transcriptase
Stage 3: Hybridization to DNA arrays

• Mixing of equal amounts of cDNA from a reference and a sample
• Load the solution to DNA microarray
• Incubation for hybridization followed by washing and drying

Stage 4: Image analysis

• Gene expression levels are quantified
• Fluorescence intensities are measured with a scanner, or radioactivity with a phosphoimage analyzer
Stage 5: Microarray data analysis

- How can arrays be compared?
- Which genes are regulated?
- Are differences authentic?
- What are the criteria for statistical significance?
- Are there meaningful patterns in the data (such as groups)?

Stage 6: Biological confirmation

- Microarray experiments can be thought of as “hypothesis-generating” experiments: Clues
- Differential up- or down-regulation of specific genes can be confirmed using independent assays:
  - Northern blots (detection of RNA or isolated mRNA)
  - Reverse transcription polymerase chain reaction (RT-PCR)
  - in situ hybridization to localize a specific DNA or RNA sequence using a labeled cDNA or RNA in a portion or section of tissue
Example of an approximately 37,500 probe-spotted oligo microarray with enlarged inset to show detail
Use of DNA microarray

Comparison of gene expression levels

- Different tissues
- Different environmental conditions (treated with drug)
- Normal and cancer cells
• Search for a specific gene(s) responsible for biological phenomenon → Starting point for basic research
• Search for biopharmaceuticals/drug targets
• Identification of potential biomarkers for diagnosis
• SNP detection

However, validation is needed
Search for a gene responsible for a disease

Normal

- Rett syndrome: childhood neuro-developmental disorder
- Normal early development followed by loss of purposeful use of the hands, distinctive hand movements, slowed brain and head growth, gait abnormalities, seizures, and mental retardation
- It affects females almost exclusively.

αB Crystallin is over-expressed in Rett Syndrome
### mRNA Expression profiling between lung cancer patient and normal person

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>alcohol dehydrogenase IB (class I), beta polypeptide</td>
<td>DOWN</td>
</tr>
<tr>
<td>mucolipin 1</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>UP</td>
</tr>
<tr>
<td>creatine kinase, brain</td>
<td>DOWN</td>
</tr>
<tr>
<td>artemin</td>
<td>DOWN</td>
</tr>
<tr>
<td>SP110 nuclear body protein</td>
<td>DOWN</td>
</tr>
<tr>
<td>apoptosis antagonizing transcription factor</td>
<td>UP</td>
</tr>
<tr>
<td>killer cell lectin–like receptor subfamily D, member 1</td>
<td>DOWN</td>
</tr>
<tr>
<td>fatty acid desaturase 3</td>
<td>DOWN</td>
</tr>
<tr>
<td>SH2 domain protein 2A</td>
<td>DOWN</td>
</tr>
<tr>
<td>cholinergic receptor, nicotinic, epsilon polypeptide</td>
<td>DOWN</td>
</tr>
<tr>
<td>ribosomal protein L29</td>
<td>UP</td>
</tr>
<tr>
<td>TGFB–induced factor 2 (TALE family homeobox)</td>
<td>DOWN</td>
</tr>
<tr>
<td>ectonucleoside triphosphate diphosphohydrolase 2</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>TBC1 domain family, member 8 (with GRAM domain)</td>
<td>DOWN</td>
</tr>
<tr>
<td>3-hydroxymethyl–3-methylglutaryl–Coenzyme A lyase (hydroxymethylglutaricuria)</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>homeo box D4</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>eukaryotic translation initiation factor 3, subunit 8, 110kDa</td>
<td>UP</td>
</tr>
<tr>
<td>Rho guanine nucleotide exchange factor (GEF) 10</td>
<td>DOWN</td>
</tr>
<tr>
<td>aquaporin 5</td>
<td>DOWN</td>
</tr>
</tbody>
</table>

**Choi et al., J Thorac Oncol (2006), 1, 622-628**

...
Expression profiles under different nutritional conditions

cDNA microarray chip containing 2,500 genes from yeast

- Expression profiling of genes from Yeast grown at 2% galactose and glucose
  - **Green**: up-regulation in yeast grown at galactose
  - **Red**: up-regulation in yeast grown at glucose
  - **Yellow**: equally expressed

Lashkari *et al.*, *PNAS* (1997)
<table>
<thead>
<tr>
<th>Advantage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast</td>
<td>Data on &gt;20,000 genes in several weeks</td>
</tr>
<tr>
<td>Comprehensive</td>
<td>Entire yeast, mouse, and human genome on a chip</td>
</tr>
<tr>
<td>Flexible</td>
<td>- As more genomes are sequenced, more arrays can be made.</td>
</tr>
<tr>
<td></td>
<td>- Custom arrays can be made to represent genes of interest</td>
</tr>
<tr>
<td>Easy</td>
<td>Submit RNA samples to a core facility</td>
</tr>
<tr>
<td>Cheap</td>
<td>Chip representing 20,000 genes for $350; Robotic spotter/scanner cost $50,000</td>
</tr>
</tbody>
</table>
## Disadvantages of microarray experiments

**Cost:** Some researchers can’t afford to do appropriate controls, replicates.

### mRNA significance
- Final products of gene expression are proteins

### Quality control
- Impossible to assess elements on array surface
- Artifacts with image analysis
- Artifacts with data analysis
Why Proteomics?

- Proteins: directly involved in most of biological functions and drug targets
- Protein expression levels can not be accurately detected or measured via DNA array technology

- Levels of mRNA-encoded polypeptides are not directly correlated with those of mRNAs
- A significant proportion of eukaryote mRNA undergoes differential splicing, and can yield more than one polypeptide product
- No detailed information regarding how the functional activity of expressed proteins will be regulated.

(e.g., post-translational modifications: phosphorylation, glycosylation, ubiquitination, sumoylation, proteolysis)
Proteins are responsible for most of biological functions, drug targets, or potential biomarkers:

→ More successfully identified by direct analysis of the expressed proteins in the cell

Systematic and comprehensive analysis of the proteins (proteom) expressed in the cell and their functions

- Direct comparison of protein expression levels
- Changes in cellular protein profiles with cellular conditions
Proteomics approach by 2-D protein gels

- **General procedure**
  - Extraction of the total protein contents from the target cell/tissue and corresponding reference
  - Separation of proteins by 2-D gel electrophoresis
    - Dimension one: iso-electric focusing
    - Dimension two: SDS-PAGE (polyacrylamide gel)
  - Elution of protein spots with different expression levels
  - Analysis of eluted proteins for identification
Electro-focusing: a technique for separating different molecules by differences in their isoelectric point (pI)

- Acrylamide gel matrix co-polymerized with the pH gradient

- A protein that is in a pH region below its isoelectric point (pI) will be positively charged and will migrate towards the CATHODE (negative).

- As it migrates through a gradient of increasing pH, however, the protein's overall charge will decrease until the protein reaches the pH region that corresponds to its pI. At this point it has no net charge and so migration stops.

- The technique is capable of extremely high resolution with proteins differing by a single charge being fractionated into separate bands.
At low pH, most proteins have a positive charge while at high pH, most proteins have a negative charge.

When an electric field is present, the cathode and anode ends pull the proteins to their isoelectric point where each individual protein possesses a neutral charge.

The proteins stopped migrating because they’ve reached their isoelectric point at a unique pH level.

- ● = Isoelectric point at pH 7.5
- ○ = Isoelectric point at pH 6.8
- ○ = Isoelectric point at pH 8.5
- ▲ = Isoelectric point at pH 10.1
- ▼ = Isoelectric point at pH 5.6
2-D gel electrophoresis between two different conditions
How to figure out which spot is what?

- **Protein micro-sequencing** using Edman degradation protocol (partial amino acid sequence) : laborious and time-consuming

- **Protein analysis using mass spectrometry**
  - Molecular mass of protein : MALDI-TOF
  - Digestion pattern by Trypsin : MALDI-TOF
  - Amino acid sequence of a digested peptide : Tandum mass spectrometry

- Use a core facility, or collaborate with experts

- **Identification or assignment of protein function by sequence homology search**
Basic components of a mass spectrometer

- **Ion source**: Convert sample molecules into ions (ionization)
- **Mass analyzer**: Sorts the ions by their masses by applying electromagnetic fields
- **Detector**: Measures the value of an indicator quantity and thus provides data for calculating the abundances of each ion present
Mass spectrometers

- MALDI (Matrix Assisted Laser Desorption Ionization)
- ESI (Electrospray Ionization)
- EI (Electron Ionization)
- CI (Chemical Ionization)
- FAB (Fast Atom Bombardment)

- TOF (Time of Flight)
- Quadrupole
- FT-ICR (FTMS)
- Ion Trap
Time of Flight in mass spectrometry

- Ions are accelerated by an electric field of known strength.
  - Resulting in an ion having the same kinetic energy as any other ion that has the same charge.
  - The velocity of the ion depends on the mass-to-charge ratio.

- The time for the ion to reach a detector at a given distance is measured.
  - The elapsed time from the instant a particle leaves a source to the instant it reaches a detector.
  - This time will depend on the mass-to-charge ratio of the ions.

- From this time and the known experimental parameters, mass-to-charge of the ion is determined.
Time of Flight (TOF)

- When the charged particle is accelerated into time-of-flight tube by the voltage $U$, its kinetic energy of any mass is $\frac{1}{2} mv^2$.
- The smaller the molecular mass, the higher the velocity of a molecule; Calculate the m/z by measuring the flight time.
- Mass-to-charge (m/z) ratio of a molecule is determined by measuring the flight time in the tube.

\[ E = \frac{1}{2} m v^2 \]
\[ v = \sqrt{\frac{2E}{m}} \]
\[ t = \frac{\text{length}}{\sqrt{\frac{2E}{m}}} \]
Mass to charge ratio: m/z ratio

- **M**: mass,
- **Z**: charge number of ions. For positive ions, the number of electrons removed.
  
  ex) Single charged ions: an electron is taken from molecules.
  
  Double charged ions: two electrons are removed.

- m/z represents mass divided by charge number.
- z is almost always 1 with GC/MS. The m/z value is considered to be the mass.
Matrix-assisted laser desorption/ionization (MALDI)

- Soft ionization technique: Analysis of biomolecules (such as protein, peptides, and sugars) and large organic molecules, which tend to be fragile and fragmented when ionized by conventional ionization methods
- The matrix absorbs the laser energy, and the matrix is ionized (by addition of a proton)
- The matrix transfers a proton to the target molecules (e.g., protein molecules), charging them
- Commonly used matrix
  - 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid),
  - α-cyano-4-hydroxycinnamic acid (alpha-cyano or alpha-matrix)
  - 2,5- dihydroxybenzoic acid (DHB)
- Detection procedure
  - Ions are accelerated by an electrical field to the same kinetic energy
  - The velocity of the ion depends on the mass-to-charge ratio.
  - From the elapsed time to reach a detector, the mass-to-charge ratio can be determined.
Peptide mass fingerprinting (Protein fingerprinting)

- Analytical technique for protein identification
  - Unknown protein of interest is first cleaved into smaller peptides by proteases
  - Their absolute masses can be accurately measured with a mass spectrometer such as MALDI-TOF or ESI-TOF.
  - The peptide masses are compared to either a database containing known protein sequences or even the genome using computer programs.
  - Programs translate the known genome of the organism into proteins, then theoretically cut the proteins into peptides, and calculate the absolute masses of the peptides from each protein.
  - Compare the masses of the peptides of the unknown protein to the theoretical peptide masses of each protein encoded in the genome.
  - The results are statistically analyzed to find the best match.
Peptide Mass Fingerprinting

Protein

Trypsin

Tryptic peptide mixture. Masses measured by MS. Every peptide has a basic C-terminus.

A protein can be identified in a database by matching masses of a subset of the tryptic peptides against calculated values.
Peptide fragmentation fingerprinting

PFF = ion search MS/MS database matching

Protein(s) → Enzymatic digestion → Peptides → MS/MS spectra of peptides → Ions peaklists

In-silico digestion

...MAIIAGGHSVRFVPKAF AEVGETFYSRTVLEGSTNM FNEIIIISTNAQATQFYPN VVIDDEHNDKGPLAGTYI MKQHPEELFFVVSVDPFMI TGKAVSTLYQFLV ...

In-silico fragmentation

- MAIIAGGHSVVR - FPK - AFAEVGETFYSR - VTVLESTNMFNEIIK - YPNVVIDDENDK ...

Sequence database entry

Theoretical proteolytic peptides

Theoretical fragmented peptides

Theoretical peaklist

Match

Result: ranked list of peptide and protein candidates
Direct identification of proteins responsible for cellular functions under specific conditions:
- Differentiation, development, etc.
- Treatment of drugs, stress etc.
- Identification of key enzymes in metabolic pathways
  → Construction of new strains

Discovery of disease biomarkers
- Comparison of protein expression levels between patient and normal person
  → Protein expression profiling
  → Potential disease biomarkers
• Breast cancer
  - Most commonly diagnosed cancer among women in the US
  - Account for 31% (203 500) of all new cancer cases among women, and 39 600 death

• Despite the availability and recommended use of mammography as a routine screening method for women 40 years of age and older, its effectiveness in reducing overall population mortality from breast cancer is still being investigated

• Currently, serum tumor markers, such as CA15.3 and CA27.29, are used in breast cancer detection, but still lack the adequate sensitivity (23%) and specificity (69%) to be applicable in detecting early-stage carcinoma in a large population

• Approved tumor markers: CA15.3 and CA27.29
  - only for monitoring therapy of advanced breast cancer or recurrence
- **CA15-3 (Cancer Antigen 15-3):**
  - A tumor marker used to monitor breast cancer.
  - Expressed on the surface of cancer cells and released into the blood.
  - Main uses: preclinically detecting recurrent breast cancer and monitoring the treatment of patients with advanced breast cancer, i.e., metastatic cancer.
  - Criteria: >30.4 U/mL

- **CA 27-29 (Cancer antigen 27-29):**
  - Approved by the FDA for monitoring people with breast cancer.
  - Criteria:
    - < 38 U/mL: no active breast cancer.
    - >=38 U/mL: active breast cancer, or recurrence of the cancer or metastasis.
• Discovery of new biomarkers that could be used individually or in combination with an existing modality for cost-effective diagnosis of breast cancer

• Proteomics approach
  - Total 169 serum samples
  - 103 samples from a cancer group of 103 breast cancer patients at different clinical stages: stage 0 (n = 4), stage I (n = 38), stage II (n = 37), and stage III (n = 24)
  - Control group of 41 healthy women, and from 25 patients with benign breast diseases

• Results
  - Identification of three potential biomarkers for breast cancer
    BC1 (4.3 kDa), BC2 (8.1 kDa). BC3 (8.9kDa)
  - Needs further validation for approval
L-Threonine

- Essential amino acid
- Feed and food additives
- Raw material for synthesis of various medicines

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Production (MT / annum)</th>
<th>Capacity (MT / annum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-Lysine-HCl</td>
<td>583,000</td>
<td>704,000</td>
</tr>
<tr>
<td>DL-Methionine</td>
<td>496,000</td>
<td>680,000</td>
</tr>
<tr>
<td>L-Threonine</td>
<td>27,000</td>
<td>49,000</td>
</tr>
</tbody>
</table>

Source: Feedinfo. 2002

Use of proteomics in metabolic pathway engineering
Biosynthetic pathway of L-Thr in *E. coli*

- **Glucose** → **Phosphoenolpyruvate** → **Pyruvate**
- **TCA cycle**: **Oxaloacetate** → **AspC**
- **Feedback repression**

**Key Pathway Enzymes**:
- **metL**: L-Aspartate
- **thrA**: L-Aspartate semialdehyde
- **asd**: Homoserine
- **thrB**: Homoserine phosphate
- **thrC**: L-Threonine
- **ilvA**: L-Isoleucine
- **dapA**: L-Lysine
- **metA**: L-Methionine
Development of an L-Threonine-overproducing strain

- Conventional mutagenic method

- Use of protein expression profiles in biosynthetic pathway between parent and an L-threonine-producing strain

  - Production level of L-threonine
    - W3110 (Wild-type *E. Coli*) : < 0.001 g/L
    - TF 5015 (Mutant) : ~ 20 g/L
Proteome analysis of two strains

Identification of protein spots by MALDI-TDF

Lee et al., J Bacteriol (2004)