Introduction
Proteins

Major functional molecules

- Consist of linear polymers built from series of up to 20 different L-amino acids
- Play central roles in biological events through interactions with cognate proteins / ligands
  - Cell signaling processes
- Enzymes catalyze the reactions with high specificity and catalytic efficiency in cellular metabolic pathways: major components

Practical use

- Therapeutic agents
  - Many diseases caused by mutations and loss of their functions
  - Cancers, lysosomal storage disorders, Alzheimer's disease etc.
- Industrial Biotechnology: Use of enzymes for bio-based processes
- Biomaterials: Biocompatible materials
Protein science and engineering

• Protein sciences: Understand and demonstrate the structure and function of proteins/enzymes *in vivo*

• Protein engineering: design and produce novel therapeutic proteins, enzymes, materials that can benefit humans
Protein-protein interaction network

Therapeutic agents for treating diseases caused by abnormal cell signaling
- Antibodies
- Designed protein scaffold/ peptides
- Naturally occurring proteins
Cellular metabolic pathways

Enzymes
- Synthesis of valuable compounds
- Treatment of diseases: lysosomal storage disorder, cancers

View the other chart: Metabolic Pathways
Basic components

Nonpolar, aliphatic side groups
- Glycine (Gly, G)
- Alanine (Ala, A)
- Valine (Val, V)
- Leucine (Leu, L)
- Methionine (Met, M)
- Isoleucine (Ile, I)

Aromatic side groups
- Phenylalanine (Phe, F)
- Tyrosine (Tyr, Y)
- Tryptophan (Trp, W)

Positively charged side groups
- Lysine (Lys, K)
- Histidine (His, H)
- Asparagine (Asn, N)

Polar, uncharged side groups
- Serine (Ser, S)
- Threonine (Thr, T)
- Cysteine (Cys, C)

Negatively charged side groups
- Glutamate (Glu, E)
- Aspartate (Asp, D)
Peptide bond: planar and trans

Planarity results from the delocalization of the lone pair of electrons of the nitrogen onto the carbonyl oxygen.

Conformation of the polypeptide backbone:
- Location in space of the three sets of atoms that are linked together, namely the alpha carbon, carboxyl carbon, and amide nitrogen atoms.
- Secondary structure is defined by the rotation of the planar peptide units around the bonds connecting the alpha carbon atoms.
- Number of conformations available to the polypeptide chains are restricted by steric clashes.
- Conformational space in a two-dimensional plot of $\psi$ and $\phi$: Ramachandran diagram.

C-N double bond character in amide (peptide) bonds.

Planar peptide bond segments.
Protein structure

- **Primary structure**: amino acid sequence
- **Secondary structure**: Polypeptides are organized into hydrogen-bonded structures
  → regularly repeating local structures stabilized by hydrogen bonds. Alpha helix, beta sheet and turns
- **Tertiary structure**: the overall shape of a single protein molecule
- **Quaternary structure**: the structure formed by several subunits
Protein Structure

(a) Primary structure  (b) Secondary structure

β (beta) sheet

α (alpha) helix

Disulfide bond

(c) Tertiary structure

Hydrogen bond

(d) Quaternary structure

Peptide bond

Amino acid

COO⁻
3D structure of the protein myoglobin showing turquoise alpha helices. This protein was the first to have its structure solved by X-ray crystallography in 1958. Towards the right-center among the coils, a prosthetic group called a heme group (shown in gray) with a bound oxygen molecule (red).
<table>
<thead>
<tr>
<th>Experimental Method</th>
<th>Proteins</th>
<th>Nucleic acids</th>
<th>Protein/Nucleic Acid complexes</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray</td>
<td>92427</td>
<td>1645</td>
<td>4600</td>
<td>4</td>
<td>98676</td>
</tr>
<tr>
<td>NMR</td>
<td>9686</td>
<td>1124</td>
<td>227</td>
<td>8</td>
<td>11045</td>
</tr>
<tr>
<td>EM</td>
<td>584</td>
<td>29</td>
<td>191</td>
<td>0</td>
<td>804</td>
</tr>
<tr>
<td>Hybrid</td>
<td>70</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>76</td>
</tr>
<tr>
<td>Other</td>
<td>166</td>
<td>4</td>
<td>6</td>
<td>13</td>
<td>189</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>102933</strong></td>
<td><strong>2805</strong></td>
<td><strong>5026</strong></td>
<td><strong>26</strong></td>
<td><strong>110790</strong></td>
</tr>
</tbody>
</table>

August 2015
Examples of protein structures from PDB
Method to develop more useful or valuable proteins

- Alteration of a single amino acid residues at specific site
- Insertion or deletion of a single amino acid residue
- Alteration or deletion of a segment or an entire domain
- Generation of a novel fusion protein
- Incorporation of unnatural amino acids at specific site
Why Protein Engineering?

• Proteins/Enzymes: Evolved for host itself, not for human
  - Most proficient catalysts with high specificity

• Need further improvement for practical use:
  - Specificity: Cognate ligands, substrates
  - Binding affinity
  - Stability
  - Catalytic activity
  - Folding/Expression level etc.

• Goal of protein engineering:
  - Design of protein/enzyme with desired function and property for practical applications

Designer proteins/Enzymes

Ex) Therapeutic proteins, Industrial enzymes, Fluorescent proteins, Protein binders
Biomolecular Eng. Lab.

Technology Development

Random approach
- Screening from nature
- Random mutations

Structure-based rational approach
- Structure-function relationship
- Site-directed/saturation mutagenesis

Evolutionary approach
- Directed evolution
  - Accumulation of beneficial mutations
  - No structural data
  - HTS system
  - Construction of diverse library

Computational (in silico) method
- Virtual screening of large sequence space
- Large structural data: >~30,000
- High computing power
- Mechanistic knowledge

Combinatorial approach
- Structure-based design
- Evolutionary method
- Computational method
Therapeutic proteins

• Small molecule-based drugs: Efficacy, side effect, safety
• Therapeutic proteins: High efficacy and safety, less toxicity
  - Monoclonal antibodies, proteins, enzymes, peptides etc.
    ex) EPO, Interferon, Insulin, Avastin, Enbrel, Remicade, Herceptin,
      
      EPO (Erythropoietin): Stimulating the proliferation of red blood cells
      Herceptin: Mab against EGFR2 (Epidermal growth factor receptor 2)
      Avastin: Mab against VEGF (Vascular endothelial growth factor)
      Remicade: Mab against TNF-α (Tumor necrosis factor- α)

• World market
  - EPO alone: ~ $11 billion per year
  - Remicade: ~ $9 billion per year
    - Intensive investment in monoclonal antibodies: Biosimilar

Therapeutic proteins will form the back-born of future biotech market
New version of therapeutic proteins by Protein Eng

- EPO (Erythropoietin) with a longer plasma half-life by incorporation of additional N-glycosylation sites

- Human insulin
  - Faster-acting insulin by modification of amino acid sequence
  - Long-standing insulin: fusion to Fc

- Faster-acting tissue plasminogen activator (t-PA) by removal of three of the five native domains → higher clot-degrading activity

- Ontak: A fusion protein consisting of the diphtheria toxin linked to IL-2
  → Selectively kills cells expressing the IL-2 receptor
  → Approved for the treatment of cutaneous T cell lymphoma in 1999 in US

- Bi-specific monoclonal antibodies for dual targets → higher efficacy
Erythropoietin (EPO)

- **Growth factor** (166 amino acids, MW 34 kDa) produced in kidney → Promote the formation of red blood cells (erythrocytes) in the bone marrow

- Binds to the erythropoietin receptor on the red cell progenitor surface and activates a JAK2 signaling cascade

- Clinically used in treating anemia resulting from chronic kidney disease, inflammatory bowel disease (Crohn's disease and ulcer colitis), and myelodysplasia from the treatment of cancer (chemotherapy and radiation)
Structural aspect

- About 50% of EPO’s secondary structure: α-Helix
- Glycoprotein: High carbohydrate content (~40%)

- Carbohydrate moiety:
  - *in vivo* activity, stability, solubility, cellular processing and secretion, immunogenicity
  - Three N-glycosylation sites and one O-glycosylation site
Glycosylation sites in EPO
History of the EPO development

- **1971**: First purified from the plasma of anemic sheep

- **1985**: Produced by recombinant DNA technology

- **1989**: Approved by FDA for treatment of anemia resulting from chronic kidney disease and cancer treatment (chemotherapy and radiation)

- **Total sales**: $11 billion (2005)

- **Major EPO brands**: Biosimilars
  - Epogen by Amgen ($2.5 billion)
  - Procrit by Ortho Biotech ($3.5 billion)
  - Neorecormon by Boehringer-Mannheim ($1.5 billion)
Glycoproteins are proteins that contain oligosaccharide chains (glycans) covalently attached to polypeptide side-chains.

- One of the most important post-translational modifications (PTMs) are N-glycosylation / O-glycosylation in Mammalian / Yeast.

- Essential roles in *in vivo* include: Biological activity, folding, solubility, protease resistance, immunogenicity, signal transduction, and pharmacokinetics.

- Carbohydrates on cell surface: Cell signaling, cell attachment, cell adhesion, recognition, and inflammation.

- About 60% of therapeutic proteins are glycoprotein.
  - Therapeutic proteins: 140 approved.
Glycan Profile of Glycoprotein

- Glycan profile: very complex and varies broadly, depending on cell types and production conditions: Glycan moiety, occupation number, length of glycosylation chain.

- Therapeutic proteins require proper glycosylation for biological efficacy.
  → Analysis of glycan profile, its role/function in vivo, glycosylation pathway, and property of glycoproteins are a key to Glycobiology.
Monosaccharide structures

Fucose, Fuc
\[ C_6O_5H_{12} \]

Galactose, Gal
\[ C_6O_6H_{12} \]

Mannose, Man
\[ C_6O_6H_{12} \]

Sialic Acid, NAcNeuA
\[ C_{11}O_9NH_{19} \]

N-Acetylgalactosamine, GalNAc
\[ C_8O_6NH_{15} \]

N-Acetylglucosamine, GlcNAc
\[ C_8O_6NH_{15} \]
**N-Linked glycan structures**

<table>
<thead>
<tr>
<th>N-Linked Structures</th>
<th>complex sialylated fucosylated diantennary</th>
<th>complex sialylated fucosylated triantennary</th>
<th>complex sialylated fucosylated tetraantennary</th>
<th>high mannose (man 7)</th>
<th>hybrid</th>
<th>common core str. subunit</th>
<th>sialylated branch subunit</th>
<th>branch subunit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composition</td>
<td>$C_{90}O_{65}N_{6}H_{146}$</td>
<td>$C_{115}O_{83}N_{8}H_{186}$</td>
<td>$C_{140}O_{101}N_{10}H_{225}$</td>
<td>$C_{64}O_{49}N_{2}H_{106}$</td>
<td>$C_{83}O_{62}N_{4}H_{136}$</td>
<td>$C_{34}O_{25}N_{2}H_{56}$</td>
<td>$C_{25}O_{18}N_{2}H_{40}$</td>
<td>$C_{14}O_{10}N_{1}H_{23}$</td>
</tr>
</tbody>
</table>

Example of tetraantennary complex glycan that contains terminal sialic acid residues, a bisecting GlcNAc on the pentasaccharide core, and fucosylation on the core GlcNAc.
As the patent becomes expired, Amgen wanted to prolong the market share by developing a new version of EPO by protein engineering.

- **New version of EPO**: Prolonged serum half-life
  - How: Introduction of two additional N-glycosylation
  - How to design? Which sites?
  - What benefit to patients and developer?

- Other approaches?
Design of hyper-glycosylated EPO

- Relationship between sialic acid content, in vivo activity, and serum half-life
- Hypothesis: Increasing the sialic acid-containing carbohydrate of EPO → increase in serum half-life → in vivo biological efficacy
- Incorporation of additional N-glycosylation sites

The structure of the carbohydrate chains is variable

Isolate molecules based on sialic acid content
Test the isolated EPO isoforms for biological activity
Design strategy

- N-glycosylation occurs at the polypeptide backbone at a consensus sequence: Asn-Xxx-Ser/Thr (Middle amino acid can not be proline)

- Critical factors for selecting N-glycosylation sites
  - No effect on protein folding and conformation (structure/function relation)
  - No interference with receptor binding
  - Stability
Structure-based rational engineering

- Selection and mutation of the sites based on structural /functional analysis: site-directed mutagenesis
  - Ala30ASn, His32Thr, Pro87Thr, Trp88Asn, Pro90Thr

- Expression in CHO cells and analysis of the EPO variants
  - Bioactivity
  - Folding
  - Interaction with EPO receptor

- Final selection of two additional N-glycosylation sites: positions 30 and 88:
  - Prolonged serum half-life from 4-6 up to 21 hrs

- Preclinical and clinical trials of new version EPO: Aranesp
- Approval by FDA
- Start of clinical use in 2001: Current sale of $3.5 billion
Immunoglobulin Ab

- Widely used in many areas: Therapeutics, immunoassays, diagnostics, purification
  - Therapeutics: mAb
Engineering of mAb for therapeutics

- Reduced immunogenicity: Humanization, Human Ab
- Improved affinity: Engineering of variable domains (< nM)
- Antibody fragment: Fab, single-chain Fv (scFv), minibody, diabody
- Novel effector function: Conjugation with radioisotope, cytotoxic drug
  - Antibody-drug conjugates: Targeted delivery of a cytotoxic drug
- Improved effector function: Fc engineering
- Longer half-life: Fc engineering (Increase in binding affinity for FcRn)
- Bi-specific antibody: Dual-targeting mAb
Engineering of mAb for humanization

Nature Reviews | Cancer
Antibody Fragments

- Better tissue penetration, fast clearance, no effector function
- Local delivery
- Production in bacterial system
Lucentis

- A monoclonal antibody fragment (Fab) derived from the same parent mouse antibody as Avastin
- Much smaller than the parent molecule and has been affinity matured to provide stronger binding to VEGF-A
- An anti-angiogenic protein to treat the "wet" type of age-related macular degeneration (AMD, also ARMD), Common form of an age-related vision loss.
- Cost $1,593 per dose, compared to Avastin that cost $42.
- Developed by Genentech and is marketed in the United States by Genentech and elsewhere by Novartis

Intermediate age-related macular degeneration
Antibody with prolonged-half life

- Half-life of antibodies: 2~3 weeks
- pH-dependent interaction of Fc with FcRn: Histidine residues and salt bridges
- Increase in the binding affinity of Fc with FcRn at acidic pH (pH 6.0) in endosome while maintaining the binding affinity at pH 7.4
- Rational design and random mutations/selection: prolonged up to 90 days in human
Enzymes: Biocatalysts

- Most proficient catalysts with high specificity
- Competitive and cost-effective processes

Use in daily life
- Cleaning (Detergents)
- Textiles
- Starch Processing
- Leather
- Baking
- Pulp and Paper
- Food and Specialties
- Cosmetics

Use in biosciences
- DNA polymerase: Thermostability, fidelity
- Restriction enzymes: Specificity
- Alkaline phosphatase, Peroxidase

Use in drug or specialty chemicals
- Chiral drugs
- Chiral intermediates
- Semisynthetic antibiotics
- Organic acids
Enzymes play a key role in industrial biotechnology

Energy and Environmental issues
- Depletion of fossil fuels
- Limitation to CO$_2$ emission (Kyoto protocol)

Petrochemical-based economy
Chemical process

Renewable source-based economy
Bio-based process

Enzymes

Use of enzymes for biofuel and biochemicals from renewable sources such as starch and cellulose → amylase, cellulase etc.
# Therapeutic Enzymes: Enzyme replacement treatment

<table>
<thead>
<tr>
<th>Disease</th>
<th>Product name</th>
<th>Developer</th>
<th>Sales (US$Millions)</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>2004</td>
<td>2007</td>
</tr>
</tbody>
</table>
| Gaucher’s| Ceredase®    | Genzyme   | 443  | N/A     | ▪ Glucocerebrosidase (β-Glucosidase)  
          |              |           |      |         | ▪ Purified from human placenta  
|          | Cerezyme®    | Genzyme   | 932  | 1,048   | ▪ Produced in CHO cells  
          |              |           | (2005)|         | ▪ 3 Exoglucosidases process  
|          |              |           |      |         | for Terminal Mannose  
| Fabry’s  | Fabrazyme®   | Genzyme   | 209  | 397     | ▪ α-galactosidase  
|          | Replagal     | TKT       | 57   | 168     | ▪ Mannose-6-phosphate for  
|          |              |           |      |         | Glycotargeting  
| MPS-1    | Aldurazyme®  | Genzyme   | 12   | 204     | ▪ α–L-iduronidase  
| Pompe    | Myozyme®     | Genzyme   | Approved |         | ▪ α-glucosidase  
|          |              |           | (2006)|         |          |

Treatment of Gaucher’s disease by Cerezyme costs up to $550,000 annually: Orphan drug and life-long treatment

Most of therapeutic enzymes: glycoproteins
Gaucher’s Disease: Lysosomal Storage Disease

- Caused by a recessive mutation in a gene located on chromosome 1, affecting both males and females
- Most common type in LSD
- Found by Phillipe Gaucher in 1882
- Biochemical basis for the disease in 1965 by Brady et al.

Autosomal recessive inheritance

Glucocerebroside: Constituent of red and white blood cell membranes
Gaucher’s disease: Occurrence and symptoms

- 1/40,000~60,000 (Jew 1/500)
- Swollen vacuoles → Gaucher cells
- Accumulation in spleen, liver, kidney, brain
- Enlarged spleen and liver, liver malfunction, neurological complications etc..

Normal cells

- Glucocerebrosides
- Digestive vacuole
- Glucocerebrosidase: glucose + ceramide → Residual vacuole
- Exocytosis

Gaucher cells

- Glucocerebrosides
- Digestive vacuole
- Incomplete digestion
- Residual vacuole accumulated
- No exocytosis

Distended abdomen
Nucleases: Genome editing

• Nucleases or artificially engineered "molecular scissors."
  - A type of gene engineering by which DNA is inserted, replaced, or removed from a genome

• The nucleases create specific double-stranded break (DSBs) at desired locations in the genome, and harness the cell’s endogenous mechanisms to repair the induced break by natural processes of homogeneous recombination (HR) and nonhomogeneous end-joining (NHEJ).

• Four families of engineered nucleases
  - Meganucleases: Endodeoxyribonuclease with a large recognition site (double-stranded DNA sequences of 12 to 40 base pairs)
  - Zinc finger nucleases (ZFNs)
  - Transcription Activator-Like Effector Nucleases (TALENs)
  - CRISPR (Clustered regularly interspaced short palindromic repeats)/Cas system
• **Issues:**
  - Off-target effect: non-specific mutations
  - First approval in UK for human healthy embryos to alter genes active after fertilization
  - Ethical issues: Editing of human embryos
Practical application to crops

- Control of browning by editing polyphenol oxidase (PPO) using Crisper/Cas system
  - Nonbrowning potatoes and apples
  - Potential target in other crops that suffer from enzymatic browning: lettuce, cherries, avocados and bananas.
Development of enzyme process for drug intermediate

- Atorvastatin by Pfizer: Lipitor
  - The world's best-selling drug of all time, with more than US$125 B in sales over approximately 14.5 years from 1996 to 2012
  - $10.7 billion in 2015 even after launch of generic drug in 2013

- A competitive inhibitor of HMG-CoA reductase (3-hydroxy-3-methylglutaryl-CoA reductase)
  - HMG-CoA reductase catalyzes the reduction of 3-hydroxy-3-methylglutaryl-coenzymeA (HMG-CoA) to mevalonate, which is the rate-limiting step in hepatic cholesterol biosynthesis.
Cholesterol biosynthetic pathway

Acetyl CoA → HMG CoA → Mevalonate → Farnesyl pyrophosphate → Squalene → Dolichol, Cholesterol, Coenzyme Q10

Statins inhibit the conversion of HMG CoA to Mevalonate.

HMG-CoA Reductase

mevalonate
• Inhibition of the enzyme decreases *de novo* cholesterol synthesis, increasing expression of low-density lipoprotein receptors (L-receptors) on hepatocytes.
→ Increase in LDL uptake by the hepatocytes, decreasing the amount of LDL-cholesterol in the blood.

• Like other statins, atorvastatin also reduces blood levels of triglycerides and slightly increases the levels of HDL-cholesterol.

• Generic drug: Simvastatin by Merck
Pfizer fight against a simvastatin generic

• Doctors and patients began switching to a cheaper generic alternative drug called simvastatin from Merck.

• Pfizer launched a campaign including advertisements, lobbying efforts, and a paid speaking tour by Dr. Louis W. Sullivan, a former secretary of the federal Depart. of Health and Human Services, to discourage the trend.

• Studies show that at commonly prescribed doses Lipitor and simvastatin are equally effective at reducing LDL cholesterol.

• Pfizer has begun promoting a study, conducted by Pfizer’s own researchers, concluding switching increased the rate of heart attacks among British patients.
Economic process using an enzyme with higher efficiency

- **Ethyl(R)-4-cyano-3-hydroxybutyrate (NH)**
  - Starting material for the production of Atorvastatin (Lipitor)

![Chemical structure of NH](image)

- Manufacture of NH from ethyl ethyl (S) -4 –chloro-3-hydroxybutyrate(ECHB)
  - Hydroxyl group in HN is defined the second stereo-center in atorovastatin, and high chemical purity is essential for downstream chemistry

- Enzymatic process
  - Nitrilase, ADH, Aldolase
  - Halohydrin dehalogenase (HHDH): More economic process
    - Catalyze the nucleophilic displacement of a halogen by a vicinal hydroxyl group in halohydrins, yielding an epoxide
    - Interconverts halohydrins and epoxides
Halohydrin and Epoxide

• A type of organic compound or functional group in which one carbon atom has a halogen substituent, and an adjacent carbon atom has a hydroxyl substituent.

• An epoxide is a cyclic ether with three ring atoms.

General structure of a halohydrin, where $X = \text{I, Br, F, or Cl}$

Epoxide
Cyanation reaction: attachment or substitution of a cyanide group on various substrates

Scheme 3

Figure 1  HHDH interconverts halohydrins and epoxides\textsuperscript{18} and can accept alternative nucleophiles\textsuperscript{17}. Chemical cyanation is typically performed at elevated temperatures (80 °C) and pH 9. HHDH catalyzes the single-vessel enzymatic conversion of ethyl (S)-4-chloro-3-hydroxybutyrate (ECHB) to ethyl (R)-4-cyano-3-hydroxybutyrate (HN) in the presence of cyanide at pH 7 where no chemical reaction is observed. The HCl produced in the reaction causes the pH to drop, although both the pH and the cyanide concentration can be maintained by running the reaction in a pHstat with NaCN as the base.
Tetramer

Monomer

HHDH from *Agrobacterium radiobacter*
Design criteria

- **Enzyme source**
  - Expression of HHDH from *Agrobacterium radiobacter* in *E. coli*
  - Volumetric productivity: $6 \times 10^{-3}$ g product/L/hour/gram of biocatalyst

- **Requirement for developing the enzyme process at commercial scale**
  - Yield: Complete conversion (100%) of at least 100 g per liter substrate
  - Volumetric productivity: > 20 g product/liter/hour/gram of biocatalyst

Figure 2. Multivariate optimization of enzymes. At any point, ~50 mutations (variables) are evaluated in the combinatorial libraries (in the hopper). The best variants for the desired activity, and a fraction of less improved variants, are sequenced and analyzed as described. After ProSAR analysis, individual mutations are parsed into four classes: 'Beneficial', which are fixed into the population by retention in the next round parental enzyme, 'Potentially beneficial', which are sent back into the hopper for retesting, 'Deleterious', which are discarded and 'Neutral', which have little or no effect on protein function and are discarded. The amount of the diversity under investigation is maintained by addition of additional diversity discovered, for example, through rational design, homologous sequences, saturation or PCR mutagenic libraries or other evolution programs.
Increase in the volumetric productivity: ~ 4,000 fold

Development of a economically feasible process
Integration of computational analysis with experimental screening to identify 37 mutations (yellow) that increase enzymatic activity $\sim 4,000$-fold.
• Proteins/enzymes form backbone in life/medical sciences and biotechnology

• Basic knowledge on proteins and protein engineering technology are crucial for creating proteins/enzymes with greater potential