Concentrations of mRNAs within a cell are poorly correlated with the actual abundances of the corresponding proteins.

Protein chip:
A speedy and high throughput means to profile expression levels of proteins, and to study **protein-protein interactions** and **protein-drug interactions**

Key components
- Diverse proteins are printed onto a solid surface to make arrays of them
- Capture agents to recognize and bind the target ligands
- Analysis of bound proteins by various methods
✓ Critical issues
  - Different Protein molecules should be deposited in a biologically active form at separate locations

  - Non-specific protein binding should be minimized

  - Detection methods must have a much larger range of detection. Protein concentrations in a biological sample may be many orders of magnitude different from that for mRNAs.

  ex) Protein concentrations in a single biological sample vary far more than $10^{14}$ mRNA : a factor of $10^4$
Construction of protein microarray

Surface chemistry
- Coated slides: e.g. polylysine, aldehyde
- Proprietary surface chemistries

Capture agents
- Antibodies: not very specific but several large libraries are available
- Aptamers: nucleic acids that bind to proteins
- Fibronectins: used to generate antibody mimics
- Phage display peptides: peptides that bind to proteins

Detection methods
- Chemiluminescence: sensitive but requires enzymatic reaction
- Fluorescence: sensitive but requires labelling of the protein
- Mass spectrometry: low throughput but no labelling required
- Surface plasmon resonance: low throughput but no labelling required
Protein Microarray

Features
- High throughput analysis
- Small amounts of reagents

Protein expression profiling
- Analysis of protein expression levels between a reference and a sample
- Discovery of disease biomarker
- Physiological response to toxin, drug and environmental conditions

Protein interaction analysis
- Discovery of protein binding partner
- Protein interaction network
- High-throughput screening of drug target
- Understanding of basic biology
Core Technologies in Protein Microarray

Protein chips for practical use

Detecting the biomolecular interaction with high sensitivity and reliability

How to construct the monolayers of capture molecules on a solid surface

- Maximizing the binding efficiency
- Maximizing the fraction of active biomolecules
- Minimizing the nonspecific protein binding
Protein immobilization

A. Random Attachment

- Random immobilization via different chemistries including aldehyde- and epoxy-treated slides that covalently attach protein by their primary amines or by adsorption onto slides coated with nitrocellulose or acrylamide gel pads.

B. Uniformly orientated immobilization onto slides coated with a ligand.
  - His6X-tagged proteins can be bound to nickel-derivatized slides
  - Biotinylated proteins can be attached to streptavidin-coated slides.
  - Orientate the binding site of protein away from the slide surface.
Protein chips and their analysis

- **Antibody chip**
- **Antigen chip**
- **Protein chip**
- **Ligand chip**
- **Carbohydrate chip**

1. **Sera + antibodies**
   - **Antigen–antibody**

2. **Antibodies**
   - **Antigen–antibody**

3. **Labeled proteins**
   - **Protein–liposome**
   - **Protein–drug**
   - **Enzyme–substrate**

4. **Liposome**
   - **Protein–liposome**

5. **Drugs**
   - **Protein–drug**

6. **Enzyme**
   - **Enzyme–substrate**

7. **Proteins**
   - **Receptor–ligand**

8. **Sera**
   - **Antibody–carbohydrate**
Global Analysis of Yeast Proteom by Protein Chips

- Protein microarray with 5,800 ORFs from Yeast (~80 % whole proteins)

- Major hurdle in proteom analysis: Cloning and expression of clones and purification of proteins in a high throughput manner

- GST-HisX6-fused proteins → Proteom microarray on nickel-coated or aldehyde-treated slide (purification using glutathione-agarose beads)

Identification of new calmodulin and phospholipid-interacting proteins

- Yeast proteom chip: commercially available

Snyder et al., Science (2001)
Probing the chip with antibodies to GST (Anti-GST Ab)
- How much a fusion protein was covalently attached
- Reproducibility of the protein attachment to the slide

A. Immunoblot analysis of purified proteins using anti-GST antibody
B. Proving of 5800 proteins with Cy5 labeled anti-GST antibody
C. An enlarged image of one of the 48 blocks

Each spot contained 10 to 950 fg of protein
Analysis of protein-protein interactions: Calmodulin binding proteins

- **Calmodulin**: Highly conserved calcium-binding protein involved in many calcium-regulated cellular processes
- **Yeast proteom was probed with biotinylated calmodulin in the presence of calcium**
- **Bound calmodulin was detected with Cy3-labeled streptavidin**: Streptavidin-biotin interaction

Analysis of protein-lipid interactions

- **Phosphoinositide (PI)**: Second-messenger which regulates the cellular process like growth, differentiation, and cytoskeletal rearrangement
- Important constituents of cellular membrane
- Probing with anti-GST antibody labeled with Cy5
- Loading with biotinylated calmodulin in the presence of $\text{Ca}^{2+}$
- Detection of calmodulin-binding proteins by Cy3-labeled streptavidin
- Loading with biotinylated phosphoinositide (PI)
- Detection using Cy3-labeled streptavidin

**Putative calmodulin binding motifs of 14 positive proteins**

Conserved sequence is (I/L)QXK(K/X)GB

The size of the letter: relative frequency of the amino acid indicated

- X: any residue
- B: basic residue
Analysis of PI (phosphoinositide)-binding proteins

43 strong
19 weak
Protein kinase

- **Enzymes that mediate phosphorylation of target proteins by transferring phosphorous group from ATP to the target proteins**
  - Up to 30% of all proteins: Substrates of various protein kinases
  - Regulate the majority of cellular pathways, especially involved in signal transduction, the transmission of signals within the cells.
  - **Human genome**: about 500 protein kinase genes: they constitute about 2% of all eukaryotic genes.

- **Disregulated kinase activity**: Frequent cause of disease, particularly cancer, where kinases regulate many aspects that control cell growth, movement, and death.

- **Important drug target**: Drugs which inhibit specific kinases are currently in clinical use → TKI (Tyrosine kinase inhibitor)
  - ex) Gleevec by Novartis (leukemia)
  - Iressa by AstraZeneca (lung cancer)
Analysis of Yeast Protein Kinases using protein chip: Substrate specificity of Kinases

**Yeast genome**: 6,200 ORFs

- 122 Protein kinases:
  - Ser/Thr family (120 Kinases), Tyr family, Poly (Tyr-Glu)

**Experimental Procedures for protein kinase assay**

- **Expression and purification of 119 kinases** in GST-fusion proteins in yeast
- **Protein chips composed of an array of microwells** using silicone elastomer
  - **10×14 array** on (PDMS :poly(dimethylsiloxane))
    (each well : 1.4 mm in diameter, 300 μm deep, 300 nL)
  - Covalent attachment of **17 substrates on 17 protein chips** (8×10⁻⁹ μg/μm² in each spot) using a cross-linker (GTPS)
- **Kinase reaction and high-throughput assay**
  - $^{33}$Pγ-ATP and protein kinase, incubation for 30 min at 30 °C, washed, exposed to X-ray film and phosphoimager

Protein chip fabrication and kinase assays

- PDMS is poured onto etched mold
- The chip containing the wells is peeled away and mounted on a glass slide for easy handling
- Modification of the surface and covalent attachment of kinase substrates to the wells:
  - 17 different substrates
  - Each substrate on respective protein chip

- Each protein kinase and $^{33}\text{P}_\gamma$-ATP are added into the microwell followed by incubation for 30 min
- Extensive washing and expose to both X-ray film and a phosphoimager

Master mold

(3-glycidoxypropyl) trimethoxy silane
Kinase assays on different kinase Substrates

- Kinase themselves (autophosphorylation)
- Bovine casein
- Bovine histone H1
- Myelin basic protein
- Ax12 Carboxy-terminus GST
- Rad 9 (a phospho-protein involved in the DNA damage checkpoint)
- Gic2 (involved in budding)
- Red1 (involved in chromosome synapsis)
- Mek1 (involved in chromosome synapsis)
- Poly (Tyr-Glu)
- Ptk2 (transport protein)
- Hsl1 (involved in cell cycle regulation)
- Swi6 (involved in G1/S control)
- Tub4 (involved in microtuble nucleation)
- Hog1 (involved in osmoregulation)
- Inactive form of Hog1
- GST (Control)
New findings

✓ 18 predicted protein kinases phosphorylate one or more substrates

✓ Many yeast kinases phosphorylate poly(Tyr-Glu)

✓ Large-scale analysis of yeast protein kinases enables correlation between functional specificity and amino acid sequences of the poly(Tyr-Glu) kinases
Perspective

✓ Biochemical assay of protein kinases in a high throughput way: Extremely powerful to analyze thousands of proteins using a single protein chip

✓ Applicable to wide variety of additional assays: Nuclease, helicase, protein-protein interaction assays

✓ Facilitate high throughput drug screening for inhibitors or activators of any enzymes
**Biomarker**

- **Biomarker**: A biochemical feature or facet that can be used to predict the progress of disease or the effects of treatment.

- **Non-invasive diagnosis vs biopsy**
  - Thousands of proteins in serum offer an opportunity to identify potential biomarkers.
  - Detection of biomarkers in serum.
  - General method to discover disease-related biomarkers.
    - Expression profiling of proteins in serum between patients and healthy persons.
Profiling of Proteins in Human Sera For Identification of Lung Cancer Biomarkers using Antibody microarray

In collaboration with Samsung Medical Center, SKKU

Han et al., Proteomics (2009)
Protein Profiling

To understand the condition of cell, mechanism of disease, know how proteins function in cells, and identify biomarkers for diagnosis.

Profiling of protein expression in a highly parallel manner

Protein Microarray
Lung Cancer (2006)

- 175,000 new cases (USA) ; 1.2 million worldwide
- Cure rate for all patients: 15%
- Most common cause of cancer death in US for both men and women
- 85% in current/former smokers; 10-15% in never smokers
- Median survival is 8-10 months with chemotherapy
Lung Cancer Mortality in USA

Male

Female
Early diagnosis

- Biomarkers: Lung cancer vs. Abnormal expression of proteins
- Analysis of serum
  Several thousands of proteins in serum give opportunity to find biomarkers

Acquired resistance:
- Reduction in effectiveness of a drug in curing a disease or improving patients’ symptoms
- Major lung cancer deaths: Acquired resistance against chemotherapy
## mRNA Profiling 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>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 (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>null</td>
<td>UP</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>defensin, alpha 6, Paneth cell–specific</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>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
Target Proteins for Ab Microarray

- 19 target proteins were selected based on mRNA profiling (DNA array) and commercial availability of Ab.
  - MMP7\(^1\), SAA\(^2\) and VEGF\(^3\) were chosen from previous reports.
  - Anti-Human IgG as positive control, Anti-OppA as negative control.

<table>
<thead>
<tr>
<th>no.</th>
<th>Abbreviation</th>
<th>Protein full name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>hIgG</td>
<td>human Immunoglobulin G</td>
</tr>
<tr>
<td>2</td>
<td>AATF</td>
<td>apoptosis-antagonizing transcription factor</td>
</tr>
<tr>
<td>3</td>
<td>ADH1B</td>
<td>alcohol dehydrogenase 1B</td>
</tr>
<tr>
<td>4</td>
<td>AQP5</td>
<td>aquaporin 5</td>
</tr>
<tr>
<td>5</td>
<td>ARHGEF10</td>
<td>rho guanine nucleotide exchange factor 10</td>
</tr>
<tr>
<td>6</td>
<td>ARID4A</td>
<td>ARID domain-containing protein 4A</td>
</tr>
<tr>
<td>7</td>
<td>ARTN</td>
<td>artemin precursor</td>
</tr>
<tr>
<td>8</td>
<td>CHRNE</td>
<td>acetylcholine receptor protein subunit epsilon precursor</td>
</tr>
<tr>
<td>9</td>
<td>CKB</td>
<td>creatine kinase brain</td>
</tr>
<tr>
<td>10</td>
<td>DEFA6</td>
<td>defensin 6 precursor</td>
</tr>
<tr>
<td>11</td>
<td>EIF3S8</td>
<td>eukaryotic translation initiation factor 3 subunit 8</td>
</tr>
<tr>
<td>12</td>
<td>FKSG14</td>
<td>centromere protein K</td>
</tr>
<tr>
<td>13</td>
<td>GPR50</td>
<td>G protein-coupled receptor 50</td>
</tr>
<tr>
<td>14</td>
<td>HMGCL</td>
<td>hydroxymethylglutaryl-CoA lyase</td>
</tr>
<tr>
<td>15</td>
<td>KLRD1</td>
<td>CD94</td>
</tr>
<tr>
<td>16</td>
<td>MAP3K6</td>
<td>mitogen-activated protein kinase kinase kinase 6</td>
</tr>
<tr>
<td>17</td>
<td>MCM3</td>
<td>minichromosome maintenance protein</td>
</tr>
<tr>
<td>18</td>
<td>RFP2</td>
<td>arfaptin-2</td>
</tr>
<tr>
<td>19</td>
<td>TAF9</td>
<td>TATA-binding protein</td>
</tr>
<tr>
<td>20</td>
<td>TGIF2</td>
<td>5'-TG-3' interacting protein</td>
</tr>
<tr>
<td>21</td>
<td>MMP7</td>
<td>matrix metalloproteinase7</td>
</tr>
<tr>
<td>22</td>
<td>SAA</td>
<td>serum amyloid A</td>
</tr>
<tr>
<td>23</td>
<td>VEGF</td>
<td>vascular endothelial growth factor</td>
</tr>
<tr>
<td>24</td>
<td>OppA</td>
<td>periplasmic oligopeptide-binding protein precursor</td>
</tr>
</tbody>
</table>

Factors affecting the performance of protein microarray

Surface chemistry
- Non-specific binding & low background

Capturing agent
- Cross-reactivity & Loss-of-function

Detection methods
- Dye-labeling efficiency

Reliable microarray data

Han et al., Proteomics (2006)
Dye-labeling and Array Analysis

- Antibody arrays on glass slides by robotic arrayer
- Both dye labeling to reference (pooled healthy serum) and samples (cancer serum)
- 15-fold diluted serum with dye-conjugation buffer
- Analysis of the Internally Normalized Ratio (INR)

Cancer serum → Cy3-NHS or Cy5-NHS labeling → Centrifugation using Centricon (cutoff: 10 kDa) → Recovery → Equally mixing → Cy5-Cancer / Cy3-Pooled → #1

Pooled Healthy serum → Centrifugation using Centricon (cutoff: 10 kDa) → Recovery → Equally mixing → Cy3-Cancer / Cy5-Pooled → #2

< Hydrogel-coated glass slide >
The ratio of the ratios from both slides can be defined as follows:
- \( X \) : the Cy5/Cy3 ratio obtained slide #1
- \( Y \) : the Cy5/Cy3 ratio obtained slide #2

\[
X / Y = \frac{\text{Cancer}^{\text{Cy5}}}{\text{Normal}^{\text{Cy3}}} / \frac{\text{Normal}^{\text{Cy3}}}{\text{Cancer}^{\text{Cy3}}} = \frac{\text{Cancer}^{\text{Cy5}}}{\text{Normal}^{\text{Cy3}}} \times \frac{\text{Cancer}^{\text{Cy3}}}{\text{Normal}^{\text{Cy5}}} = R_1 \times R_2
\]

In ideal case, INR can be defined as follows:

\[
INR = \sqrt{R_1 \times R_2} = R
\]

- Reduction in the number of false positives
- Useful as universal conditions without normalization
Protein Profiling in Sera:
19 Lung Cancer Patients vs. 8 The Healthy

Reference: Pooled serum of 8 normal

Test\textsuperscript{Cy5} / Reference\textsuperscript{Cy3}

Test\textsuperscript{Cy3} / Reference\textsuperscript{Cy5}

Healthy

Cancer

Biomarker candidates
Validation of identified proteins:
Western blotting

Han et al., Proteomics (2009)
Protein Profiling in Sera:
Lung cancer vs. other types of samples

P-value (t-test) < 0.03

<table>
<thead>
<tr>
<th></th>
<th>AQP5</th>
<th>ARTN</th>
<th>CKB</th>
<th>MCM3</th>
<th>TAF9</th>
<th>TGIF2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>-</td>
<td>0.0096</td>
<td>0.0004</td>
<td>0.0164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-</td>
<td>0.1013</td>
<td>0.0746</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>0.3020</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AQP5</th>
<th>ARTN</th>
<th>CKB</th>
<th>MCM3</th>
<th>TAF9</th>
<th>TGIF2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>-</td>
<td>0.0078</td>
<td>0.0072</td>
<td>0.0142</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-</td>
<td>0.2188</td>
<td>0.0831</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>0.2036</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AQP5</th>
<th>ARTN</th>
<th>CKB</th>
<th>MCM3</th>
<th>TAF9</th>
<th>TGIF2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>-</td>
<td>0.0153</td>
<td>0.0094</td>
<td>0.0620</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-</td>
<td>0.1032</td>
<td>0.0627</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>0.2610</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AQP5</th>
<th>ARTN</th>
<th>CKB</th>
<th>MCM3</th>
<th>TAF9</th>
<th>TGIF2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>-</td>
<td>0.0143</td>
<td>0.0056</td>
<td>0.0340</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-</td>
<td>0.0851</td>
<td>0.0903</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>0.4821</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AQP5</th>
<th>ARTN</th>
<th>CKB</th>
<th>MCM3</th>
<th>TAF9</th>
<th>TGIF2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>-</td>
<td>0.0114</td>
<td>0.0069</td>
<td>0.0780</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-</td>
<td>0.0751</td>
<td>0.0501</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>0.2738</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering: Blind test of 32 sera through pattern of protein profile.

Assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters.

<table>
<thead>
<tr>
<th>17 - cancer</th>
<th>15 - normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 - cancer</td>
<td>14 - normal</td>
</tr>
</tbody>
</table>

Sensitivity: 88%
Specificity: 80%
Accuracy: 84%

Han et al., Proteomics (2009)
Measuring the performance of a diagnostic test

- **Sensitivity**: Probability of true positives that are correctly identified by the test \( \rightarrow \frac{TP}{TP+FN} \)

- **Specificity**: Probability of true negatives that are correctly identified by the test \( \rightarrow \frac{TN}{TN+FP} \)

- **True positives** (TP): Some of people have the disease, and the test says they are positive.
- **False negatives** (FN): Some have the disease, but the test claims they don't.
- **True negatives** (TN): Some don't have the disease, and the test says they don't
- **False positives** (FP): Healthy people are diagnosed to have a positive test result
Structural Genomics

- Large-scale, systematic study of protein structures
- Provide a complete 3-D structure of proteins
  → Establish the structure-function relationships of proteins

- General procedure
  - Cloning and functional expression of target protein
  - Purification
  - Resolution of 3-D structure
    - X-ray crystallography for protein crystal
    - NMR for soluble protein (less than 20-25 kDa)

- About 30,000 protein structures: deposited in protein data bank
  - ~ 2,000 proteins are truly different
Pharmacogenetics

✓ Study on correlation of specific gene sequence information (specifically sequence variation) to drug response

✓ Personalized medicine: doctors make better-informed decisions as to what drug to prescribe to individual patients

✓ Different people respond differently to any given drug if they present with essentially identical disease symptoms
  
  ex) dose, side effect, more effective drug
  
  - Irresa is effective for lung cancer patients with primary mutation (L858R) at TK domain and deletion of exon 19.
- Differential response:
  - Not all patients respond positively to a specific drug
  - Non-genetic factor: state of health
  - Genetic variation amongst individuals remain the predominant factor

- Single nucleotide polymorphisms (SNPs):
  - All humans display almost identical genome sequences, but some differences are evident
  - The most prominent widespread-type variations among individuals
  - Occur in the general population at an average incidence of one in every 1,000 nucleotide bases
  - Entire human genome harbors about 3 million SNPs
  - SNPs occurring in structural or regulatory genes alter amino acid sequence/expression levels of a protein, affecting functional attributes (height, color etc..)
• The presence of SNPs within the regulatory or structural regions of a gene coding for a protein which interacts with a drug obviously affects the effectiveness of the drug on the body.

• Identifying and comparing SNP patterns from a group of patients responsive to a particular drug with patterns displayed by a group of un-responsive patients
  → identify SNP characteristics linked to drug efficacy
  → uncover SNP patterns associated with adverse effect

• **Personalized medicine**
  - Drug treatment can be tailored
  - Different drugs can be developed according to patient sub-type
Initial characterization of drug candidate

- Physicochemical and other properties of any new drug candidate: characterization prior to pre/clinical trials

- Major biopharmaceuticals: proteins
  - Prerequisite to characterization: purification to homogeneity based on physicochemical property (pI, size, quaternary structure, hydrophobicity)
  - High-resolution chromatographic processes
    - Gel-filtration, Ion-exchange, Hydrophobic, Affinity-chromatography
  - Careful design of purification process: critical for economics of a drug development: Purity vs Yield
  - From bench to pilot and process-scale systems
✓ **Full characterization**: Basis of the quality control (QC) test
   - Task tree for the full structural characterization of a therapeutic protein: **Figure 2.7**

✓ **Additional characteristics**
   - Stability of the protein (pH, temp, additives etc..)
     → Required in order to select a suitable final product formulation, and estimate the shelf-life