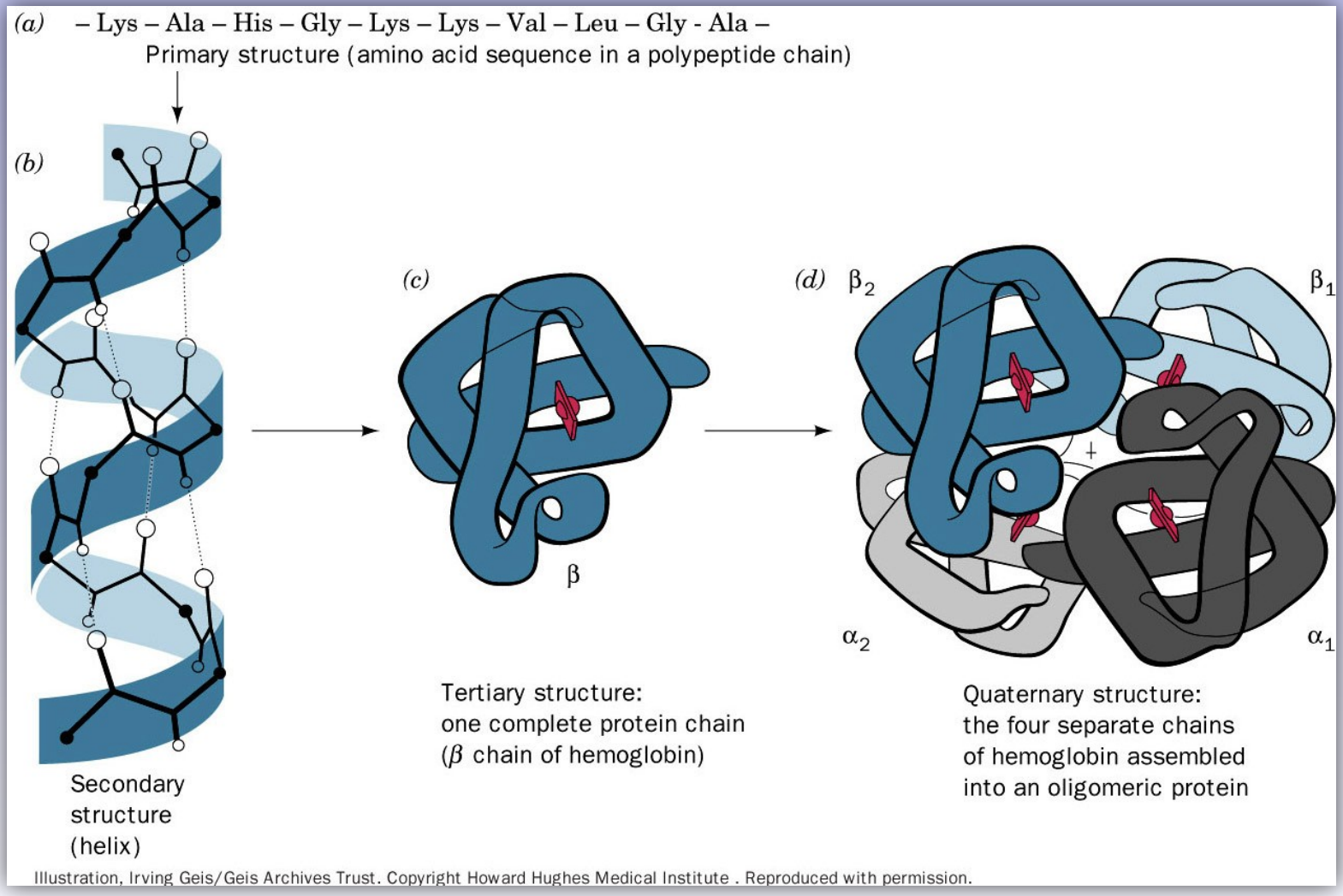
Proteins Primary Structure

Peptide/Protein Sequencing; Chemical Synthesis

CHEM 420 – Principles of Biochemistry
Instructor – Anthony S. Serianni

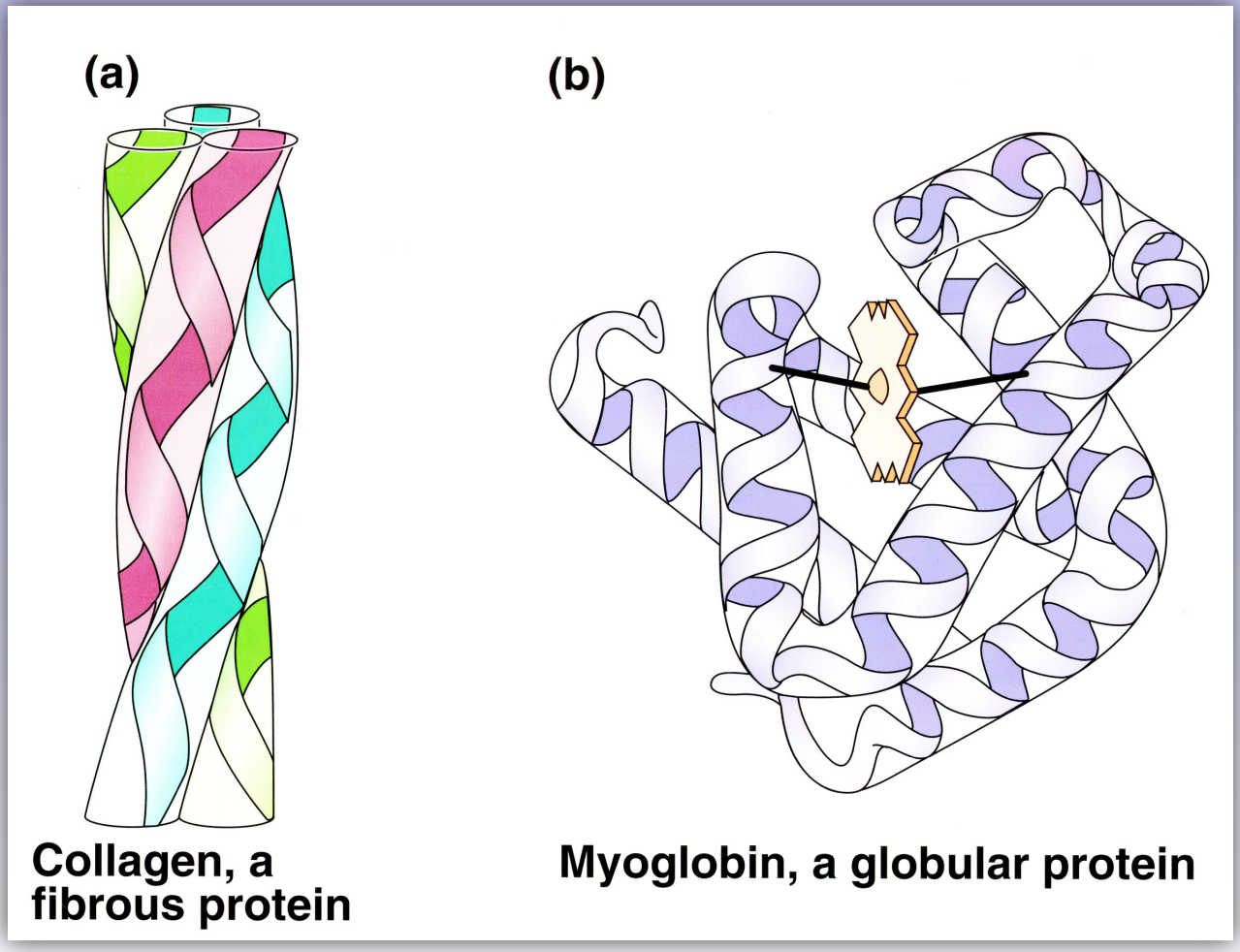
Chapter 7: Voet/Voet, *Biochemistry*, 2011
Fall 2015

September 7 & 9



Structural hierarchy of proteins

Representations of fibrous and globular proteins



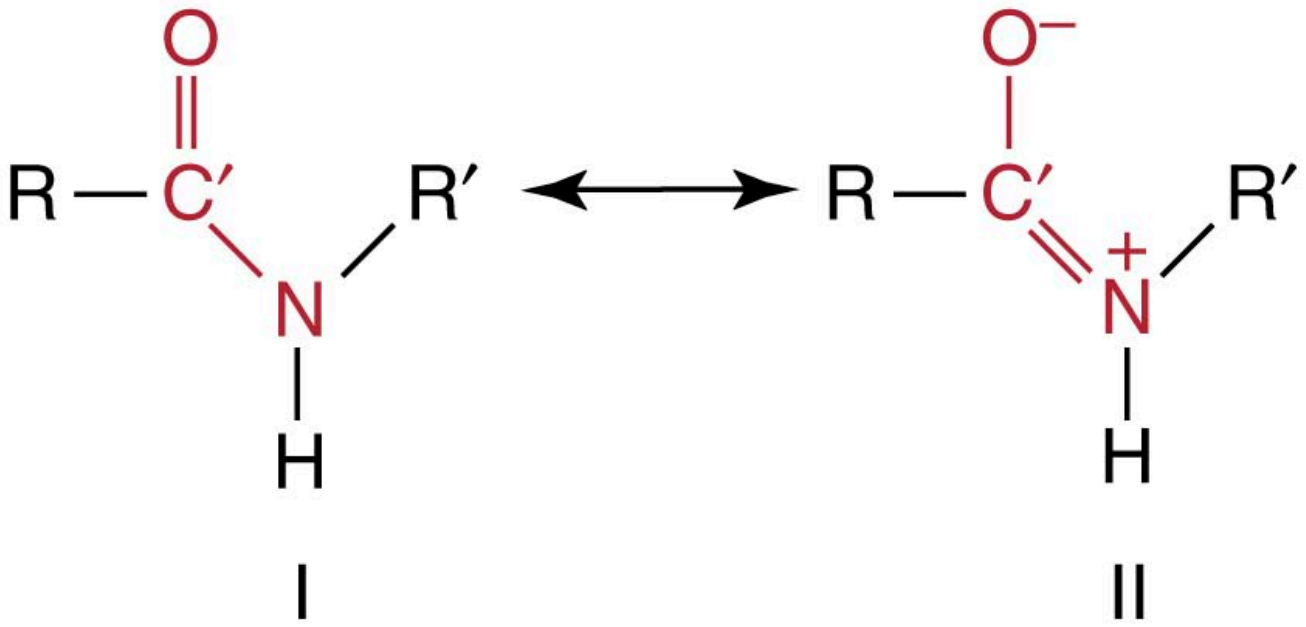


Figure 3.10. Electronic isomer structures of a peptide bond.

ϕ, ψ and peptide bonds along the backbone of a polypeptide

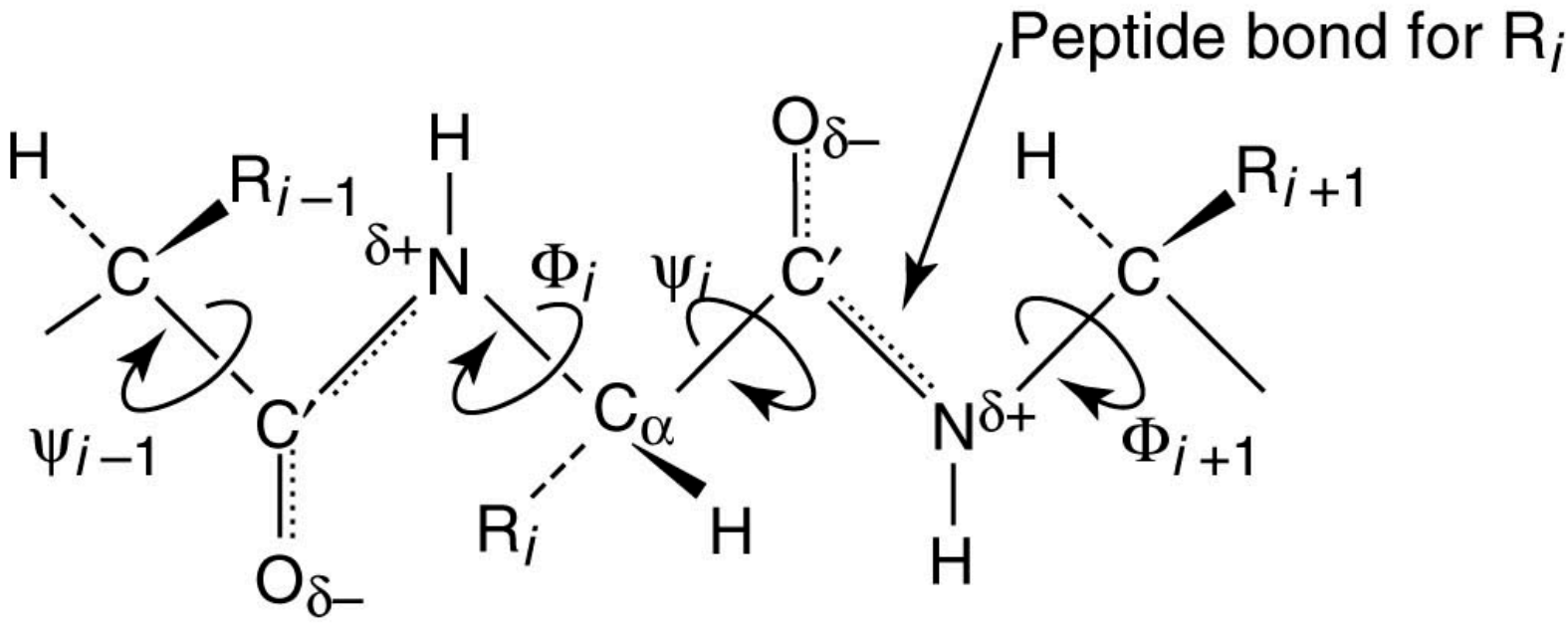
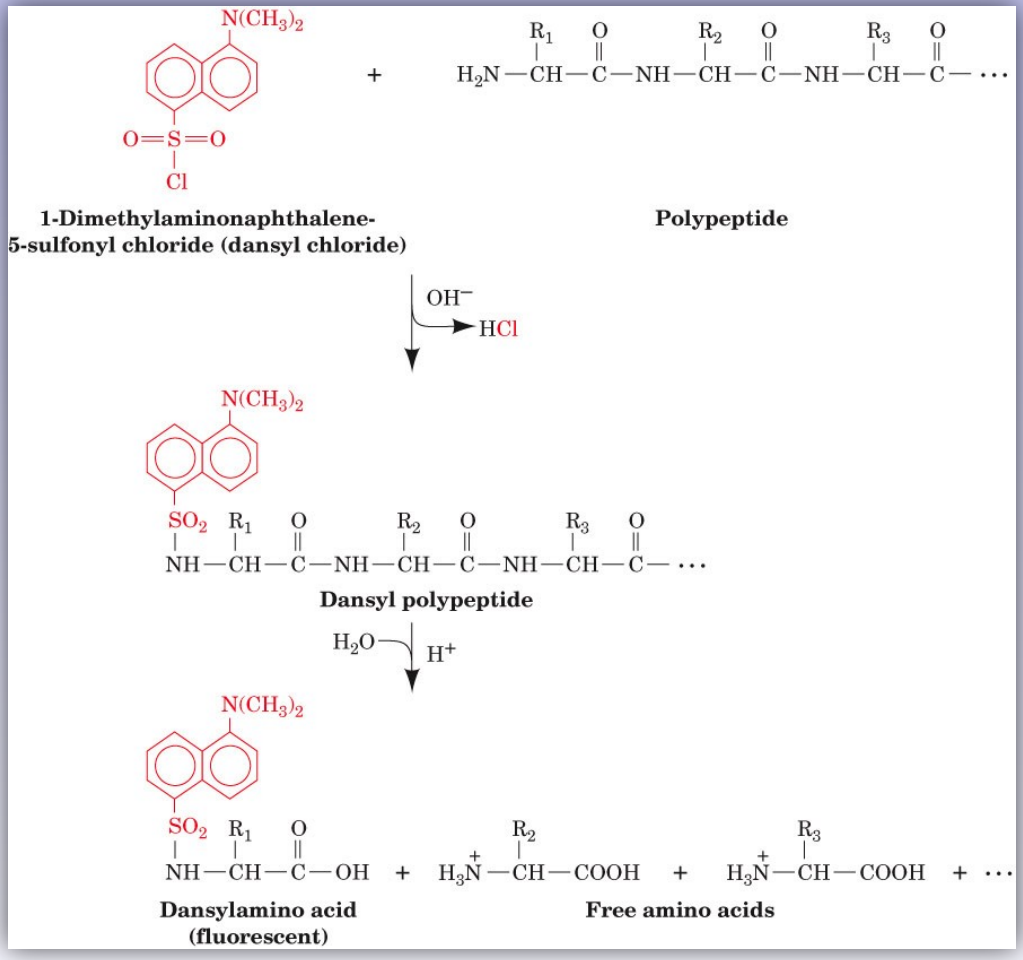
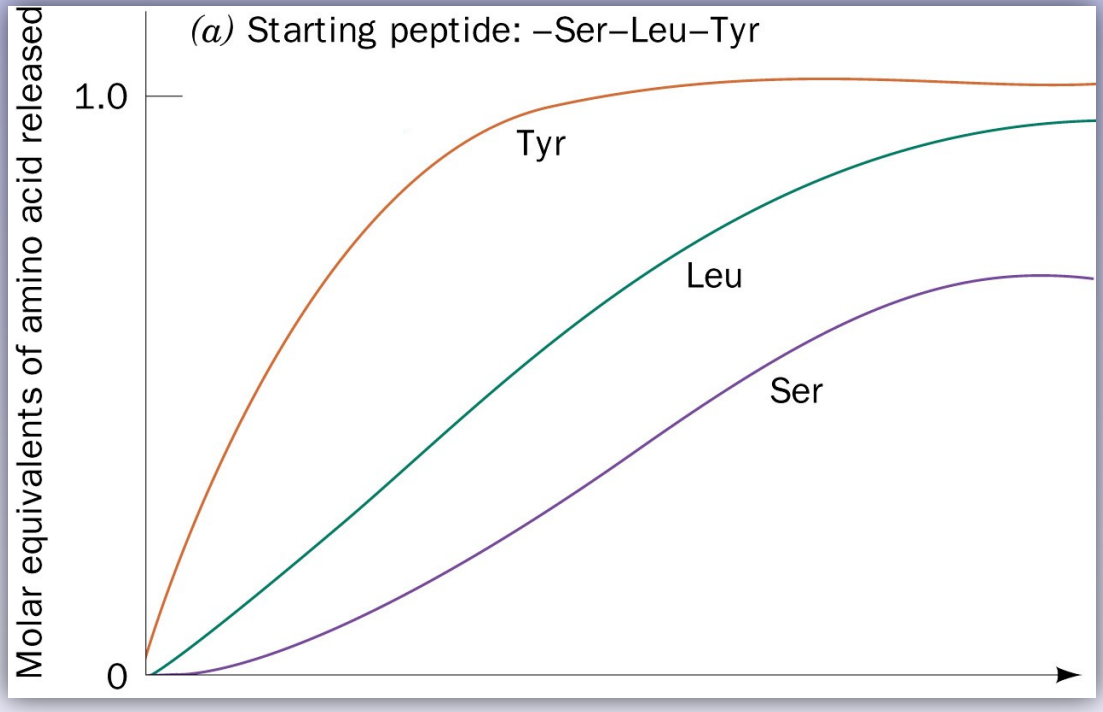


Figure 3.11. Amino acid residue within a polypeptide chain.

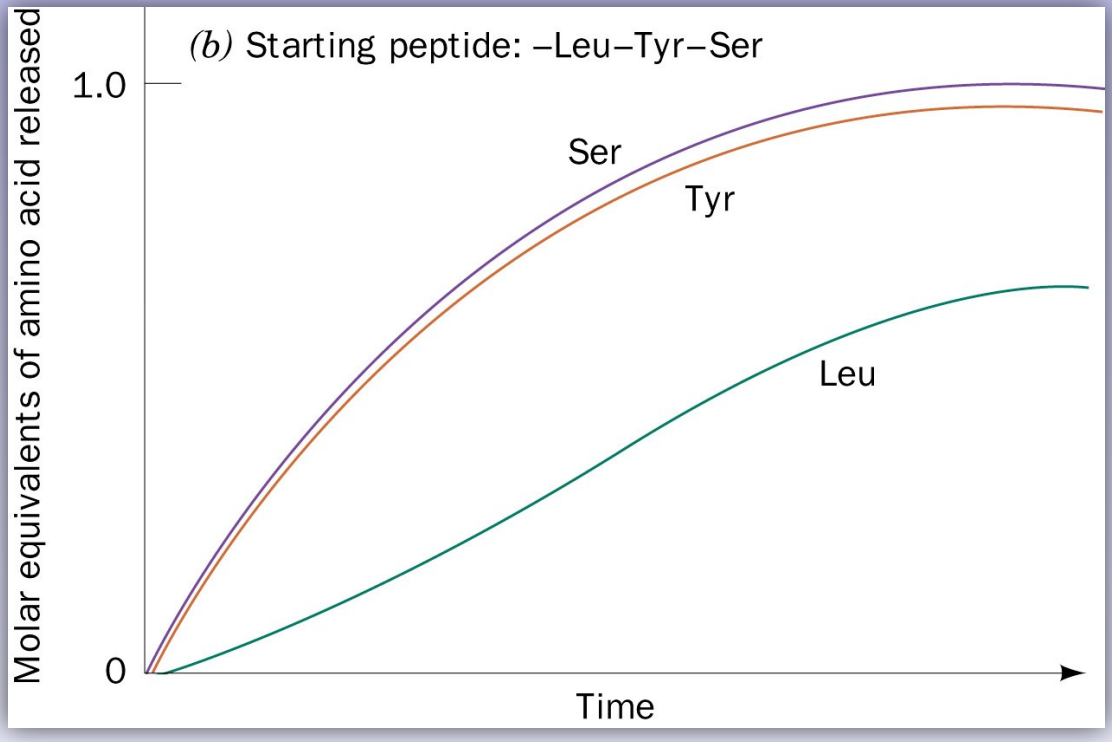
Determining the 1° structure (sequence) of a polypeptide



The reaction of dansyl chloride in end-group (*N*-terminus) analysis of a protein



Hypothetical rates of carboxypeptidase-catalyzed release of amino acids from the C-terminus of a protein: all bonds cleaved at the same rate

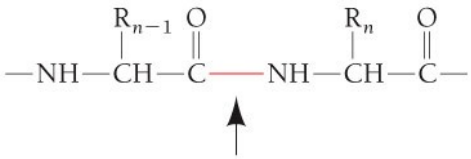


Hypothetical rates of carboxypeptidase-catalyzed release of amino acids from the C-terminus of a protein: Ser slow, Tyr fast, and Leu intermediate

Enzyme	Source	Specificity[a]
Carboxypeptidase A	Bovine pancreas	$R_n \neq$ Arg, Lys, Pro; $R_{n=1} \neq$ Pro
Carboxypeptidase B	Bovine pancreas	$R_n =$ Arg, Lys; $R_{n=1} \neq$ Pro
Carboxypeptidase C	Citrus leaves	All free C-terminal residues; pH optimum = 3.5
Carboxypeptidase Y	Yeast	All free C-terminal residues, but slowly with $R_n =$ Gly
Leucine aminopeptidase	Porcine kidney	$R_1 \neq$ Pro
Aminopeptidase M	Porcine kidney	All free N-terminal residues

[a] R_1 = the N-terminal residue; R_n = the C-terminal residue.

Specificities of various exopeptidases: C-terminus and N-terminus

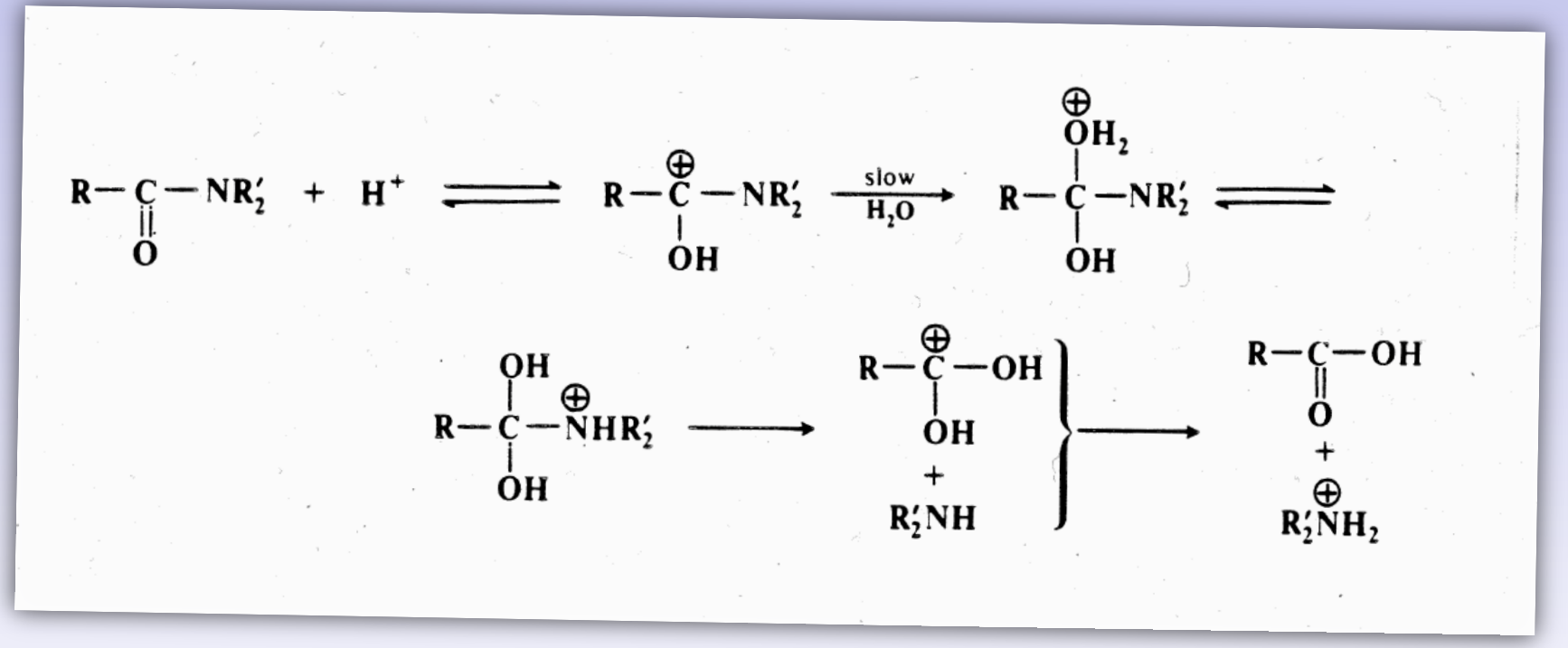
Enzyme	Source	Specificity	Comments
			
Trypsin	Bovine pancreas	R_{n-1} = positively charged residues: Arg, Lys; $R_n \neq$ Pro	Highly specific
Chymotrypsin	Bovine pancreas	R_{n-1} = bulky hydrophobic residues: Phe, Trp, Tyr; $R_n \neq$ Pro	Cleaves more slowly for R_{n-1} = Asn, His, Met, Leu
Elastase	Bovine pancreas	R_{n-1} = small neutral residues: Ala, Gly, Ser, Val; R_n = 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 Arg-C	Mouse submaxillary gland	R_{n-1} = Arg	May cleave at $R_{n\ 1}$ = Lys
Endopeptidase Asp-N	*Pseudomonas fragi*	R_n = Asp	May cleave at R_n = Glu
Endopeptidase Glu-C	*Staphylococcus aureus*	R_{n-1} = Glu	May cleave at R_{n-1} = Gly
Endopeptidase Lys-C	*Lysobacter enzymogenes*	R_{n-1} = Lys	May cleave at R_{n-1} = Asn

Specificities of various endopeptidases: R_{n-1} and R_n recognition

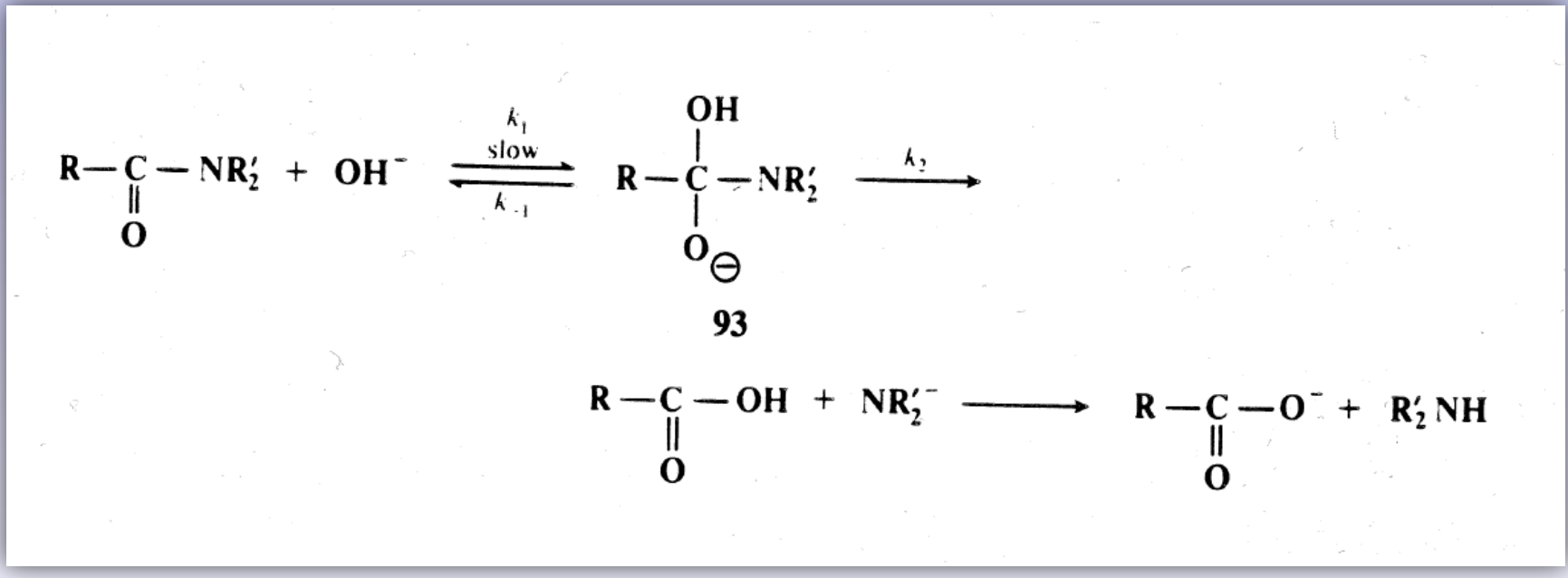
Enzyme mechanism: Serine proteases

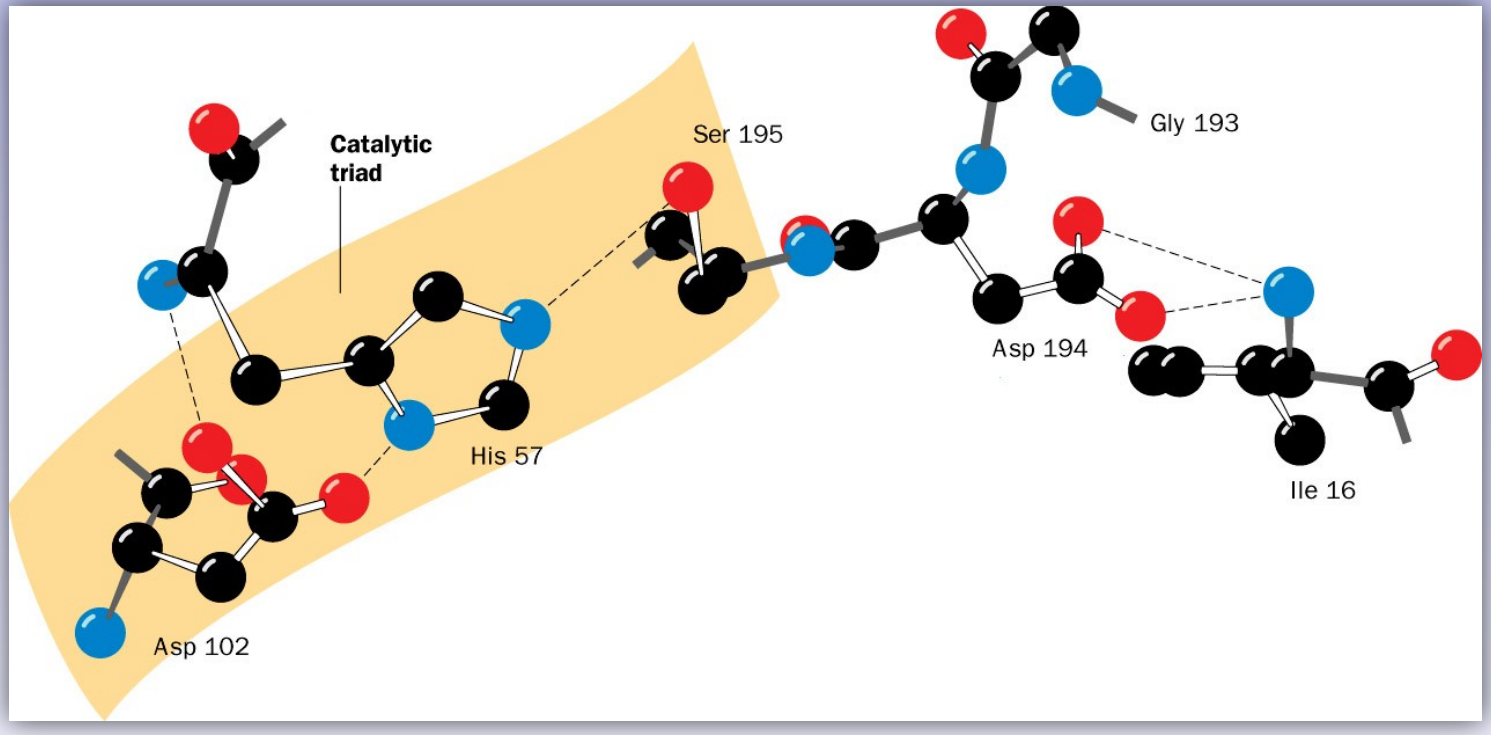
Trypsin and Chymotrypsin

Amide Bond Hydrolysis: Acid-catalyzed Mechanism



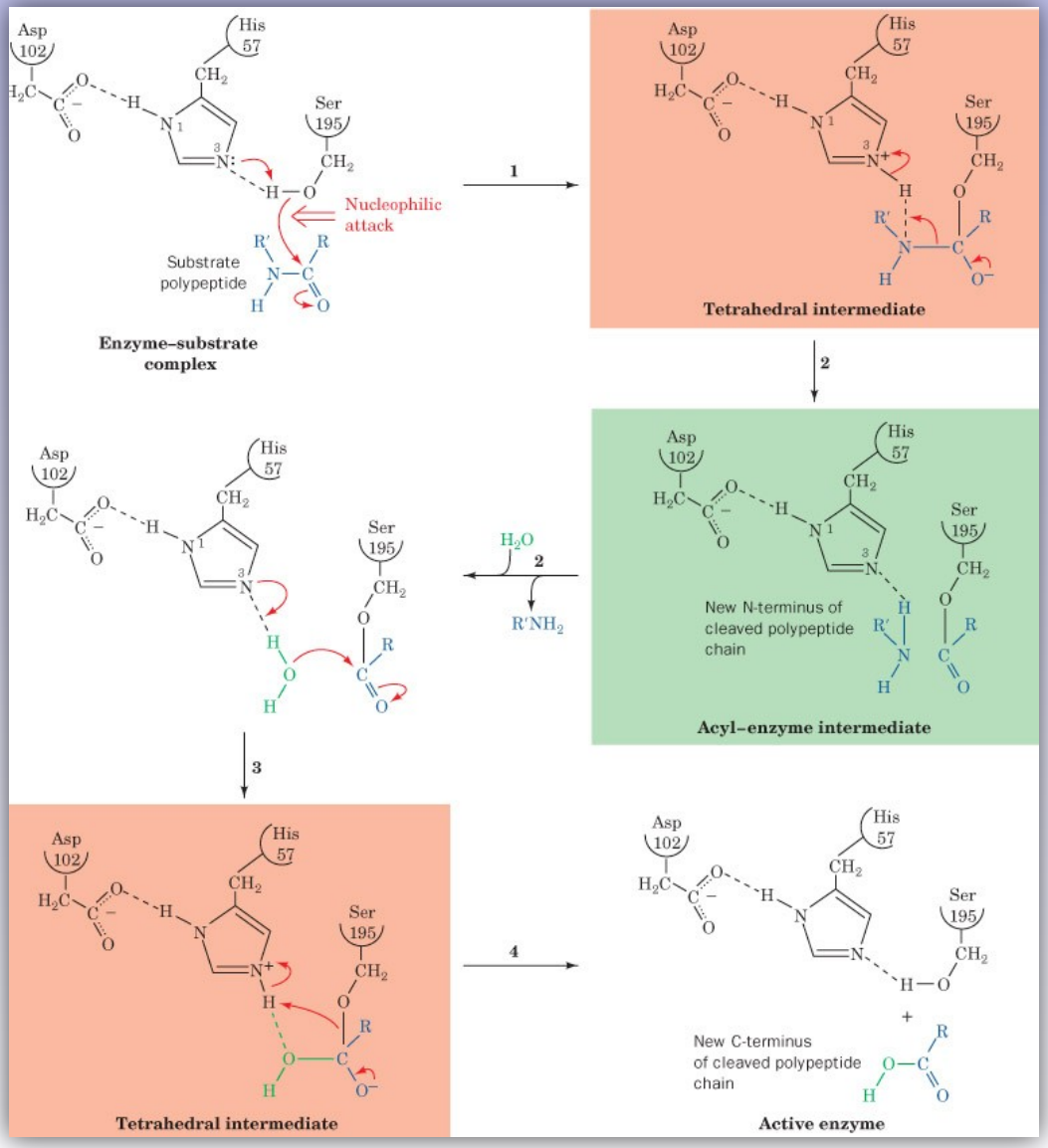
Amide Bond Hydrolysis: Base-catalyzed Mechanism



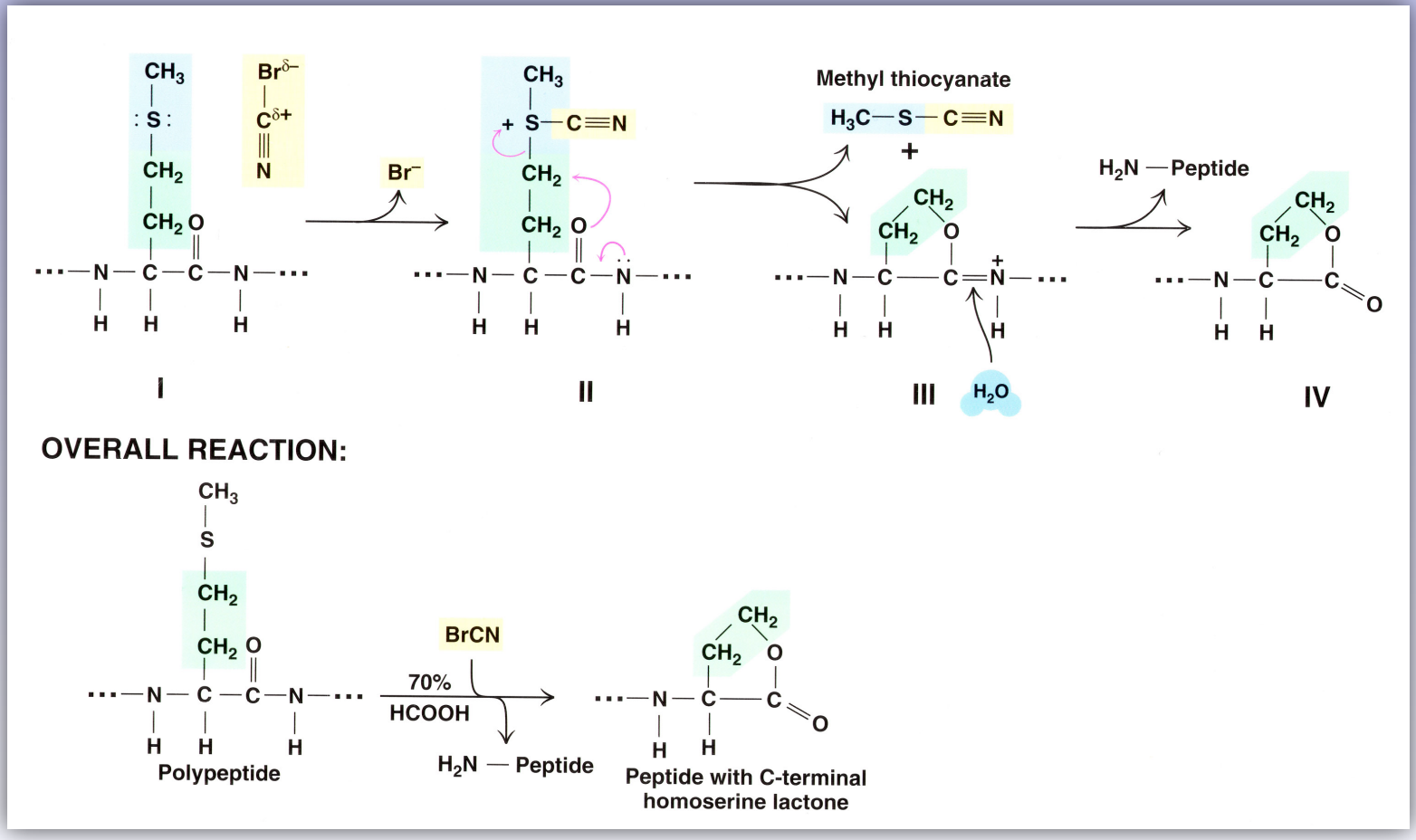


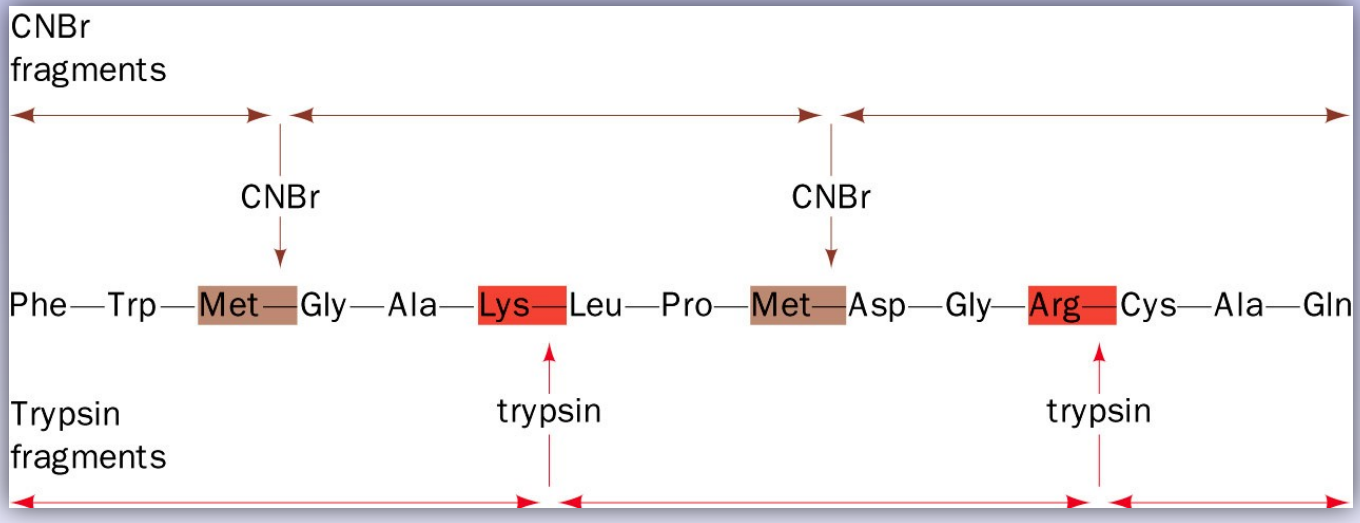
Active site residues of chymotrypsin (catalytic triad): a serine protease

Catalytic mechanism of serine proteases

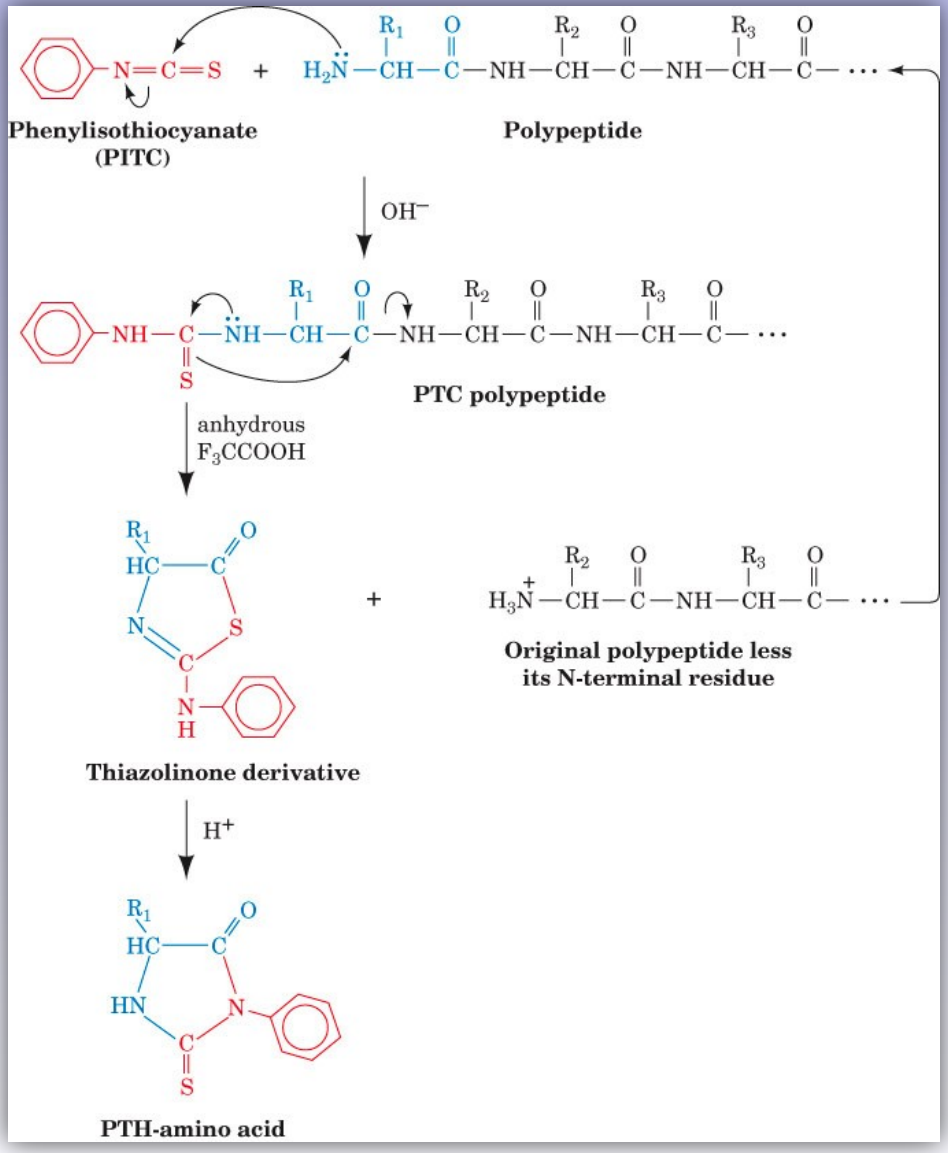


Cyanogen bromide mediated cleavage of a peptide bond



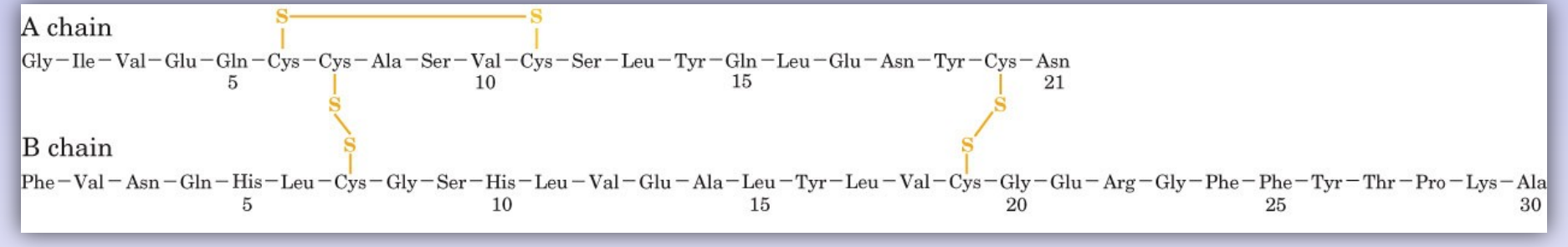


The amino acid sequence of an oligopeptide showing its susceptibility to CNBr and trypsin cleavage



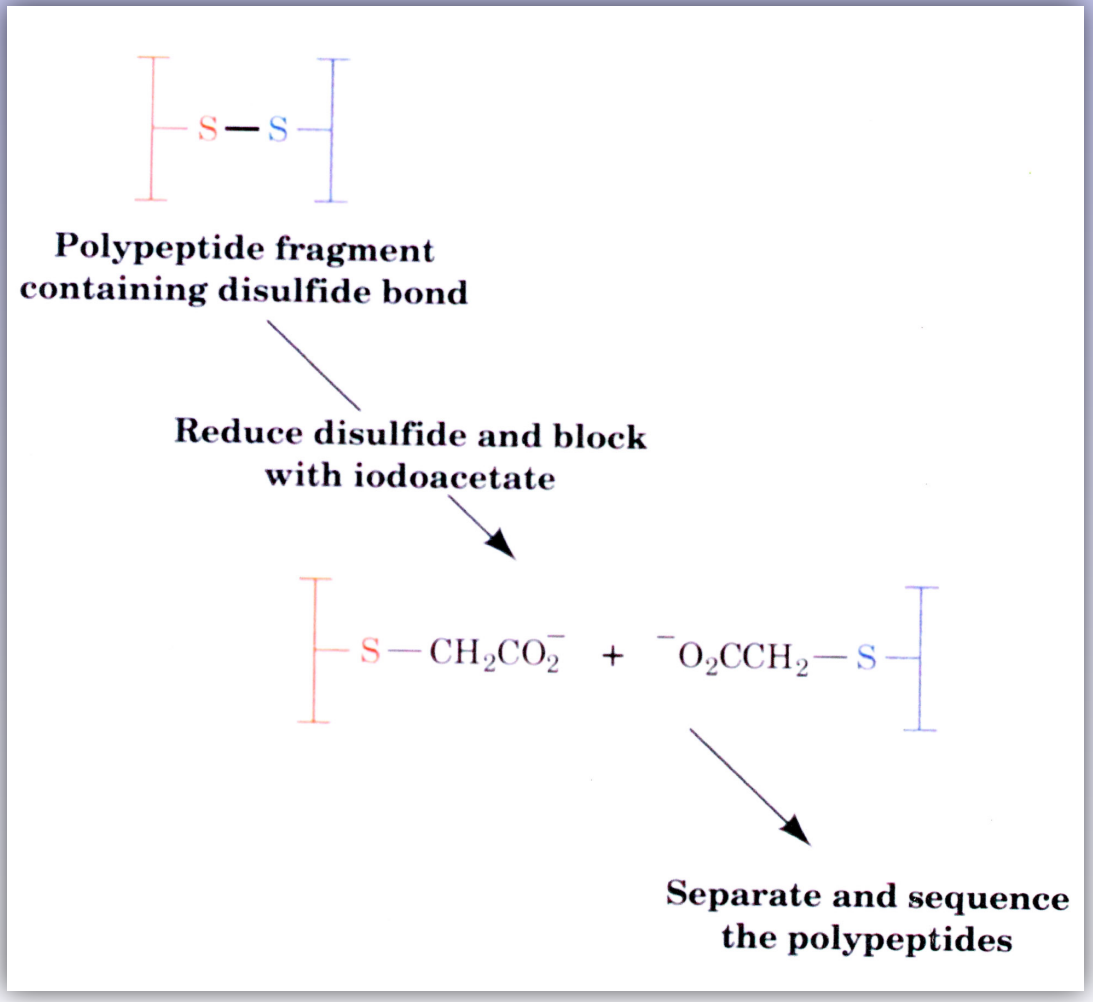
Sequencing the proteolytic fragments

Edman degradation. Acid hydrolysis of the PTC polypeptide yields the PTH amino acid (which is identified analytically) and the intact polypeptide minus one amino acid from the N-terminus.



Primary structure of bovine insulin. The native molecule is comprised of two separate oligopeptide chains linked at two sites by interchain disulfide bonds (cystine). An intrachain disulfide bond is also present in the native structure.

Reductive cleavage and alkylation of a disulfide bond



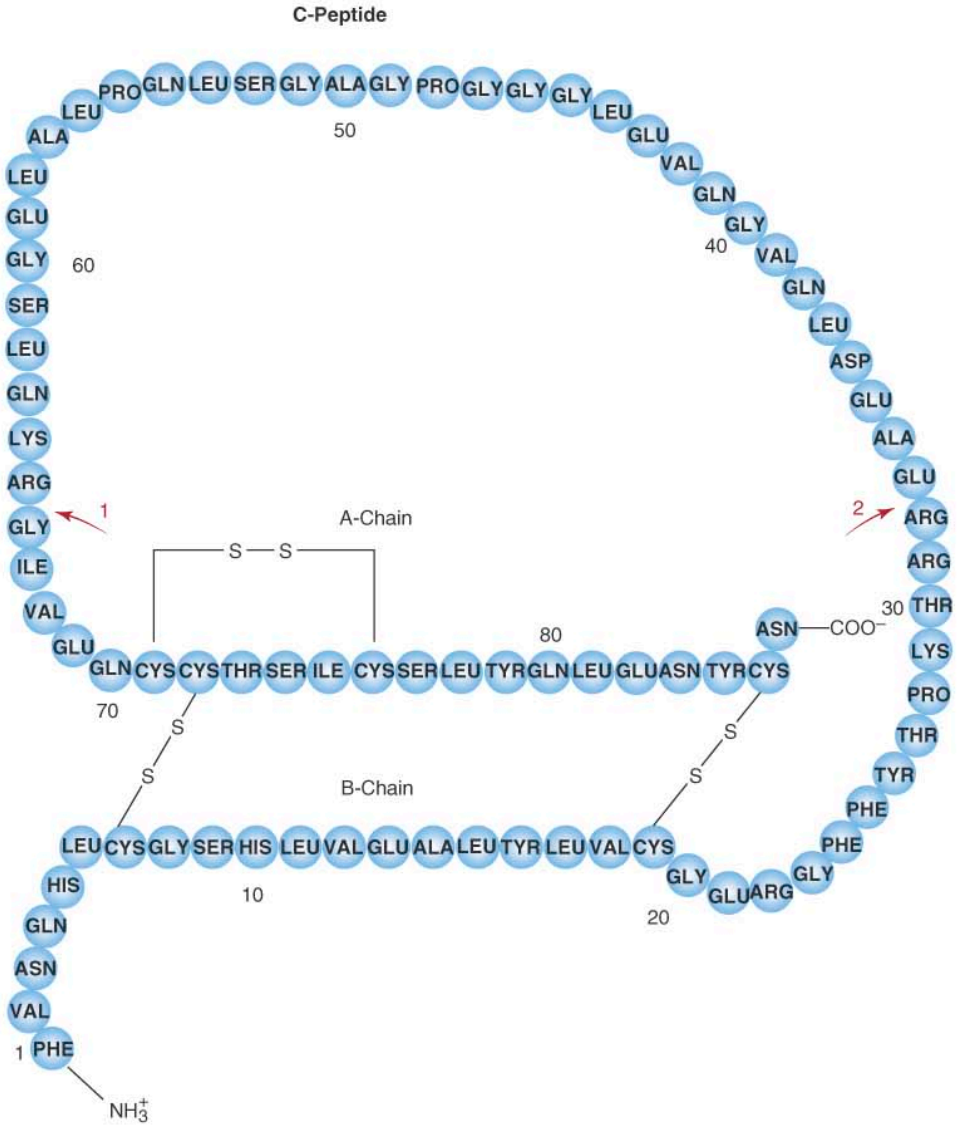


Figure 6.19. Maturation of human proinsulin. Redrawn from Bell, G. I., Swain, W. F., Pictet, R., Cordell, B., Goodman, H.M., and Rutter, W. *J. Nature* 282:525, 1979.

Figure 5.21 Structure of preproinsulin and its conversion to insulin

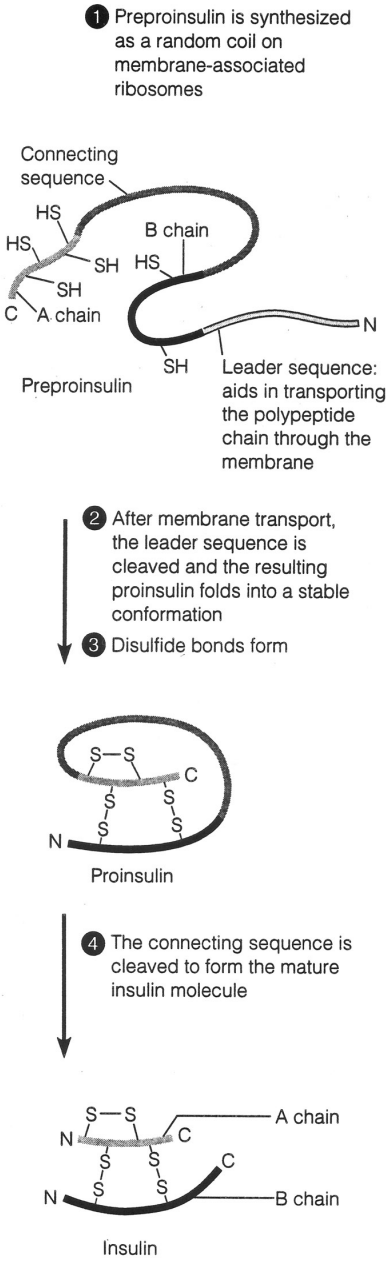
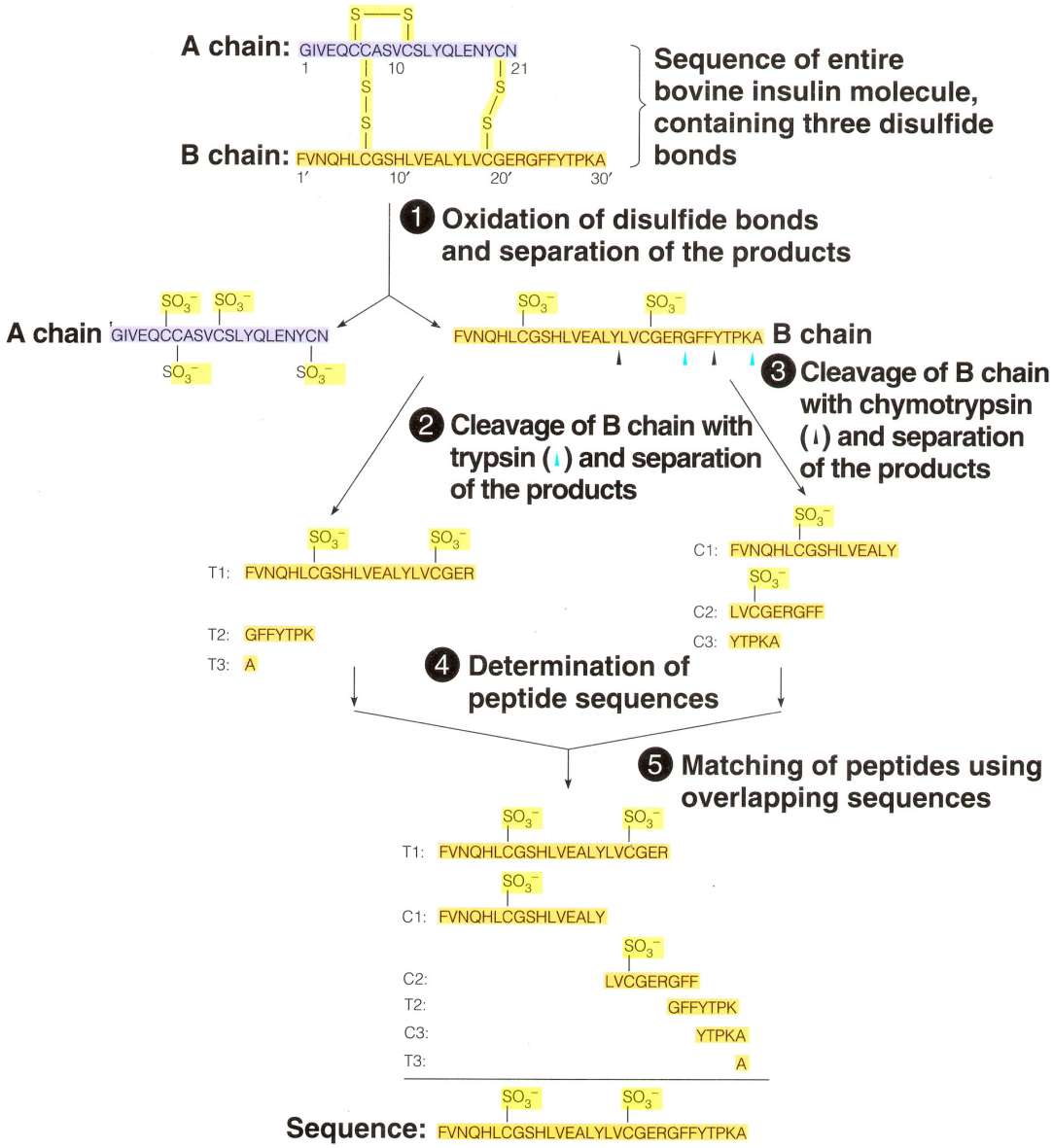
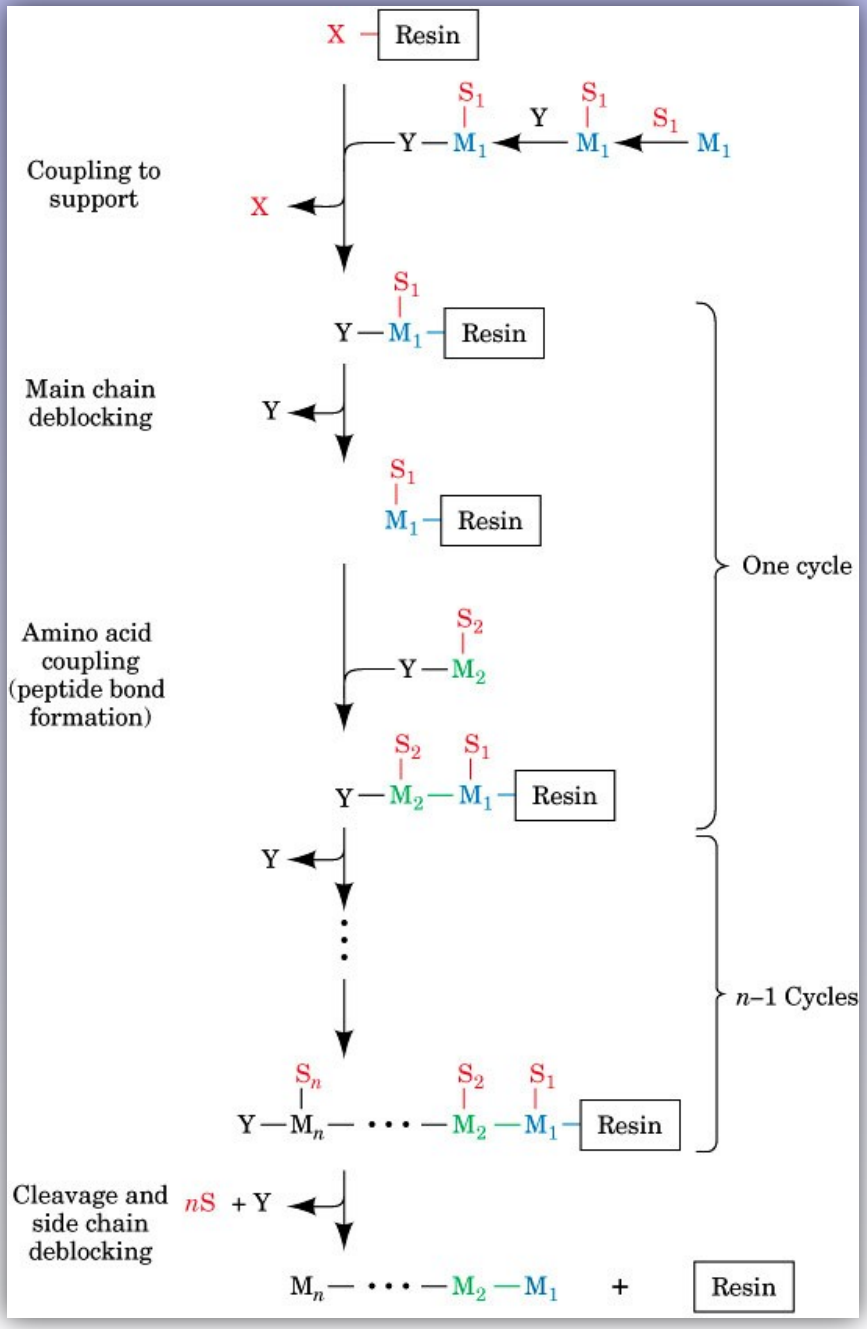


Figure 5D.1 Sequencing the β chain of insulin

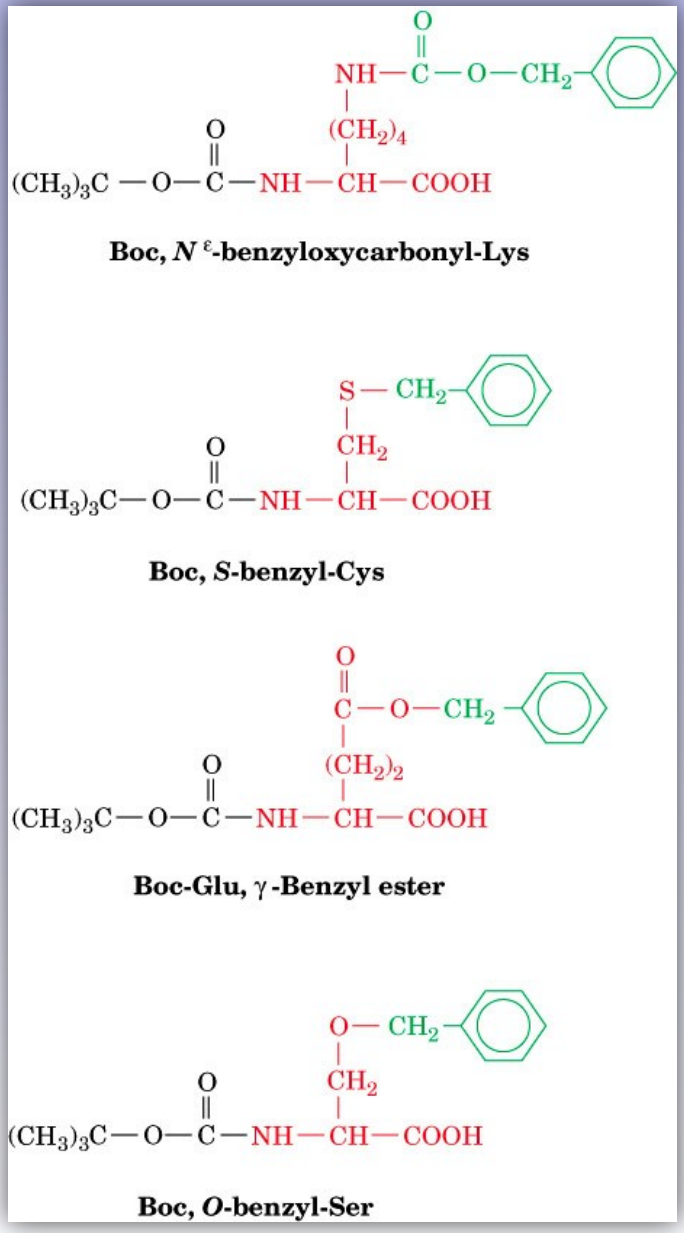




Solid-phase synthesis of peptides and proteins

Flow diagram for the chemical synthesis of a polypeptide by the solid phase method (Merrifield synthesis)

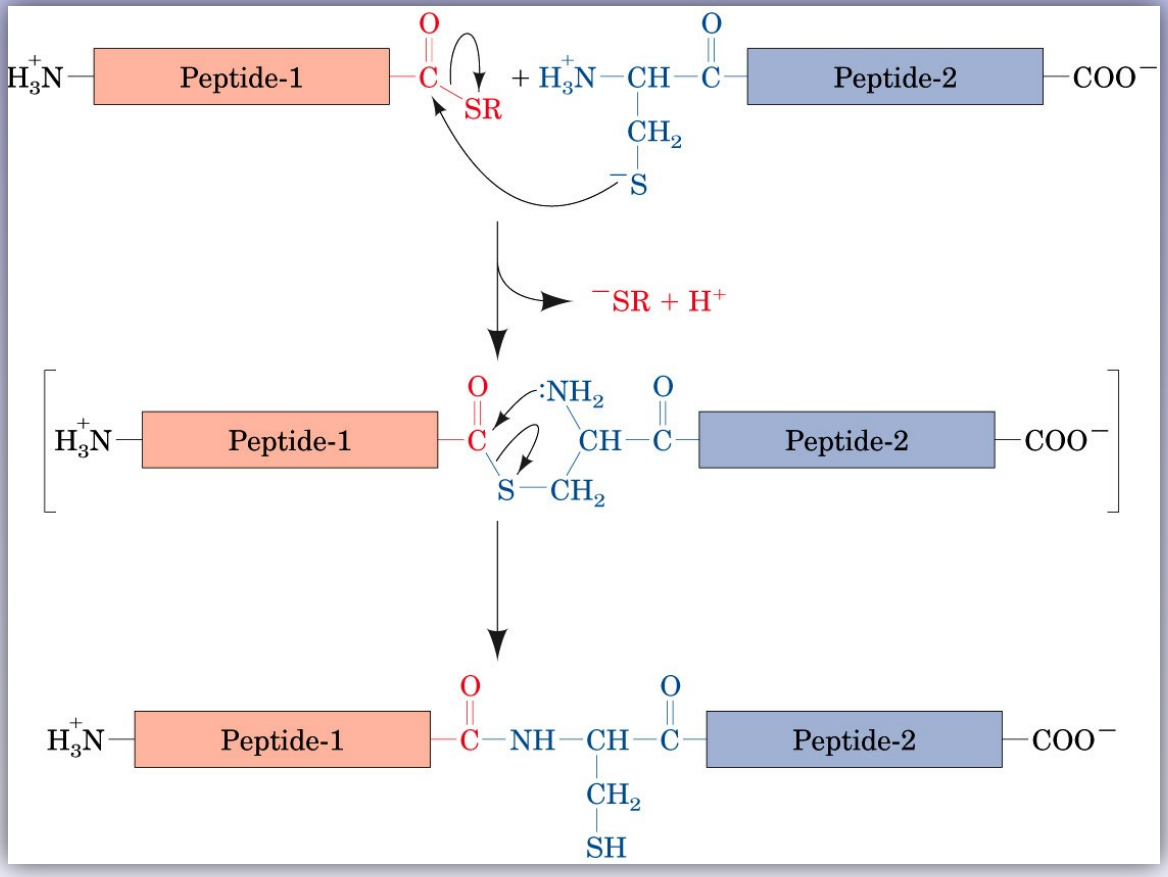
Synthesis direction: C-terminus to N-terminus



Some amino acid derivatives containing benzyl-protected sidechains and BOC-protected α-amino groups used in solid-phase peptide synthesis

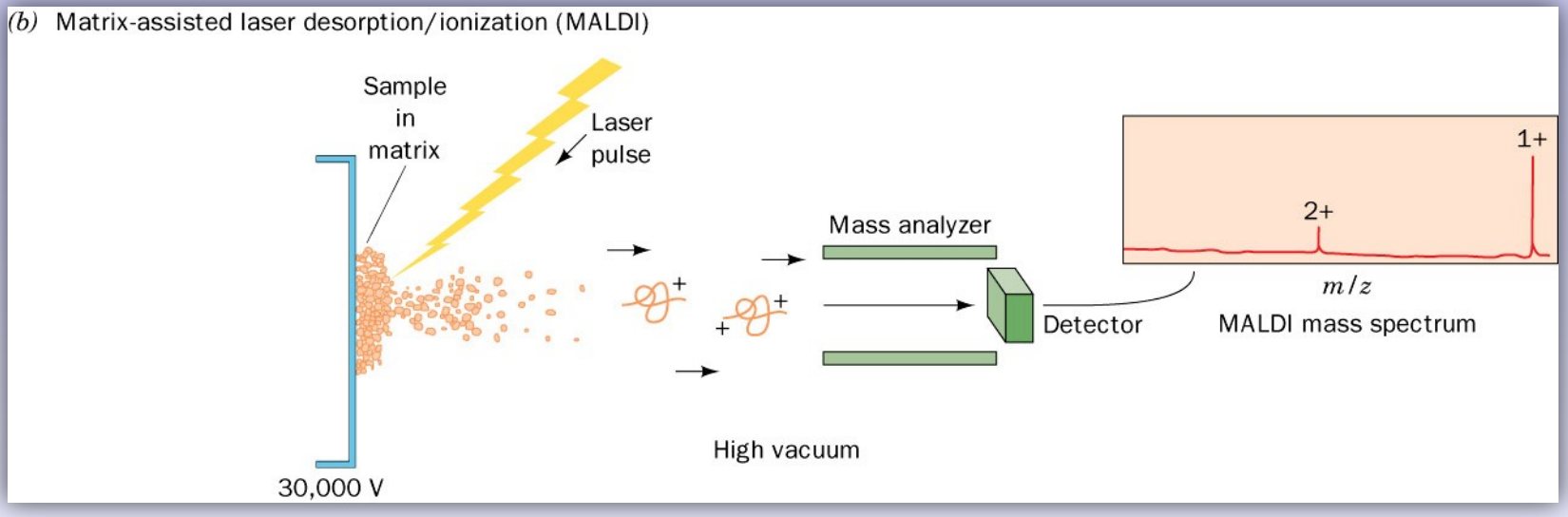
BOC = *t*-butyloxycarbonyl

Linking two peptides together chemically



Connecting two peptides via chemical coupling:
native chemical ligation

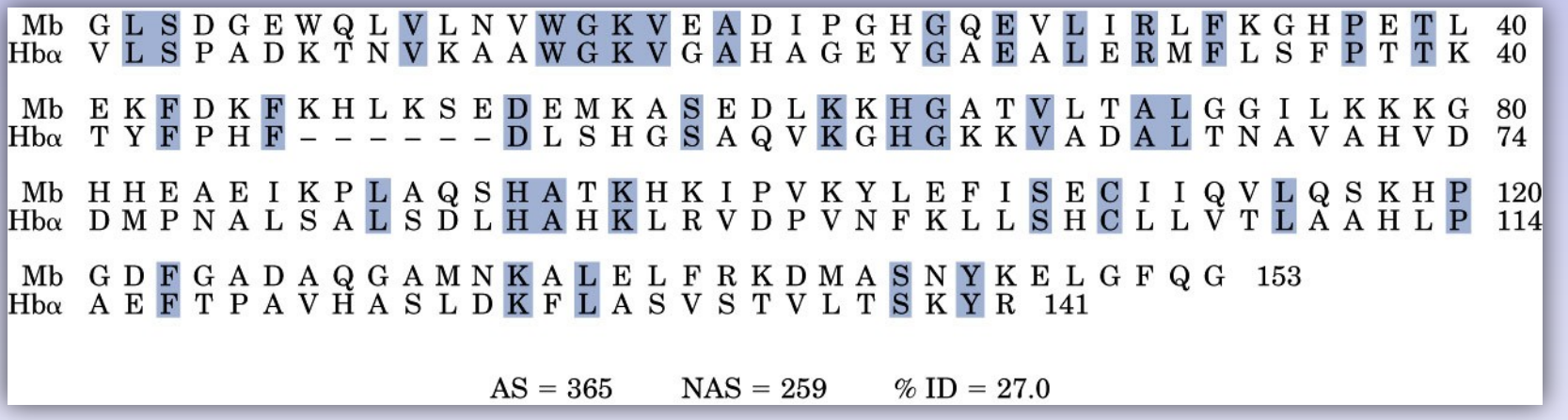
Analytical tools for peptide/protein characterization:
Mass spectrometry



Generation of gas-phase ions required for mass spectrometric analysis of proteins: matrix-assisted laser desorption/ionization (MALDI)

Bioinformatics: intersection of biotechnology and computer science; computational tools and methods used to extract useful structural information and relationships from protein and DNA sequence data

- sequence databases
- sequence alignment (homology; pairwise and multiple alignments)
- phylogenetic relationships



Optical alignments of human myoglobin (Mb, 153 residues) and the human hemoglobin α chain (Hbα, 141 residues)

AS = alignment score; NAS = normalized alignment score; %ID = percent identical

(a) **BLAST pairwise alignment**

```
>sp|P38524|HPI2_ECTVA HIGH POTENTIAL IRON-SULFUR PROTEIN, ISOZYME 2 (HIPIP 2)
          Length = 71

Score = 50.4 bits (118), Expect = 6e-07
Identities = 27/69 (39%), Positives = 35/69 (50%), Gaps = 4/69 (5%)

Query: 1  EPRAEDGHAHDYVNEAADASG--HPRYQEGQLCENCAFWGEAVQDGWGRCTHPDFDEVLVKAEGWCSVY  67
           E   +ED   A    +      DAS    HP  Y+EGQ  C  NC   +  +A     WG  C+    F    LV  A  GWC+  +
Sbjct: 2  ERLSEDDPAAQALEYRHDASSVQHPAYEEGQTCLNCLLYTDASAQDWGPCS--VFPGKLVSANGWCTAW  68
```

(b) **FASTA pairwise alignment**

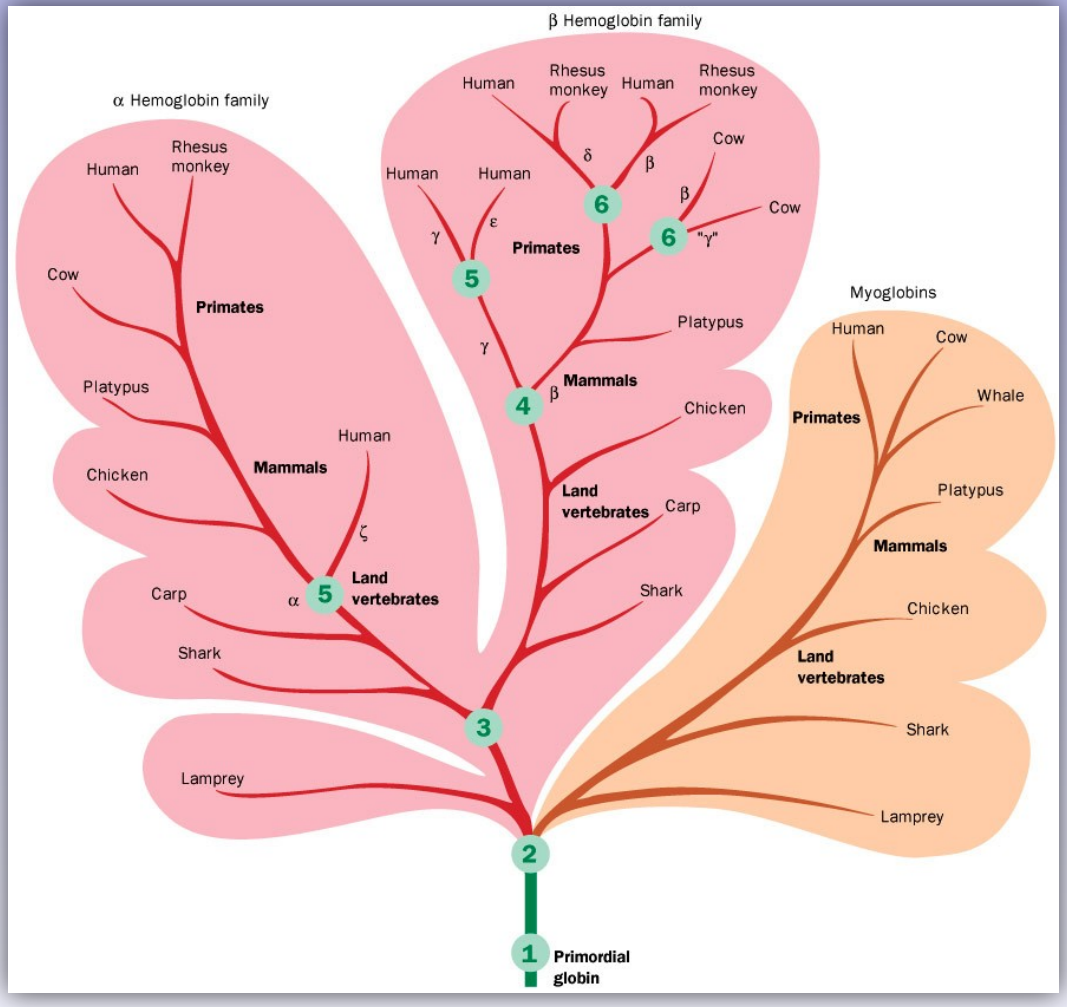
```
>>SWALL:HPI2_ECTVA P38524 HIGH POTENTIAL IRON-SULFUR PRO (71 aa)
initn: 102 init1: 77 opt: 116 Z-score: 278.0 expect() 4e-08
Smith-Waterman score: 116; 39.130% identity in 69 aa overlap (1-67:2-68)

                10        20        30        40        50        60        70
Sequen    EPRAEDGHAHDYVNEAADASG--HPRYQEGQLCENCAFWGEAVQDGWGRCTHPDFDEVLVKAEGWCSVYAPAS
          :   .::   :  .  ..    :::.   :: :.:::: : :: .. .:   .  :: :.   :    :::.:.:::.... .
SWALL:    MERLSEDDPAAQALEYRHDASSVQHPAYEEGQTCLNCLLYTDASAQDWGPCSV--FPGKLVSANGWCTAWVAR
                10        20        30        40        50        60        70
```

(c) **CLUSTAL X multiple-sequence alignment**

```
                                        *   *   :.:... :. .       *. * ** ::          * *   . *     *   :***:.:.
1  sp|P38524|HPI2  ----MERLSEDDPAAQALEYRHDASSVQ-HPAYE---EGQTCLNCLLYTDASAQDWGPC--SVFPGKLVSANGWCTAWVAR-
2  sp|P38941|HPI1  ----AERLDENSPEALALNYKHDGASVD-HPSHA---AGQKCINCLLYTDPSATEWGGC--AVFPNKLVNANGWCTAYVARG
3  sp|P00265|HPIS  -----APVDEKNPQAVALGYVSDAAKAD-KAKYKQFVAGSHCGNCALFQGKATDAVGGC--PLFAGKQVANKGWCSAWAKKA
4  sp|P04168|HPI1  --------EPRAEDGHAHDYVNEAADASGHPRYQ---EGQLCENCAFWGEAVQDGWGRCTHPDFDEVLVKAEGWCSVYAPAS
5  sp|P04169|HPI2  GLPDGVEDLPKAEDDHAHDYVNDAADTD-HARFQ---EGQLCENCQFWVDYVN-GWGYCQHPDFTDVLVRGEGWCSVYAPA-
```

Examples of peptide *pairwise* and *multiple* sequence alignments: BLAST (*b*asic *l*ocal *a*lignment *s*earch *t*ool); FASTA; CLUSTAL)



Phylogenetic tree of the globin family. Circled branch points represent gene duplications, and unmarked branch points are species divergences.

Web addresses for the major protein and DNA sequence data banks

Data Banks Containing Protein Sequences

ExPASy Molecular Biology Server (SWISS-PROT):
http://expasy.ch/
Protein Information Resource (PIR):
http://pir.georgetown.edu/
Protein Research Foundation (PRF):
http://www.prf.or.jp/en/

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/LinkDB Integrated Database Retrieval System:
http://www.genome.ad.jp/dbget/