Chapter 5
Polypeptide Sequences, Analysis, and Evolution

1. Polypeptide diversity
2. Protein purification and analysis
3. Protein sequencing
4. Protein evolution
Introduction

- Protein function is central to life
- \( \Rightarrow \) understand their structure & function
- *Sequence directly protein or the corresponding DNA/cDNA*
1) Polypeptide Diversity

- **Primary structure**: sequence of amino acids
- $20^n$ possibilities, $n =$ number of residues
- $20^{100}$, more possibilities than are atoms in the universe

Range from:
- $<40$ to $>10,000$

Peptides, multi-subunit complexes
Primary structure of insulin

Note the intrachain and interchain disulfide bridges
<table>
<thead>
<tr>
<th>Protein</th>
<th>Amino Acid Residues</th>
<th>Subunits</th>
<th>Protein Molecular Mass (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteinase inhibitor III (bitter gourd)</td>
<td>30</td>
<td>1</td>
<td>3,427</td>
</tr>
<tr>
<td>Cytochrome c (human)</td>
<td>104</td>
<td>1</td>
<td>11,617</td>
</tr>
<tr>
<td>Myoglobin (horse)</td>
<td>153</td>
<td>1</td>
<td>16,951</td>
</tr>
<tr>
<td>Interferon-γ (rabbit)</td>
<td>288</td>
<td>2</td>
<td>33,842</td>
</tr>
<tr>
<td>Chorismate mutase (<em>Bacillus subtilis</em>)</td>
<td>381</td>
<td>3</td>
<td>43,551</td>
</tr>
<tr>
<td>Triose phosphate isomerase (<em>E. coli</em>)</td>
<td>510</td>
<td>2</td>
<td>53,944</td>
</tr>
<tr>
<td>Hemoglobin (human)</td>
<td>574</td>
<td>4</td>
<td>61,986</td>
</tr>
<tr>
<td>RNA polymerase (bacteriophage T7)</td>
<td>883</td>
<td>1</td>
<td>98,885</td>
</tr>
<tr>
<td>Nucleoside diphosphate kinase (<em>Dictyostelium discoideum</em>)</td>
<td>930</td>
<td>6</td>
<td>100,764</td>
</tr>
<tr>
<td>Pyruvate decarboxylase (yeast)</td>
<td>2,252</td>
<td>4</td>
<td>245,456</td>
</tr>
<tr>
<td>Glutamine synthetase (<em>E. coli</em>)</td>
<td>5,616</td>
<td>12</td>
<td>621,264</td>
</tr>
<tr>
<td>Titin (human)</td>
<td>34,350</td>
<td>1</td>
<td>3,816,188</td>
</tr>
</tbody>
</table>
Constrains of protein structure

1. Size: minimal for function, max for synthesis

2. Not all amino acids are equally frequent in proteins
   a) Most abundant Leu, Ala, Gly, Ser, Val, Glu
   b) Rarest: Trp, Cys, Met, His

Hydrophobic residues cluster inside
Hydrophilic residues are exposed to water, *ie* on surface of protein

Characteristic is AA sequence rather than AA composition
2) Protein Purification and Analysis

Purify: enrich something that makes up <0.1% of a tissue to ~98% purity, ie 1000x

What does this mean: 1000x?

A) Purifying a protein requires a strategy
- Earlier: purify abundant proteins, ie haemoglobin
- Today: express gene in E. coli,
  -> purify recombinant protein
  Can constitute up to 40% of host cell protein

Often in inclusion bodies
If the protein is outside its natural environment, pH, temperature etc. become important

Protein environment must be controlled
Important for protein stability
1. pH -> buffers
2. Temperature -> the lower, the more stable, 0°C
3. Degradative enzymes, proteases, nucleases, should be inactivated, inhibited
4. Surface adsorption, denaturation at air-water surface and at plastic surface
5. Long-term storage, prevent microbial growth, oxidation etc, -80°C, or -196°C (fl N₂) under argon
Quantification of proteins

To purify a protein you need to be able to “see” it somehow, to follow its enrichment during different purification steps

Assay:
- monitor catalytic activity of your protein
- use of an antibody against your protein (ELISA)
ELISA

test for detecting 
Chorion Gonadotropin in urine

1. Immobilize first antibody on solid support
2. Incubate with protein-containing sample
3. Add a second antibody that is covalently linked to an assayable enzyme
4. Wash and assay the enzyme

Bsp. Schwangerschaftstest basierend auf ELISA detektieren Chorion Gonadotropin im Urin

Animated Figure
Determination of protein concentration by spectroscopy

Determination of the concentration of a solute that absorbs light

By Beer-Lambert law: \( A = \log \left( \frac{I_0}{I} \right) = \varepsilon cl \)

- \( A \), absolute absorbance (optical density)
- \( I_0 \), intensity of light at given \( \lambda \)
- \( I \), transmitted intensity
- \( \varepsilon \), extinction coefficient
- \( c \), concentration
- \( l \), length of light path

Polypeptides absorb strongly in the 200-400 nm region (UV), due to aromatic side chains (Trp, Phe, Tyr) but not in the visible light region (400-800 nm) colorless unless it has a chromophore, such as cytochromes
UV absorbance spectra

\( \varepsilon \) vs. \( \lambda \) (nm)

- Trp
- Tyr
- Phe

Absorbance values range from 0 to 40,000.
Determination of Protein concentration

Monitor absorbance at 280 nm to follow total protein concentration (But DNA and RNA also absorb at this wavelength)

More specific Assays:
Bradford Assay: Use of *Coomassie brilliant blue*
shifts absorbance from 465 to 595 when bound to protein
sensitive: 1µg of protein/ml
- Will also thightly stain your fingers ...
Coomassie brilliant blue

R250: $R = H$
G250: $R = CH_3$
Protein purification is a stepwise process

Protein purification by fractionation:
Series of independent steps: eliminate everybody else

Protein purification is Art & Science

Principle: Use the unique physicochemical properties of your protein to separate it from all the others
<table>
<thead>
<tr>
<th>Protein Characteristic</th>
<th>Purification Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solubility</td>
<td>Salting out</td>
</tr>
<tr>
<td>Ionic Charge</td>
<td>Ion exchange chromatography</td>
</tr>
<tr>
<td></td>
<td>Electrophoresis</td>
</tr>
<tr>
<td></td>
<td>Isoelectric focusing</td>
</tr>
<tr>
<td>Polarity</td>
<td>Hydrophobic interaction chromatography</td>
</tr>
<tr>
<td>Size</td>
<td>Gel filtration chromatography</td>
</tr>
<tr>
<td></td>
<td>SDS-PAGE</td>
</tr>
<tr>
<td>Binding Specificity</td>
<td>Affinity chromatography</td>
</tr>
</tbody>
</table>

© 2008 John Wiley & Sons, Inc. All rights reserved.
B) Salting out separates proteins by their solubility

- What does solubility mean?

- Salting out, remove protein from the solution by centrifugation/precipitation

- Ammoniumsulfate, \((\text{NH}_4)_2\text{SO}_4\), soluble to 3.9M in water

- Every Protein is least soluble at its isoelectric point, pI (has no netto charge)
Salting out

(a) Target protein

(b) Supernatant

(c) Precipitate
### Table 5-2: Isoelectric Points of Several Common Proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>pI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pepsin</td>
<td>&lt;1.0</td>
</tr>
<tr>
<td>Ovalbumin (hen)</td>
<td>4.6</td>
</tr>
<tr>
<td>Serum albumin (human)</td>
<td>4.9</td>
</tr>
<tr>
<td>Tropomyosin</td>
<td>5.1</td>
</tr>
<tr>
<td>Insulin (bovine)</td>
<td>5.4</td>
</tr>
<tr>
<td>Fibrinogen (human)</td>
<td>5.8</td>
</tr>
<tr>
<td>γ-Globulin (human)</td>
<td>6.6</td>
</tr>
<tr>
<td>Collagen</td>
<td>6.6</td>
</tr>
<tr>
<td>Myoglobin (horse)</td>
<td>7.0</td>
</tr>
<tr>
<td>Hemoglobin (human)</td>
<td>7.1</td>
</tr>
<tr>
<td>Ribonuclease A (bovine)</td>
<td>9.4</td>
</tr>
<tr>
<td>Cytochrome c (horse)</td>
<td>10.6</td>
</tr>
<tr>
<td>Histone (bovine)</td>
<td>10.8</td>
</tr>
<tr>
<td>Lysozyme (hen)</td>
<td>11.0</td>
</tr>
<tr>
<td>Salmine (salmon)</td>
<td>12.1</td>
</tr>
</tbody>
</table>
C) Chromatography

- Chromatography: Gr. *chroma*, color + *graphein*, to write
- 1903, separation of plant pigments
- Principle: Mobile phase - stationary phase
  - Interaction of the analyte between stationary (matrix) and mobile phases
- Classification according to the type of interaction, hydrophobic, ionic, affinity, etc
- On paper, beads in columns
- High-performance liquid chromatography, HPLC, up to 5000 psi
Ion Exchange Chromatography separates anions and cations

- Stationary phase carries ionic groups, linked to cellulose or agarose
  \[ R^+A^- + B^- \leftrightarrow R^+B^- + A^- \] (anion exchanger, R stationary)

- Proteins are poly-electrolytes and can be separated on anion, or cation exchanger

- Strength of interaction depends on pH and salt concentration of buffer

- Steps: 1. Binding of proteins under given conditions A
  2. Selective elution of protein under condition B
Examples of ion exchanger

- anion exchanger: DEAE, diethylaminoethyl
- cation exchanger: CM, caboxymethyl

Proteins are polyelectrolytes that can bind to both anion and cation exchangers, and their binding affinities depends on the pH of the buffer and ist salt content.
Ion exchange chromatography

- Eluant, elution buffer, reduces affinity between protein and stationary phase, increases affinity of protein to buffer (eluant)

- Proteins with high affinity for the matrix elute slow, those with low affinity elute fast
Hydrophobic interaction chromatography

- To purify non-polar substances, i.e. lipids, hydrophobic proteins (membrane proteins)

- Matrix is substituted with octyl- or phenyl-groups

- Hyrophobic interaction is increased by salt in the buffer

- Eluant, low salt, may contain detergents to disrupt hydrophobic interaction between protein and matrix
Gel-filtration chromatography

- Gel-Filtration = Size Exclusion Chromatography = Molecular sieve, separation according to size and shape

- Stationary phase, gel beads with small pores

- Pore size of beads is determined by the manufacturer by the degree of crosslinking

- If protein too big for entry into pores -> fast elution
Animation gel filtration
Affinity chromatography

- Exploits specific binding properties of the protein
- Ligand is coupled to matrix via spacer
- Immuno-affinity chromatography, attach antibody to matrix
- Metal chelate affinity chromatography, Zn\(^{2+}\) or Ni\(^{2+}\) are attached to matrix, bind to poly-His tags (6xHis)
D) Electrophoresis

- Migration of ions in an electric field

- Polyacrylamide gel electrophoresis, PAGE (chapter 3-4B)

- At pH~9 -> most proteins have net negative charge and thus move to anode

- Basis of separation: gel filtration (size and shape) and electrophoretic mobility (charge)

- Visualization of proteins, coomassie blue, autoradiography,
**SDS-PAGE**

- Separates proteins by mass
- SDS denatures proteins, confers negative charge
  
  \[ \text{[CH}_3\text{—(CH}_2\text{)}_{10}\text{—CH}_2\text{—O—SO}_3\text{]Na}^+ \]
  Sodium dodecyl sulfate (SDS)
- Separation according to mass, like gel filtration
Logarithmic relationship between the molecular mass of a protein and its electrophoretic mobility in SDS-PAGE

- SDS-PAGE can be used to determine the molecular mass of an unknown protein.
- Preparing samples in presence or absence of 2-mercaptoethanol allows to determine whether subunits are linked by disulfide bridges.
Two-dimensional electrophoresis

- To resolve complex protein mixtures

- First dimension: isoelectric focusing (IEF), separation of proteins according to their $pI$ in a stable pH gradient

- Second dimension: SDS-PAGE (separation ac. To molecular mass)

- Can resolve up to 5'000 protein spots, i.e. *E.coli* or *yeast* proteome
Zwei-dimensionale Gelelektrophorese

- Isoelektrische Fokusierung (IEF) gepaart mit SDS-PAGE
3) Protein Sequencing

Why?
1) Sequence -> understand function, determine structure of protein
2) Sequence comparison of related proteins
3) Diagnostic test for inherited diseases caused by mutations in a given protein

- 1953, first sequence of bovine insulin, took 10 years, 100g protein
- Principle: Need a pure sample of your protein, break down protein into pieces that can be sequenced by a chemical method

See guided exploration 4
Overview of protein sequencing

Protein (two different polypeptide chains linked by disulfide bonds)

1. Reduce disulfide bonds
   Separate the chains

2a. Use chemical or enzymatic methods to break each polypeptide into smaller peptides

2b. Use different methods to generate a different set of peptide fragments

3a. Determine the sequence of each peptide fragment

3b. Determine the sequence of each peptide fragment

4. Use the two sets of overlapping peptide sequences to reconstruct the sequence of each polypeptide

5. Repeat fragmentation without breaking disulfide bonds to identify the Cys-containing sequences involved in the disulfide linkages
A) First step: separate subunits

- How many subunits is my protein composed of?

- N-terminal amino acid analysis reveals the number of different types of subunits:
  - Modify N-terminal AA with dansyl chloride, hydrolyze all peptide bonds, determine the nature of the dansyl-modified AA
  - Or performe 1 cycle of Edman degradation
  - If I get Gly and Phe, I know that my protein has at least two nonidentical subunits
5-Dimethylamino-1-naphthalenesulfonyl chloride (dansyl chloride) + Polypeptide → Dansyl polypeptide

Dansyl polypeptide → Dansylamino acid (fluorescent) + Free amino acids
Disulfide bonds between and within polypeptides are cleaved by 2-mercaptoethanol

Cystine + 2H_{2}SH \rightarrow 2 \text{Cysteine} + 2 \text{HSCH}_{2}CH_{2}OH
Alkylierung der Cysteine mit Iodessigsäure

\[
\text{Cys} \quad \text{CH}_2 \quad \text{SH} \quad + \quad \text{ICH}_2\text{COO}^- \quad \rightarrow \quad \text{Cys} \quad \text{CH}_2 \quad \text{S} \quad \text{CH}_2\text{COO}^- \quad + \quad \text{HI}
\]

S-Carboxymethylcysteine
B) The polypeptide chains are then cleaved

- Sequence determination has a size-limit of about 40 AA

- Peptides longer than 40 to 100 AA need to be shortened: cleaved into pieces, fragments that can be subject to chemical sequencing

- Chemical or Enzymatic cleavage of proteins by: endopeptidases, exopeptidases, CnBr

- Trypsin as example for an endopeptidase that cleaves after Arg or Lys
Typsin cleavage

\[ \begin{align*}
\text{NH}_3^+ \quad &\text{Lys (or Arg)} \\
\text{CH}_2 \quad &\text{NH} \quad \text{CH} \quad \text{C} \quad \text{O} \\
\text{CH}_2 \quad &\text{NH} \quad \text{CH} \quad \text{C} \quad \text{O} \\
\text{CH}_2 \quad &\text{NH} \quad \text{CH} \quad \text{C} \quad \text{O} \\
\text{CH}_2 \quad &\text{NH} \quad \text{CH} \quad \text{C} \quad \text{O} \\
\text{H}_2\text{O} \quad &\text{trypsin} \\
\text{NH}_3^+ \quad &\text{NH} \quad \text{CH} \quad \text{C} \quad \text{O}^- \\
\text{CH}_2 \quad &\text{NH} \quad \text{CH} \quad \text{C} \quad \text{O}^- \\
\text{CH}_2 \quad &\text{NH} \quad \text{CH} \quad \text{C} \quad \text{O}^- \
\text{H}_3\text{N} \quad &\text{NH} \quad \text{CH} \quad \text{C} \quad \text{O}^- \\
\text{© 2008 John Wiley & Sons, Inc. All rights reserved.}
\end{align*} \]
<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Source</th>
<th>Specificity</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trypsin</td>
<td>Bovine pancreas</td>
<td>$R_{n-1} = \text{positively charged residues: Arg, Lys; } R_n \neq \text{ Pro}$</td>
<td>Highly specific</td>
</tr>
<tr>
<td>Chymotrypsin</td>
<td>Bovine pancreas</td>
<td>$R_{n-1} = \text{bulky hydrophobic residues: Phe, Trp, Tyr; } R_n \neq \text{ Pro}$</td>
<td>Cleaves more slowly for $R_{n-1} = \text{Asn, His, Met, Leu}$</td>
</tr>
<tr>
<td>Elastase</td>
<td>Bovine pancreas</td>
<td>$R_{n-1} = \text{small neutral residues: Ala, Gly, Ser, Val; } R_n \neq \text{ Pro}$</td>
<td></td>
</tr>
<tr>
<td>Thermolysin</td>
<td><em>Bacillus thermoproteolyticus</em></td>
<td>$R_n = \text{Ile, Met, Phe, Trp, Tyr, Val; } R_{n-1} \neq \text{ Pro}$</td>
<td>Occasionally cleaves at $R_n = \text{Ala, Asp, His, Thr; heat stable}$</td>
</tr>
<tr>
<td>Pepsin</td>
<td>Bovine gastric mucosa</td>
<td>$R_n = \text{Leu, Phe, Trp, Tyr; } R_{n-1} \neq \text{ Pro}$</td>
<td>Also others; quite nonspecific; pH optimum $= 2$</td>
</tr>
<tr>
<td>Endopeptidase V8</td>
<td><em>Staphylococcus aureus</em></td>
<td>$R_{n-1} = \text{Glu}$</td>
<td></td>
</tr>
</tbody>
</table>

© 2008 John Wiley & Sons, Inc. All rights reserved.
Cyanogen bromide (CNBr) cleavage

Chemical cleavage after Met

Cyanogen bromide

Peptidyl homoserine lactone
C) Edman degradation removes a peptide’s first amino acid residue

- Stepwise remove one amino acid after the other in a cyclic chemical process from the N-terminus of a purified peptide: 1 = Ala, 2 = Gly, 3 = Phe (sequence)
- Edmans reagents: **PITC**, phenylisothiocyanate => phenylthiocarbamyl adduct, PTC
- Hydrolyze with anhydrous trifluoroacetic acid => thiazolinone derivative
- Determine nature of phenylthiohydantoin, PTH-amino acid after each cycle
- Automated process, up to 100 Aa in a run
Phenylisothiocyanate (PITC)

Polypeptide

PTC polypeptide

Thiazolinone derivative

Original polypeptide less its N-terminal residue

- identify after each cycle

PTH-amino acid

© 2008 John Wiley & Sons, Inc. All rights reserved.
D) Mass spectrometry determines the molecular masses of peptides

- Mass spec measures mass-to-charge ration ($m/z$) of ions in the gase phase

- + charge due to Lys or Arg

- Ile = leu

- Gln-Lys, $\Delta=0.036\text{Da}$
Electrospray ionization mass spectrometry (ESI-MS)
apomyoglobin
Sample calculation

\[ p_1 = \frac{(M + z)}{z} \]
\[ p_2 = \frac{(M + z - 1)}{(z - 1)} \]

These two linear equations can readily be solved for their unknowns, M and z. Solve the first equation for M.

\[ M = z(p_1 - 1) \]

Then plug this result into the second equation.

\[ p_2 = \frac{z(p_1 - 1) + z - 1}{z - 1} = \frac{zp_1 - 1}{z - 1} \]
\[ zp_2 - p_2 = zp_1 - 1 \]
\[ z = \frac{(p_2 - 1)}{(p_2 - p_1)} \]
\[ M = \frac{(p_2 - 1)}{(p_1 - 1)} \frac{(p_2 - p_1)}{p_2 - p_1} \]

Plugging in the values for \( p_1 \) and \( p_2 \),
\[ M = \frac{(1542.3 - 1)(1414.0 - 1)}{(1542.3 - 1414.0)} \]
\[ = 16,975 \text{ D} \]
Tandem mass spectrometry in peptide sequencing

Up to 25 Aa
Reconstructed protein sequences are stored in databases

- Sequencing of overlapping peptides for reconstruction of the entire protein sequence
- Determine position of disulfide bridges
- Record sequence in public database, UniProt, Swissprot
Generating overlapping fragments to determine the amino acid sequence of a polypeptide
Determining the positions of disulfide bonds

Polypeptide fragment containing disulfide bond

Reduce disulfide and block with iodoacetate

Separate and sequence the polypeptides
Table 5-4  Internet Addresses for the Major Protein and DNA Sequence Data Banks

**Data Banks Containing Protein Sequences**
- Protein Information Resource (PIR): [http://pir.georgetown.edu/](http://pir.georgetown.edu/)
- Protein Research Foundation (PRF): [http://www4.prf.or.jp/](http://www4.prf.or.jp/)
- UniProt: [http://www.ebi.uniprot.org/](http://www.ebi.uniprot.org/)

**Data Banks Containing Gene Sequences**
- European Bioinformatics Institute (EBI): [http://www.ebi.ac.uk](http://www.ebi.ac.uk)
<table>
<thead>
<tr>
<th>ENTRY INFORMATION</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTRY NAME</td>
<td>RSN HUMAN</td>
</tr>
<tr>
<td>ACCESSION NUMBERS</td>
<td>Q9HD89; Q540D9</td>
</tr>
<tr>
<td>Integrated into Swiss-Prot on</td>
<td>2001-09-26</td>
</tr>
<tr>
<td>Sequence was last modified on</td>
<td>2001-03-01 (Sequence version 1)</td>
</tr>
<tr>
<td>Annotations were last modified on</td>
<td>2006-10-03 (Entry version 42)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NAME AND ORIGIN OF THE PROTEIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROTEIN NAME</td>
</tr>
<tr>
<td>Synonyms</td>
</tr>
<tr>
<td>Cysteine-rich secreted protein FIZZ3</td>
</tr>
<tr>
<td>Adipose tissue-specific secretory factor</td>
</tr>
<tr>
<td>ADSF</td>
</tr>
<tr>
<td>C/EBP-epsilon-regulated myeloid-specific secreted cysteine-rich protein</td>
</tr>
<tr>
<td>Cysteine-rich secreted protein A 12-alpha-like 2</td>
</tr>
<tr>
<td>GENE NAME</td>
</tr>
<tr>
<td>Synonym: FIZZ3; HXCP1; RSTN</td>
</tr>
<tr>
<td>ORF name: UNQ407/PRO1199</td>
</tr>
<tr>
<td>SOURCE ORGANISM</td>
</tr>
<tr>
<td>TAXONOMY ID</td>
</tr>
<tr>
<td>LINEAGE</td>
</tr>
</tbody>
</table>
4) Protein Evolution

• Proteins and genomes evolve through mutation and selection (Darwin)

• Random mutations in genome, can be lost or maintained if they bring an evolutionary advantage, that is through natural selection

• Lethal mutations die out (if homozygous)
A) Protein sequences reveal evolutionary relationships

• Example, cytochrome c, of mitochondrial electron transport chain

• Aerobic respiration 1.5 to 2 Mia years ago

• Compare cytochrome c sequences of more than 100 species

• What do you learn from these similarities and differences?
### Table 5-5 Amino Acid Sequences of Cytochromes c from 38 Species

<table>
<thead>
<tr>
<th>Species</th>
<th>Amino Acid Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human, chimpanzee</td>
<td>GDVEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Rhesus monkey</td>
<td>GDVEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Horse</td>
<td>GDVEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Donkey</td>
<td>GDVEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Cow, pig, sheep</td>
<td>GDVEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Dog</td>
<td>GDVEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Rabbit</td>
<td>GDVEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>California gray whale</td>
<td>GDVEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Great gray kangaroo</td>
<td>GDVEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Chicken, turkey</td>
<td>GDIEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Pigeon</td>
<td>GDIEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Pekin duck</td>
<td>GDIEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Snapping turtle</td>
<td>GDIEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Rattlesnake</td>
<td>GDIEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Bullfrog</td>
<td>GDIEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Tuna</td>
<td>GDIEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Dogfish</td>
<td>GDIEKGGKFGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Samia cynthia (a moth)</td>
<td>VPAGNAEGKGGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Tobacco hornworm moth</td>
<td>VPAGNAEGKGGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Screwworm fly</td>
<td>VPAGNAEGKGGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Drosophila (fruit fly)</td>
<td>VPAGNAEGKGGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Baker's yeast</td>
<td>PAPFEQGSSAKGGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Candida krusei (a yeast)</td>
<td>PAPFEQGSSAKGGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Neurospora crassa (a mold)</td>
<td>PAPFEQGSSAKGGKMKKCTVBEKGGKMKKTPNLHGLFGRTGQA</td>
</tr>
<tr>
<td>Wheat germ</td>
<td>ASFESAPGNPDAGAKIFKTKCACTVBEKGGKMKKTPNLHGLFGRTGSA</td>
</tr>
<tr>
<td>Buckwheat seed</td>
<td>ASFESAPGNPDAGAKIFKTKCACTVBEKGGKMKKTPNLHGLFGRTGSA</td>
</tr>
<tr>
<td>Sunflower seed</td>
<td>ASFESAPGNPDAGAKIFKTKCACTVBEKGGKMKKTPNLHGLFGRTGSA</td>
</tr>
<tr>
<td>Mung bean</td>
<td>ASFESAPGNPDAGAKIFKTKCACTVBEKGGKMKKTPNLHGLFGRTGSA</td>
</tr>
<tr>
<td>Cauliflower</td>
<td>ASFESAPGNPDAGAKIFKTKCACTVBEKGGKMKKTPNLHGLFGRTGSA</td>
</tr>
<tr>
<td>Pumpkin</td>
<td>ASFESAPGNPDAGAKIFKTKCACTVBEKGGKMKKTPNLHGLFGRTGSA</td>
</tr>
<tr>
<td>Sesame seed</td>
<td>ASFESAPGNPDAGAKIFKTKCACTVBEKGGKMKKTPNLHGLFGRTGSA</td>
</tr>
<tr>
<td>Castor bean</td>
<td>ASFESAPGNPDAGAKIFKTKCACTVBEKGGKMKKTPNLHGLFGRTGSA</td>
</tr>
<tr>
<td>Cottonseed</td>
<td>ASFESAPGNPDAGAKIFKTKCACTVBEKGGKMKKTPNLHGLFGRTGSA</td>
</tr>
<tr>
<td>Abutilon seed</td>
<td>ASFESAPGNPDAGAKIFKTKCACTVBEKGGKMKKTPNLHGLFGRTGSA</td>
</tr>
</tbody>
</table>

*The amino acid side chains have been shaded according to their polarity characteristics so that an invariant or conservatively substituted residue is identified by a vertical band of a single color. The letter a at the beginning of the chain indicates that the N-terminal amino group is acetylated; an h indicates that the acetyl group is absent.


Irving Geis/Geis Archives Trust, Copyright Howard Hughes Medical Institute. Reproduced with permission.
Table 5-5 (continued)

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>G</td>
<td>Y</td>
<td>S</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>K</td>
<td>I</td>
<td>G</td>
<td>L</td>
</tr>
<tr>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>Y</td>
<td>T</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>K</td>
<td>I</td>
</tr>
<tr>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>V</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>E</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>E</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>E</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>E</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>E</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>E</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>Q</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>Q</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
<td>G</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
<tr>
<td>H</td>
<td>P</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>D</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>G</td>
<td>I</td>
</tr>
</tbody>
</table>

Hydrophilic, acidic: D Asp  E Glu
Hydrophilic, basic: H His  K Lys  R Arg  X TrimethylLys
Polar, uncharged: B Asn or Asp  G Gly  N Asn  Q Gln
                  S Ser  T Thr  W Trp  Y Tyr  Z Gln or Glu
Hydrophobic: A Ala  C Cys  F Phe  I Ile  L Leu
              M Met  P Pro  V Val

Irving Geis/Geis Archives Trust, Copyright Howard Hughes Medical Institute. Reproduced with permission.
Sequence comparisons provide information on protein structure and function

In case of cytochrome c:

- 38 positions are invariant (fully conserved)
- Remaining positions are occupied by chemically similar Aa (conservative substitution)
- In 8 positions we find 6 or more different Aa (hypervariable residues)
- Neutral drift, change of Aa over time, these residues are not critical for the function of the protein
Phylogenetic trees depict the evolutionary history

- Comparison of homologous proteins:
  - Which Aa are important for function (conserved)

- Quantify difference, count number of different Aa
  - Normalize to differences in 100 Aa

- Note that evolutionary distance of modern forms to the earliest common forms are equal, the earliest forms thus continued to evolve and did not stand still....

- Conservation of protein domains

See bioinformatics exercise 5-1 and 5-2
Phylogenetic trees of cytochrome c
B) Proteins evolve by the duplication of genes or gene segments

- Proteins with similar function have similar sequence, orthologs
- Probably evolved from a common ancestor
- Gene duplication (genome duplication), 98% of proteins in human have paralogs, i.e. are present in at least two copies/haploid genome
- Example, globin family, $\alpha$-, $\beta$-subunit, similar to myoglobin, fetal forms, pseudogenes
Genealogy of the globin family

α, ι, γ, ε, δ, β  Myoglobin

Primordial globin
The rate of sequence divergence varies between different proteins

- Blot rate at which mutations are fixed in the proteins structure against time
- Functional constraints in protein sequence, i.e. Histone H4, cannot further evolve
- Structural constrains (binding partners)
- Folding constraints
Rates of evolution of four proteins

Number of amino acid changes/100 sites

- Fibrinopeptides
- Hemoglobin
- Cytochrome c
- Histone H4

Millions of years since divergence

0 200 400 600 800 1000 1200 1400
Many proteins contain domains that occur in other proteins

- Many proteins are composed of mosaics of protein domains that are 40-100 Aa long
- These domains are also found in other proteins, or are repeated many times in a given protein
- Functional, structural and folding modules
- Evolution by domain shuffling
Construction of some multidomain proteins

(a) Fibronectin

(b) Blood clotting proteins

Factors VII, IX, X, and protein C
Factor XII
Tissue-type plasminogen activator
Protein S

Key

△ Fibronectin domain 1
□ Fibronectin domain 2
○ Fibronectin domain 3
■ γ-Carboxyglutamate domain
◊ Epidermal growth factor domain
■ Serine protease domain
■■ Kringle domain
■■■ Unique domain