Impact of genomics and proteomics on Biotechnology / Biology

- **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

---

*Graph showing the decrease in cost per genome from September 2001 to April 2014.*
• Human genome project
  - Started in 1990
  - Completed in 2003: ~3.2 giga bases (Gb),
    1,000 times larger than a typical bacterial genome
  - Less than 1/3 of the genome is transcribed into mRNA
  - Only 5% of the RNA encodes polypeptides
    → Number of polypeptide-encoding genes: ~30,000

• Genome sequences of more than 2,000 organisms
  - Genomes of various animals and plants: mouse, rat, sheep, pig, monkey, dog, chicken, wheat, barley, Arabidopsis
Significance of genome data in Biotechnology/Biology

- **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 complex diseases and development of therapies.

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

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

  ➔ Translational research
Functional genomics

• 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 product acts, and what other cellular elements it interacts with → **Interactome**
  - How such interactions contributes 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
  - Grow crystals, solve 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
✓ Phylogenetic profiling

• 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 upon similarities and differences in their physical and/or genetic characteristics

• [http://media.hhmi.org/biointeractive/click/Phylogenetic_Trees/01.html](http://media.hhmi.org/biointeractive/click/Phylogenetic_Trees/01.html)

• Phylogenetic profiling: Study of evolutionary relationships among various biological species or other entities based on similarities and differences in their physical and/or genetic characteristics

• Closely related species should be expected to have very similar sets of genes
• Proteins that function in the same cellular context frequently 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 sequenced:

• 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, 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?

• Orthologs: 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.

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

• 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.
Glutamate decarboxylation pathway
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.
• **DNA microarray (Gene chip):**
  - Comparison of mRNA expression levels between a sample (cancer cell) and a reference (normal cell) in a high-throughput way: **mRNA expression profiling**

- **cDNA chip:** mRNA expression profiling
- **Oligo chip ( ~ 50 mers):** mRNA expression profiling
  - SNPs (Single Nucleotide Polymorphisms)
  - Complementary probes are designed from gene sequence
- Solid support (such as glass microscope slide) on which DNA of known sequence (probes) is deposited in a grid-like array
  - 250,000 different short oligonucleotide probes in cm$^2$
  - 10,000 full-length cDNA in cm$^2$

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

www.bio.davidson.edu/Courses/genomics/chip/chip.html
Comparative study using microarrays

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

- Metazoans: human, mouse, rat, worm, insect
- Fungi: yeast
- Plants: *Arabidopsis*
- Other organisms: e.g. bacteria, viruses
mRNA expression profiling using cDNA microarray

- cDNA clones
- PCR amplification
- Purification
- Robotic printing
- Sample
- Reference
- mRNA
- Reverse transcriptase
- Hybridize targets to microarray
- Laser1: exitation
- 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
Commercially available DNA chip
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
PCR (Polymerase Chain Reaction)

- Developed in 1983 by Kary Mullis
- Nobel prize in Chemistry in 1993

Melting at 95 °C
Annealing at 55 °C
Elongation at 72 °C

Thermostable DNA polymerase from thermophilic bacterium
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

- Mix 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 phosphorimage analyzer
Example of an approximately 37,500 probe-spotted oligo microarray with enlarged inset to show detail
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
  - Polymerase chain reaction (RT-PCR)
  - in situ hybridization
Use of DNA microarray

Comparison of gene expression levels

- Different tissues
- Different environmental conditions (treated with drug)
- Normal and cancer cells
Outcome of data analysis

- 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

But, need validation
Rett syndrome is a childhood neurodevelopmental disorder characterized by 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 in lung cancer patient and normal person using DNA Microarray

<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>TGFβ–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 (hydroxymethylglutaricaciduria)</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>
<tr>
<td>cytochrome b–561</td>
<td>DOWN</td>
</tr>
<tr>
<td>TATA box binding protein (TBP)–associated factor, 32kDa</td>
<td>DOWN</td>
</tr>
<tr>
<td>glypican 4</td>
<td>DOWN</td>
</tr>
<tr>
<td>AT rich interactive domain 4A (RBP1–like)</td>
<td>DOWN</td>
</tr>
<tr>
<td>TEA domain family member 4</td>
<td>UP</td>
</tr>
<tr>
<td>G protein–coupled receptor 50</td>
<td>DOWN</td>
</tr>
<tr>
<td>ret finger protein 2</td>
<td>DOWN</td>
</tr>
<tr>
<td>chromosome 11 open reading frame 24</td>
<td>UP</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>minichromosome maintenance deficient 3 (S. cerevisiae)</td>
<td>UP</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>defensin, alpha 6, Paneth cell–specific</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>small nuclear ribonucleoprotein polypeptide C</td>
<td>UP</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>HLA–B associated transcript 3</td>
<td>UP</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
<tr>
<td>mitogen–activated protein kinase kinase kinase 6</td>
<td>UP</td>
</tr>
<tr>
<td>null</td>
<td>DOWN</td>
</tr>
</tbody>
</table>

**Choi et al., J Thorac Oncol (2006), 1, 622–628**
Expression profiles under different nutritional conditions

- 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

- cDNA microarray chip containing 2,500 genes from yeast

Lashkari *et al.*, *PNAS* (1997)
## Advantages of microarray experiments

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

<table>
<thead>
<tr>
<th>mRNA significance</th>
<th>Final products of gene expression are proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality control</td>
<td>Impossible to assess elements on array surface</td>
</tr>
<tr>
<td></td>
<td>- Artifacts with image analysis</td>
</tr>
<tr>
<td></td>
<td>- Artifacts with data analysis</td>
</tr>
</tbody>
</table>
Proteins are directly involved in most of biological functions

- Drug targets: mostly proteins
- Protein expression levels cannot 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 content from the target cell/tissue

- Separation of proteins by 2-D gel electrophoresis
  Dimension one: iso-electric focusing
  Dimension two: SDS-PAGE (polyacrylamide gel)

- Elution of protein spots

- Analysis of eluted proteins for identification
Electro-focusing: a technique for separating different molecules by differences in their isoelectric point

- The gradient is established by first subjecting a solution of small molecules such as polyampholytes with varying pI values to electrophoresis.

- 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.

- As a result, the proteins become focused into sharp stationary bands with each protein positioned at a point in the pH gradient corresponding to its pI.

- The technique is capable of extremely high resolution with proteins differing by a single charge being fractionated into separate bands.
2-D gel electrophoresis between two different conditions

ISOELECTRIC FOCUSING

pH = 3  →  pH = 10

SDS - PAGE

pH = 3  →  pH = 10

MASS SPECTROMETRY
How do you 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 : Tandem mass spectrometry

- Usually have a core facility do these, 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.
  - This acceleration results 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 known distance is measured.
  - This time will depend on the mass-to-charge ratio of the ions.
  - The elapsed time from the instant a particle leaves a source to the instant it reaches a detector.

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

- When a 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
Matrix-assisted laser desorption/ionization (MALDI)

- Soft ionization technique allows the analysis of biomolecules (such as protein, peptides, and sugars) and large organic molecules, which tend to be fragile and fragmented when ionized by more conventional ionization methods

- The matrix absorbs the laser energy, and the matrix is ionized (by addition of a proton) by this event.

- The matrix then transfers proton to the analyte molecules (e.g., protein molecules), thus charging the analyte

- 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)
- 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.
Use of proteomics

• 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
L-Threonine

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

### World-wide production of amino acids

<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
Biosynthetic pathway of L-Thr in *E. coli*

1. **Glucose** → **Phosphoenolpyruvate** → **Pyruvate**
2. **TCA cycle** (via **mdh** and **aceBAK**)
3. **L-Aspartate** (via **metL**, **thra**, **lysC**)
4. **L-Aspartate phosphate** (via **aspC**)
5. **L-Aspartate semialdehyde** (via **asd**)
6. **Homoserine** (via **thrA**)
7. **Homoserine phosphate** (via **thrB**)
8. **L-Threonine** (via **thrC**)
9. **L-Isoleucine** (via **ilvA**)
10. **L-Lysine** (via **dapa**)
11. **L-Methionine** (via **metA**)

---

**Feedback repression**
Development of an L-Threonine-overproducing strain

- Conventional mutagenic method
- Use of protein expression profiles in biosynthetic pathway between a 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)