Proteins: Primary Structure
Lecture 6
Chapters 4 & 5
9/10/09

Proteins are the Body's Worker Molecules

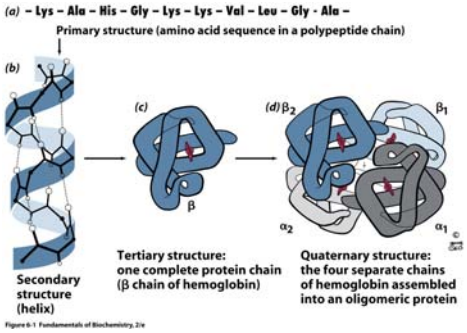
Figure from The Structures of Life (NIH)

Protein function as:

1. **Enzymes**:biological catalysts
2. **Regulators** of catalysis - hormones
3. **Transport** and store: O_2, metal ions, sugars, lipids, etc.
4. **Contractile assemblies**: Muscle fibers
   Separation of chromosomes
5. **Sensory**: Rhodopsin nerve proteins
6. **Cellular defense**
   immuoglobulins
   Antibodies
   Killer T cell
   Receptors
7. **Structural**
   Collagen
   Silk, etc.

Function is dictated by protein structure!!

Structural Hierarchy in proteins



There are four levels of protein structure

1. Primary structure
   1° = Amino acid sequence, the linear order of AA's.
   Remember from the N-terminus to the C-terminus
   Above all else this dictates the structure and function of the protein.

2. Secondary structure
   2° = Local spatial alignment of amino acids without regard to side chains.
   Usually repeated structures

   Examples: α helix, β sheets, random coil, or β turns

3. Tertiary Structure
   3° = the 3 dimensional structure of an entire peptide.
   Great in detail but vague to generalize. Can reveal the detailed chemical mechanisms of an enzyme.

4. Quaternary Structure
   4° two or more peptide chains associated with a protein.

   Spatial arrangements of subunits.

Table 5-1 Compositions of Some Proteins

Protein	Amino Acid Residues	Subunits	Polypeptide Molecular Mass (*D*)
Proteinase inhibitor III (bitter gourd)	30	1	3,427
Cytochrome *c* (human)	104	1	11,617
Myoglobin (horse)	153	1	16,951
Interferon-γ (rabbit)	288	2	33,842
Chorismate mutase (*Bacillus subtilis*)	381	3	43,551
Triose phosphate isomerase (*E. coli*)	510	2	53,944
Hemoglobin (human)	574	4	61,986
RNA polymerase (bacteriophage T7)	883	1	98,885
Nucleoside diphosphate kinase (*Dictyostelium discoideum*)	930	6	100,764
Pyruvate decarboxylase (yeast)	2,252	4	245,456
Glutamine synthetase (*E. coli*)	5,616	12	621,264
Titin (human)	26,926	1	2,993,428

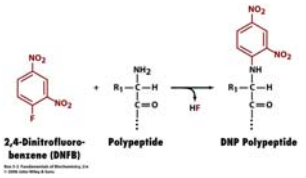
Table 5-1 Fundamentals of Biochemistry, 2/e
© 2006 John Wiley & Sons

Insulin was the first protein to be sequenced

F. Sanger won the Nobel prize for protein sequencing.

It took 10 years, many people,
and it took 100 g of protein!

Today it takes one person several days to sequence the same insulin.



Chapter 5.3 is how to determine a protein's primary structure.

"**Protein Chemistry**"

Steps towards protein sequencing

Above all else, purify it first!! Chapter 5.3 then 5.1 and 5.2

1. Prepare protein for sequencing

a. Determine number of chemically different polypeptides.
b. Cleave the protein's disulfide bonds.
c. Separate and purify each subunit.
d. Determine amino acid composition for each peptide.

Bovine insulin:
note the intra- and inter- chain disulfide linkages

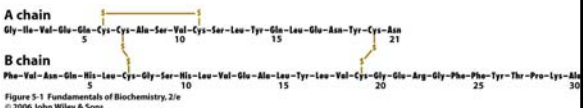


Figure 5-1 Fundamentals of Biochemistry, 2/e
© 2006 John Wiley & Sons

2. Sequencing the peptide chains:

   a. Fragment subunits into smaller peptides ≈ 50 AA's in length.

   b. Separate and purify the fragments

   c. Determine the sequence of each fragment.

   d. Repeat step 2 with different fragmentation system.

3. Organize the completed structure.

a. Span cleavage points between sets of peptides determined by each peptide sequence.

b. Elucidate disulfide bonds and modified amino acids.

At best, the automated instruments can sequence about 50 amino acids in one run!

Proteins must be cleaved into smaller pieces to obtain a complete sequence.

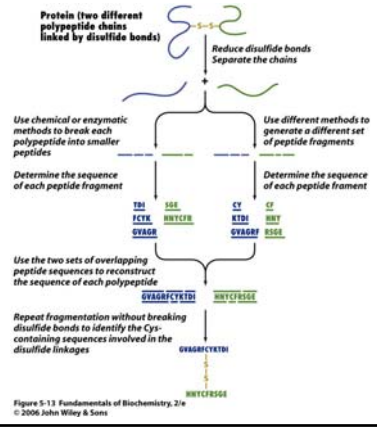


Figure 5-13 Fundamentals of Biochemistry, 2/e
© 2006 John Wiley & Sons

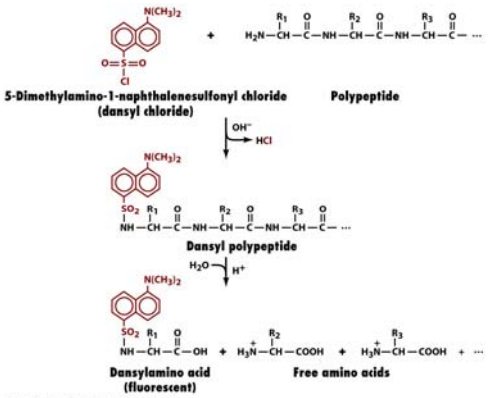
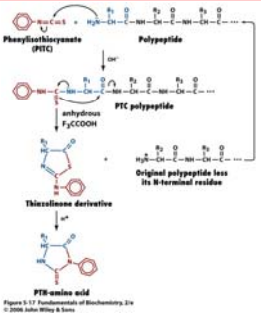


Figure 5-14 Fundamentals of Biochemistry, 2/e
© 2006 John Wiley & Sons

Disadvantage with the Dansyl-chloride method is that you must use 6M HCl to cleave off the derivatized amino acid, this also cleaves all other amide bonds (residues) as well.

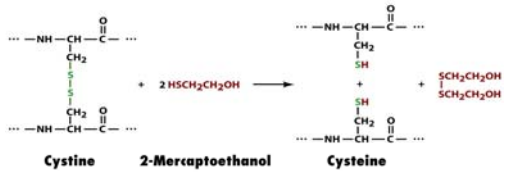
Edman degradation with Phenylisothiocyanate, PITC

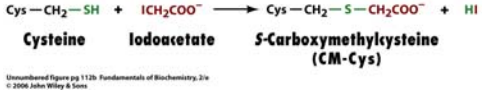


Edman degradation has been automated as a method to sequence proteins. The PTH-amino acid is soluble in solvents that the protein is not. This fact is used to separate the tagged amino acid from the remaining protein, allowing the cycle of labeling, degradation, and separation to continue.

Even with the best chemistry, the reaction is about 98% efficient. After sufficient cycles more than one amino acid is identified, making the sequence determination error-prone at longer reads.

Cleavage of disulfide bonds





Unnumbered figure pg 112b Fundamentals of Biochemistry, 2/e
© 2006 John Wiley & Sons

Amino acid composition

The amino acid composition of a peptide chain is determined by its complete hydrolysis followed by the quantitative analysis of the liberated amino acids.

Acid hydrolysis (6 N HCl) at 120 °C for 10 to 100 h

destroys Trp and partially destroys Ser, Thr, and Tyr.

Also

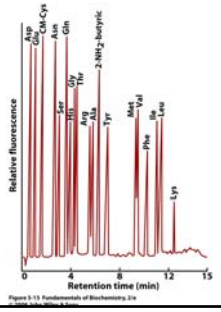
Gln and Asn yield Glu and Asp

Base hydrolysis 2 to 4 N NaOH at 100 °C for 4 - 8 h. Is problematic, destroys Cys Ser, Thr, Arg but does not harm Trp.

Amino acid analyzer

In order to quantitate the amino acid residues after hydrolysis, each must be derivatized at about 100% efficiency to a compound that is colored. Pre or post column derivatization can be done.

These can be separated using HPLC in an automated setup



Amino acid compositions are indicative of protein structures

Leu, Ala,Gly, Ser, Val, Glu, and Ile are the most common amino acids.

His, Met, Cys, and Trp are the least common.

Ratios of polar to non-polar amino acids are indicative of globular or membrane proteins.

Certain structural proteins are made of repeating peptide structures i.e. collagen.

Long peptides have to be broken to shorter ones to be sequenced

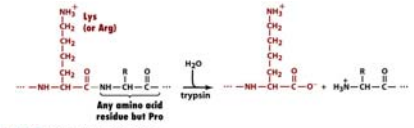


Table 5-3 Specificities of Various Endopeptidases

Enzyme	Source	Specificity	Comments
Trypsin	Bovine pancreas	R_{n-1} = positively charged residues: Arg, Lys; $R_n \neq$ Pro	Highly specific
Chymotrypsin	Bovine pancreas	R_{n-1} = bulky hydrophobic residues: Phe, Trp, Tyr; $R_n \neq$ Pro	Cleaves more slowly for R_{n-1} = Asn, His, Met, Leu
Elastase	Bovine pancreas	R_{n-1} = small neutral residues: Ala, Gly, Ser, Val; $R_n \neq$ Pro	
Thermolysin	*Bacillus thermoproteolyticus*	R_n = Ile, Met, Phe, Trp, Tyr, Val; $R_{n-1} \neq$ Pro	Occasionally cleaves at R_n = Ala, Asp, His, Thr; heat stable
Pepsin	Bovine gastric mucosa	R_n = Leu, Phe, Trp, Tyr; $R_{n-1} \neq$ Pro	Also others; quite nonspecific; pH optimum = 2
Endopeptidase V8	*Staphylococcus aureus*	R_{n-1} = Glu	

Table 5-3 Fundamentals of Biochemistry, 2/e

Cyanogen bromide cleavage of a polypeptide

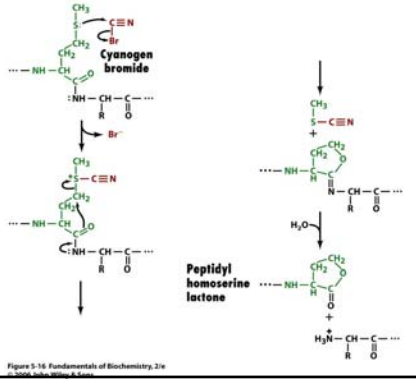


Figure 5-16 Fundamentals of Biochemistry, 2/e

Sequencing by Mass Spectrometry

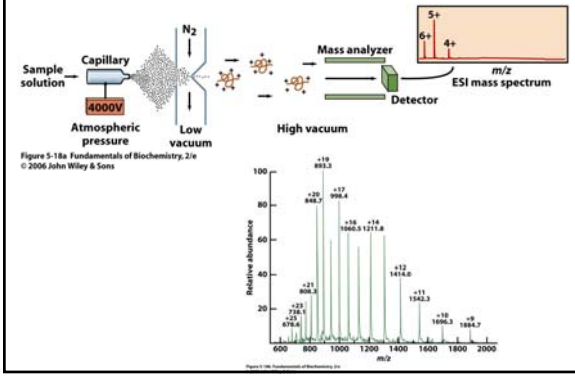
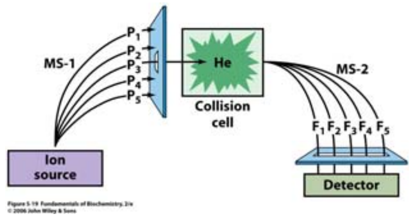


Figure 5-18a Fundamentals of Biochemistry, 2/e
© 2006 John Wiley & Sons



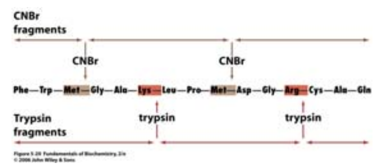
Tandem Mass Spectrometry in amino acid sequencing

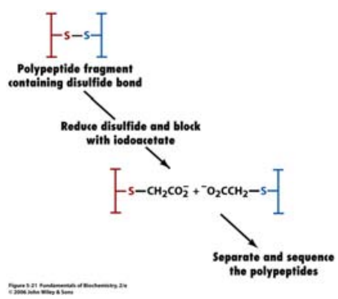
Reconstructing the protein's sequence

Specific chemical cleavage reagents

Cleave the large protein using i.e trypsin, separate fragments and sequence all of them. (We do not know the order of the fragments!!)

Cleave with a different reagent i.e. Cyanogen Bromide, separate the fragments and sequence all of them. Align the fragments with overlapping sequence to get the overall sequence.





Determining the positions of disulfide bond

How to assemble a protein sequence

1. Write a blank line for each amino acid in the sequence starting with the N-terminus.
2. Follow logically each clue and fill in the blanks.
3. Identify overlapping fragments and place in sequence blanks accordingly.
4. Make sure logically all your amino acids fit into the logical design of the experiment.
5. Double check your work.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

H_3N^+ – _ – _ – _ – _ – _ – _ – _ – _ – _ – _ – _ – _ – _ – _ – _ – COO^-

A - T
F - M - A - T
A- K - F - M
Q - M - A - K
D - I - K - Q - M
G - M - D - I - K
Y - R - G - M
Y - R

Cyanogen Bromide (CNBr) Cleaves after Met i.e M - X
D - I - K - Q - M
A - T
A - K - F - M
Y - R - G - M

Trypsin cleaves after K or R (positively charged amino acids)
Q - M - A - K
G - M - D - I - K
F - M - A - T
Y - R

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

Data Banks Containing Protein Sequences

ExPASy Molecular Biology Server (Swiss-Prot): http://au.expasy.org
Protein Information Resource (PIR): http://pir.georgetown.edu/
Protein Research Foundation (PRF): http://www4.prf.or.jp/
UniProt: http://www.ebi.uniprot.org/

Data Banks Containing Gene Sequences

GenBank: http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
European Bioinformatics Institute (EBI): http://srs.ebi.ac.uk/
DBGET/Integrated Database Retrieval System: http://www.genome.ad.jp/dbget

General information about the UniProt/Swiss-Prot entry	
Entry name	RSN_HUMAN
Primary accession number	Q9HD89
Entered in Swiss-Prot	Release 40, 16-OCT-2001
Sequence was last modified	Release 40, 16-OCT-2001
Annotations were last modified	Release 44, 05-JUL-2004
Protein description	
Protein name	Resistin precursor
Synonyms	Cysteine-rich secreted protein FIZZ3 Adipose tissue-specific secretory factor ADSF C/EBP-epsilon regulated myeloid-specific secreted cysteine-rich protein Cysteine-rich secreted protein A12-alpha-like 2 UNQ407/PRO1199
Origin of the protein	
Gene	Gene name: RETN Synonyms: RSTN, FIZZ3, HXCP1
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

Figure 5-22 Fundamentals of Biochemistry, 2/e

Lecture 7
Thursday 9/15/09
Proteins: Evolution, and Analysis