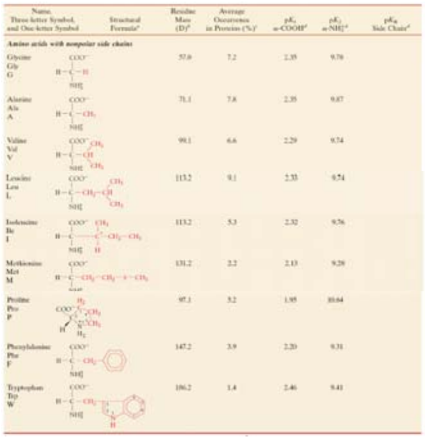
Proteins: Evolution, and Analysis

Lecture 7

9/15/2009

Chapter 4

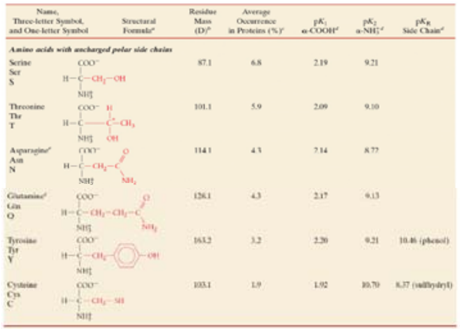
(1)



	Name, Three-letter Symbol, and One-letter Symbol	Structural Formula	Residue Mass (D)	Average Occurrence in Proteins (%)	pK_1 α-COOH	pK_2 α-NH_3^+	pK_R Side Chain
	Amino acids with nonpolar side chains						
G	Glycine Gly G	COO^- H–C–H NH_3^+	57.0	7.2	2.35	9.78	
A	Alanine Ala A	COO^- H–C–CH_3 NH_3^+	71.1	7.8	2.35	9.87	
V	Valine Val V	COO^- CH_3 H–C–CH NH_3^+ CH_3	99.1	6.6	2.29	9.74	
L	Leucine Leu L	COO^- CH_3 H–C–CH_2–CH NH_3^+ CH_3	113.2	9.1	2.33	9.74	
I	Isoleucine Ile I	COO^- CH_3 H–C–C*–CH_2–CH_3 NH_3^+ H	113.2	5.3	2.32	9.76	
M	Methionine Met M	COO^- H–C–CH_2–CH_2–S–CH_3 NH_3^+	131.2	2.2	2.13	9.28	
P	Proline Pro P		97.1	5.2	1.95	10.64	
F	Phenylalanine Phe F	COO^- H–C–CH_2– NH_3^+	147.2	3.9	2.20	9.31	
W	Tryptophan Trp W	COO^- H–C–CH_2– NH_3^+	186.2	1.4	2.46	9.41	

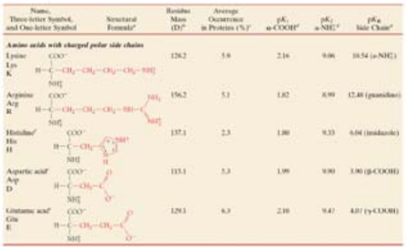
Chapter 4

(2)



	Name, Three-letter Symbol, and One-letter Symbol	Structural Formula	Residue Mass (D)	Average Occurrence in Proteins (%)	pK_1 α-COOH	pK_2 α-NH_3^+	pK_R Side Chain
	Amino acids with uncharged polar side chains						
S	Serine Ser S	COO^- H–C–CH_2–OH NH_3^+	87.1	6.8	2.19	9.21	
T	Threonine Thr T	COO^- H H–C–C*–CH_3 NH_3^+ OH	101.1	5.9	2.09	9.10	
N	Asparagine Asn N	COO^- O H–C–CH_2–C NH_3^+ NH_2	114.1	4.3	2.14	8.72	
Q	Glutamine Gln Q	COO^- O H–C–CH_2–CH_2–C NH_3^+ NH_2	128.1	4.3	2.17	9.13	
Y	Tyrosine Tyr Y	COO^- H–C–CH_2––OH NH_3^+	163.2	3.2	2.20	9.21	10.46 (phenol)
C	Cysteine Cys C	COO^- H–C–CH_2–SH NH_3^+	103.1	1.9	1.92	10.70	8.37 (sulfhydryl)

(3)



	Name, Three-letter Symbol, and One-letter Symbol	Structural Formula	Residue Mass (D)	Average Occurrence in Proteins (%)	pK_1 α-COOH	pK_2 α-NH_3^+	pK_R Side Chain
	Amino acids with charged polar side chains						
K	Lysine Lys K	COO^- H–C–CH_2–CH_2–CH_2–CH_2–NH_3^+ NH_3^+	128.2	5.9	2.16	9.06	10.54 (ε-NH_3^+)
R	Arginine Arg R	COO^- NH_2 H–C–CH_2–CH_2–CH_2–NH–C NH_3^+ NH_2^+	156.2	5.1	1.82	8.99	12.48 (guanidino)
H	Histidine His H	COO^- H–C–CH_2– NH_3^+	137.1	2.3	1.80	9.33	6.04 (imidazole)
D	Aspartic acid Asp D	COO^- O H–C–CH_2–C NH_3^+ O^-	115.1	5.3	1.99	9.90	3.90 (β-COOH)
E	Glutamic acid Glu E	COO^- O H–C–CH_2–CH_2–C NH_3^+ O^-	129.1	6.3	2.10	9.47	4.07 (γ-COOH)

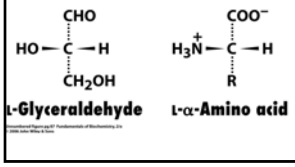
The Fischer Convention

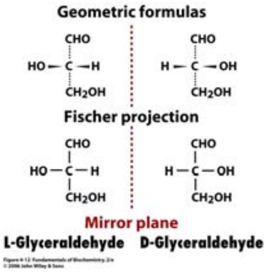
Absolute configuration about an asymmetric carbon

related to glyceraldehyde

(+) = D-Glyceraldehyde

(-) = L-Glyceraldehyde





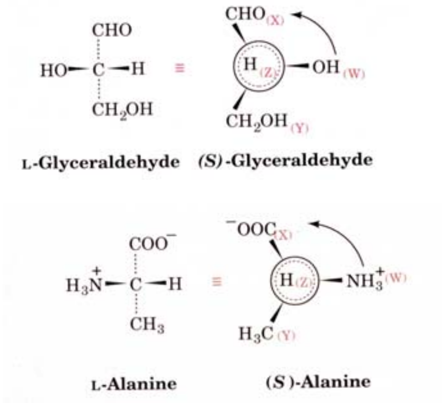
Cahn - Ingold - Prelog system

Can give absolute configuration nomenclature to multiple chiral centers.

Priority

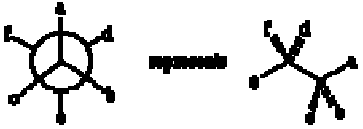
Atoms of higher atomic number bonded to a chiral center are ranked above those of lower atomic number with lowest priority away from you R highest to lowest = clockwise, S highest to lowest = counterclockwise

$SH>OH>NH_2>COOH>CHO>CH_2OH>C_6H_5>CH_3>H$



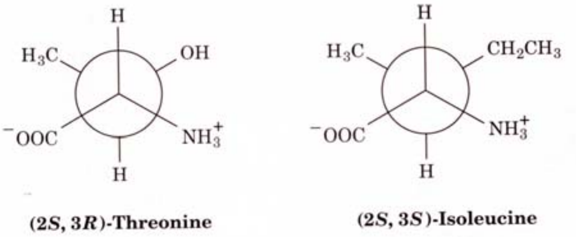
Newman Projection

- *A projection formula* representing the spatial arrangement of bonds on two adjacent atoms in a molecular entity.



- The structure appears as viewed along the bond between these two atoms, and the bonds from them to other groups are drawn as projections in the plane of the paper.
- The bonds from the atom nearer to the observer are drawn so as to meet at the centre of a circle representing that atom.
- Those from the further atom are drawn as if projecting from behind the circle.

The major advantage of the CIP or RS system is that the chiralities of compounds with multiple asymmetric centers can be unambiguously described



Structural Hierarchy in proteins

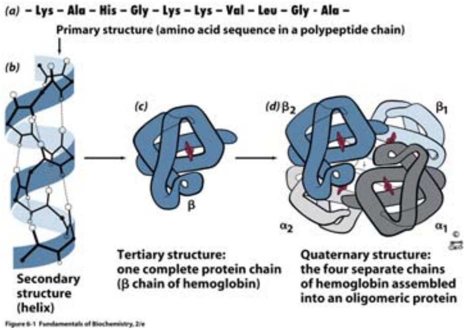

(a) – Lys – Ala – His – Gly – Lys – Lys – Val – Leu – Gly – Ala –
Primary structure (amino acid sequence in a polypeptide chain)
(b)
(c)
(d) β$_2$
β$_1$
β
α$_2$
α$_1$
Secondary structure (helix)
Tertiary structure: one complete protein chain (β chain of hemoglobin)
Quaternary structure: the four separate chains of hemoglobin assembled into an oligomeric protein

Figure 6-1 Fundamentals of Biochemistry, 2/e

Overview of Protein Sequencing

(1) Purify Protein

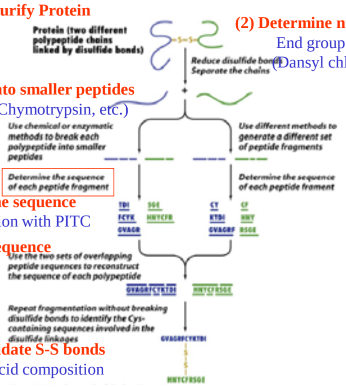
(2) Determine number of PP
End group analysis
(Dansyl chloride rxn)

(3) Fragment PP into smaller peptides
Enzymes (Trypsin, Chymotrypsin, etc.)
Chemical (CNBr)

(4) Determine the sequence
Edman Degradation with PITC

(5) Assemble a sequence

(6) Elucidate S-S bonds
Amino acid composition


Protein (two different polypeptide chains linked by disulfide bonds)
S–S
Reduce disulfide bonds
Separate the chains
+
Use chemical or enzymatic methods to break each polypeptide into smaller peptides
Use different methods to generate a different set of peptide fragments
Determine the sequence of each peptide fragment
Determine the sequence of each peptide frament
TDI
FCYK
GVAGR
SGE
HNYCFR
CY
KTDI
GVAGRF
CF
HNY
RSGE
Use the two sets of overlapping peptide sequences to reconstruct the sequence of each polypeptide
GVAGRFCYKTDI
HNYCFRSGE
Repeat fragmentation without breaking disulfide bonds to identify the Cys-containing sequences involved in the disulfide linkages
GVAGRFCYKTDI
S
S
HNYCFRSGE

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

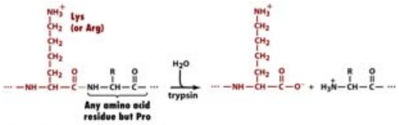
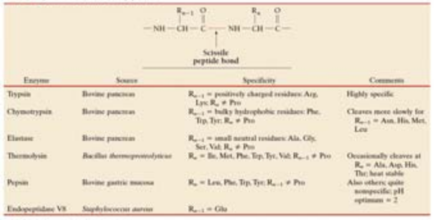


Table 5-3 Specificities of Various Endopeptidases


R$_{n-1}$
O
R$_n$
O
—NH—CH—C—NH—CH—C—
Scissile peptide bond

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
© 2006 John Wiley & Sons

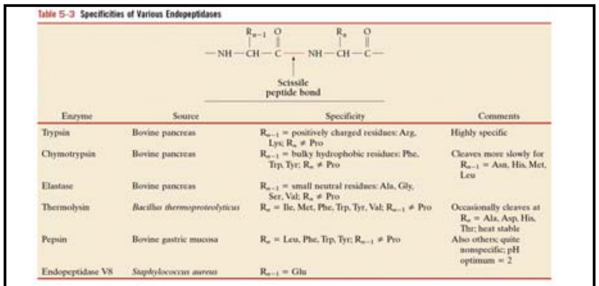


Table 5-3 Specificities of Various Endopeptidases

R_{n-1}
O
R_n
O
—NH—CH—C—NH—CH—C—
Scissile peptide bond

Enzyme	Source	Specificity	Comments
Trypsin	Bovine pancreas	R_{n-1} = positively charged residues: Arg, Lys; R_n ≠ Pro	Highly specific
Chymotrypsin	Bovine pancreas	R_{n-1} = bulky hydrophobic residues: Phe, Trp, Tyr; R_n ≠ 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} ≠ 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} ≠ Pro	Also others; quite nonspecific; pH optimum = 2
Endopeptidase V8	*Staphylococcus aureus*	R_{n-1} = Glu	

Q9. You must cleave the following peptide into smaller fragments. Which of the proteases listed in the table would be likely to yield the most fragments? The fewest?

NMTQGRCKPVNTFVHEPLVDVQNVCFKE

Cyanogen bromide cleavage of a polypeptide

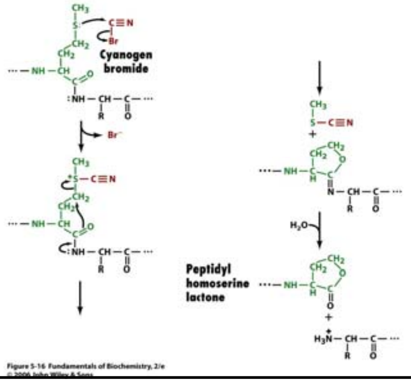


Figure 5-16 Fundamentals of Biochemistry, 2/e
© 2006 John Wiley & Sons

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.

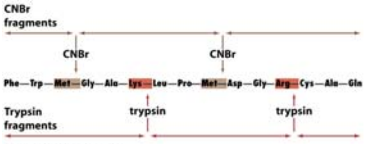
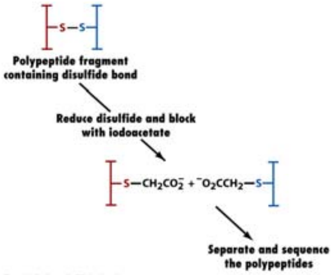


Figure 5-20 Fundamentals of Biochemistry, 2/e
© 2006 John Wiley & Sons



S—S
Polypeptide fragment containing disulfide bond
Reduce disulfide and block with iodoacetate
$S—CH_2CO_2^- + {}^-O_2CCH_2—S$
Separate and sequence the polypeptides

Figure 5-21 Fundamentals of Biochemistry, 2/e
© 2006 John Wiley & Sons

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

Q11. Separate cleavage reactions of a polypeptide by CNBr and chymotrypsin yield fragments with the following amino acid sequences. What is the the sequence of the intact polypeptide?

CNBr treatment	Chymotrypsin
1. Arg-Ala-Tyr-Gly-Asn	1. Met-Arg-Ala-Tyr
2. Leu-Phe-Met	2. Asp-Met-Leu-Phe
3. Asp-Met	3. Gly-Asn

Q13. Treatment of a polypeptide with 2-mercaptoethanol yields two PP:

1. Ala-Val-Cys-Arg-Thr-Gly-Cys-Lys-Asn-Phe-Leu

2. Tyr-Lys-Cys-Phe-Arg-His-Thr-Lys-Cys-Ser

Treatment of the intact PP with trypsin yields fragments with the following aa compositions:

3. (Ala, Arg, Cys2, Ser, Val)

5. (Asn, Leu, Phe)

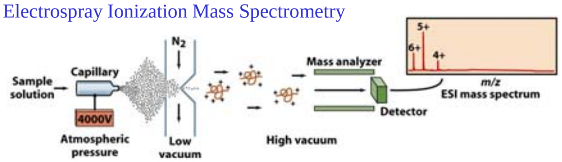
7. (Lys, Tyr)

4. (Arg, Cys2, Gly, Lys, Thr, Phe)

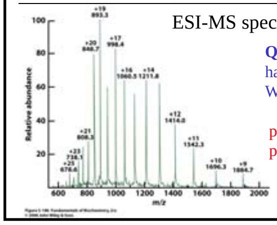
6. (His, Lys, Thr)

Sequencing by Mass Spectrometry

Electrospray Ionization Mass Spectrometry



ESI-MS spectrum of horse heart apomyoglobin

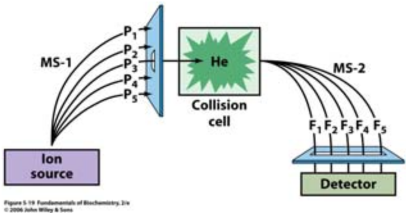


Q. Two successive peaks in the mass spectrum have measure m/z ratios of 1414.0 and 1542.3. What is the original apomyoglobin molecule?

p1= (M+z)/z

p2= (M+z-1)/z-1

M= 16,975D (16,951 D in table 5-1)



Tandem Mass Spectrometry in amino acid sequencing

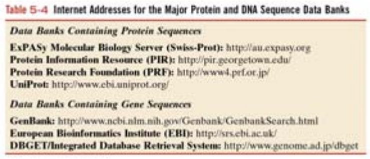


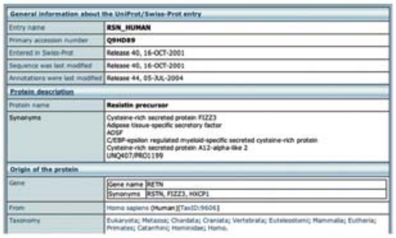
Table 5-4 Internet Addresses for the Major Protein and DNA Sequence Data Banks

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

Protein Evolution

Species variation in homologous proteins

The primary structures of a given protein from related species closely resemble one another. If one assumes, according to evolutionary theory, that related species have evolved from a common ancestor, it follows that each of their proteins must have likewise evolved from the corresponding ancestor.

A protein that is well adapted to its function, that is, one that is not subject to significant physiological improvement, nevertheless continues to evolve.

Neutral drift: changes not effecting function

Homologous proteins

(evolutionarily related proteins)

Compare protein sequences:

Conserved residues, i.e invariant residues reflect chemical necessities.

Conserved substitutions, substitutions with similar chemical properties (Asp for Glu), (Lys for Arg), (Ile for Val)

Variable regions, no requirement for chemical reactions etc.

Amino acid difference matrix for 26 species of cytochrome c

	Man,chimp	monkey	Horse	Donkey	cow,sheep	dog	gray whale	rabbit	kangaroo	Chicken,	penguin	Duck	Rattlesnake	turtle	Bullfrog	Tuna fish	worm fly	silkworm	Wheat	Bread mold	Yeast	Candida
Man,chimp	0																					
Rhesus monkey	1	0																				
Horse	12	11	0																			
Donkey	11	10	1	0																		
cow,sheep	10	9	3	2	0																	
dog	11	10	6	5	3	0																
gray whale	10	9	5	4	2	3	0															
rabbit	9	8	6	5	4	5	2	0														
kangaroo	10	11	7	8	6	7	6	6	0													
Chicken	13	12	11	10	9	10	9	8	12	0												
penguin	13	12	12	11	10	10	9	8	10	2	0											
Duck	11	10	10	9	8	8	7	6	10	3	3	0										
Rattlesnake	14	15	22	21	20	21	19	18	21	19	20	17	0									
turtle	15	14	11	10	9	9	8	9	11	8	8	7	22	0								
Bullfrog	18	17	14	13	11	12	11	11	13	11	12	11	24	10	0							
Tuna fish	21	21	19	18	17	18	17	17	18	17	18	17	26	18	15	0						
worm fly	27	26	22	22	22	21	22	21	24	23	24	22	29	24	22	24	0					
silk moth	31	30	29	28	27	25	27	26	28	28	27	27	31	28	29	32	14	0				
Wheat	43	43	46	45	45	44	44	44	47	46	46	46	46	46	48	49	45	45	0			
Bread mold	48	47	46	46	46	46	46	46	49	47	48	46	47	49	49	48	41	47	54	0		
Yeast	45	45	46	45	45	45	45	45	46	46	45	46	47	49	47	47	45	47	47	41	0	
Candida k.	51	51	51	50	50	49	50	50	51	51	50	51	51	53	51	48	47	47	50	42	27	0

Average differences

10.0

5.1

9.9

14.3

12.6

18.5

25.9

47.0

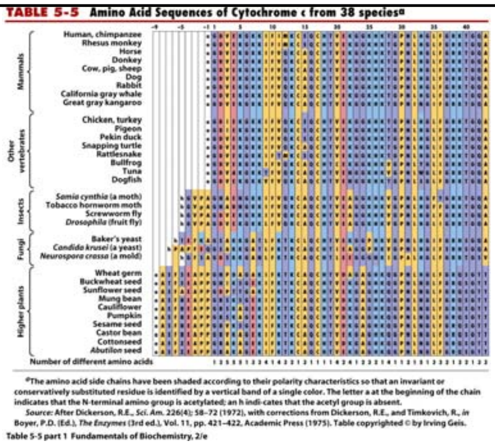


TABLE 5-5 Amino Acid Sequences of Cytochrome c from 38 species[a]

Mammals: Human, chimpanzee; Rhesus monkey; Horse; Donkey; Cow, pig, sheep; Dog; Rabbit; California gray whale; Great gray kangaroo

Other vertebrates: Chicken, turkey; Pigeon; Pekin duck; Snapping turtle; Rattlesnake; Bullfrog; Tuna; Dogfish

Insects: *Samia cynthia* (a moth); Tobacco hornworm moth; Screwworm fly; *Drosophila* (fruit fly)

Fungi: Baker's yeast; *Candida krusei* (a yeast); *Neurospora crassa* (a mold)

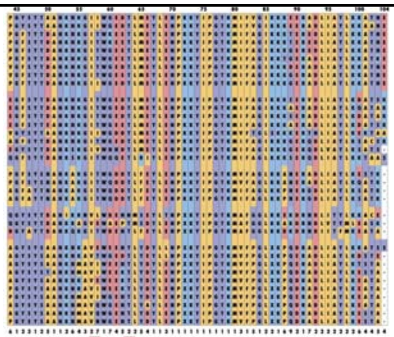
Higher plants: Wheat germ; Buckwheat seed; Sunflower seed; Mung bean; Cauliflower; Pumpkin; Sesame seed; Castor bean; Cottonseed; Abutilon seed

Number of different amino acids

[a] 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 indi-cates that the acetyl group is absent.

Source: After Dickerson, R.E., *Sci. Am.* 226(4): 58–72 (1972), with corrections from Dickerson, R.E., and Timkovich, R., *in* Boyer, P.D. (Ed.), *The Enzymes* (3rd ed.), Vol. 11, pp. 421–422, Academic Press (1975). Table copyrighted © by Irving Geis.

Table 5-5 part 1 Fundamentals of Biochemistry, 2/e



Hydrophilic, acidic: D Asp, E Glu

Hydrophilic, basic: H His, K Lys, R Arg, X TrimethylLys

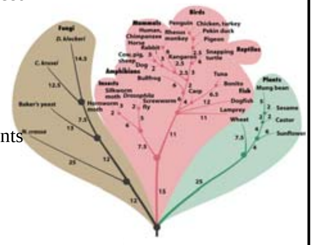
Polar, uncharged: B Asn or Asp, G Gly, N Asn, Q Gln, S Ser, T Thr, W Trp, Y Tyr, Z Gln or Glu

Hydrophobic: A Ala, C Cys, F Phe, I Ile, L Leu, M Met, P Pro, V Val

Table 5-5 part 2 Fundamentals of Biochemistry, 2/e

Phylogenetic tree

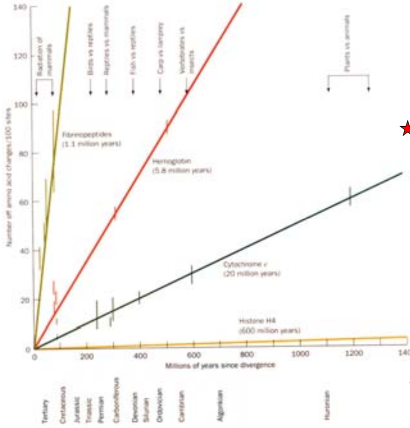
- Indicates the ancestral relationships among the organisms that produced the protein.
- Each branch point indicates a common ancestor.
- Relative evolutionary distances between neighboring branch points are expressed as the number of amino acid differences per 100 residues of the protein.



PAM units

or

Percentage of **A**ccepted **M**utations



PAM values differ for different proteins.

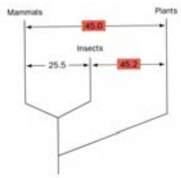
★ ★

Although DNA mutates at a assumed constant rate. Some proteins cannot accept mutations because the mutations kill the function of the protein and thus are not viable.

★ ★

Mutation rates appear constant in time

Although insects have shorter generation times that mammals and many more numbers of replication, number of mutations appear to be independent of the number of generations but dependent upon time



Cytochrome c amino acid differences between mammals, insects and plants note the similar distances

Evolution through gene duplication

Many proteins within an organism have sequence similarities with other proteins.

•These are called gene or protein families.

•The relatedness among members of a family can vary greatly.

•These families arise by gene duplication.

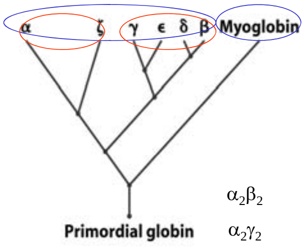
•Once duplicated, individual genes can mutate into separate genes.

•Duplicated genes may vary in their chemical properties due to mutations.

•These duplicate genes evolve with different properties.

•Example the globin family.

Genealogy of the globin family



$\alpha_2\beta_2$

$\alpha_2\gamma_2$

$\alpha_2\delta_2$

$\zeta_2\varepsilon_2$

Hemoglobin:

- is an oxygen transport protein
- it must bind and release oxygen as the cells require oxygen

Myoglobin:

- is an oxygen storage protein
- it binds oxygen tightly and releases it when oxygen concentrations are very low

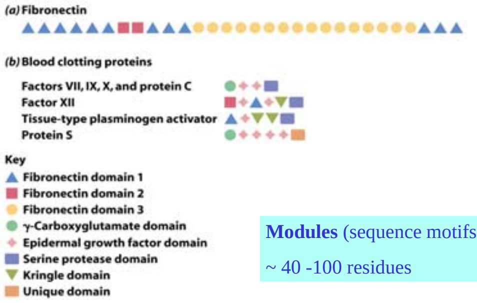
The globin family history

1. **Primordial globin gene acted as an Oxygen-storage protein.**
2. **Duplication occurred 1.1 billion years ago.**

   lower oxygen-binding affinity, monomeric protein.
3. **Developed a tetrameric structure two α and two β chains increased oxygen transport capabilities. ($\alpha_2\beta_2$).**
4. **Mammals have fetal hemoglobin with a variant β**

   chain i.e. γ ($\alpha_2\gamma_2$).
5. **Human embryos contain another hemoglobin ($\zeta_2\varepsilon_2$).**
6. **Primates also have a δ chain with no known unique function. ($\alpha_2\delta_2$).**

Modular Construction of some 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

Modules (sequence motifs):

~ 40 -100 residues

Lecture 8

(9/17/2009)

Chapter 6 - Proteins: 3-D structure

6-1. Secondary Structure