

BMOL2201/6201 PRACTICAL 3

Computational Biochemistry

AM	PM	Outline
10:05	2:05	Introductory talk
10:15	2:15	(Pre-lab Quiz on iLearn)
		1. Genetic Code
		2. Identification of an Unknown Protein
		3. Protein Structure and Function Overview
		4. Exploring α Helices
		5. Exploring β Sheets, Loops, and Turns
		6. Protein sequence and Cancer
12:55	4:55	Prepare to leave lab

Practical 3 Aims

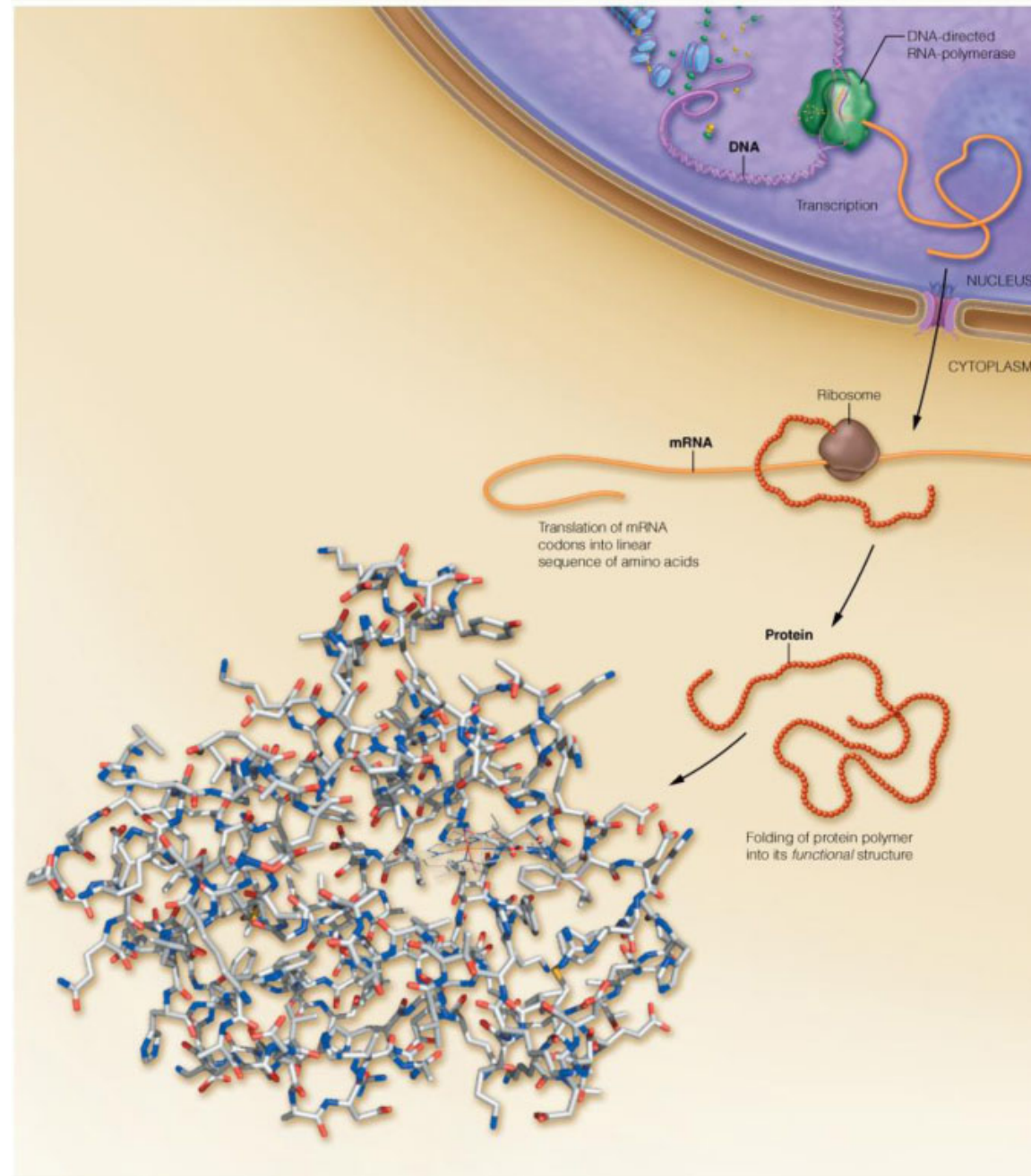
1. Genetic Code: DNA to RNA to Protein
2. Identification and Investigation of an Unknown Protein
3. Investigate protein secondary structure, using PDB's 3D Viewer
 - Visualising an α -Helix
 - Visualising β Sheets, Turns and Loops
4. Analysing a Protein linked to Cancer
 - *Associated Lectures for revision:*
 - Lectures 3-6.

More about Prac 3

- Pre-lab Quiz 3 on iLearn
- *No Prac 3 Data file*
- *No Prac 3 Quiz*
- *No lab coat/safety glasses/gloves!*
- Computational Biochemistry using Pearson Mastering accessible from iLearn
 - Answer the questions in the Prac for P3 marks!
- Bring your own laptop/iPad if you wish!

1. Genetic Code

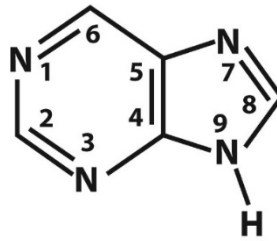
- Genes are the coding parts of DNA
- DNA is transcribed to RNA
- RNA is translated to Protein by the ribosome



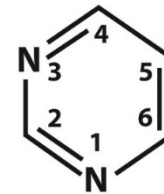
Nucleotide bases in DNA and RNA

Purines

- A: Adenine
- G: Guanine



Purine



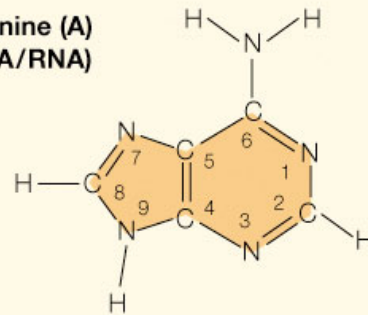
Pyrimidine

Pyrimidines

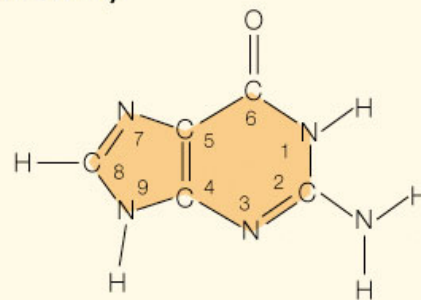
- C: Cytosine
- U: Uracil (RNA)
- T: Thymine (DNA)

PURINES

Adenine (A)
(DNA/RNA)

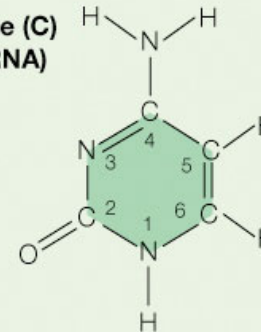


Guanine (G)
(DNA/RNA)

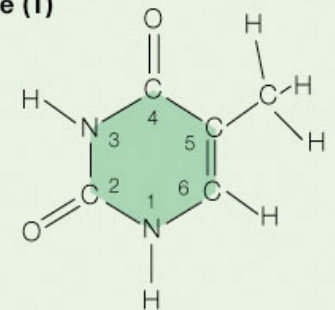


PYRIMIDINES

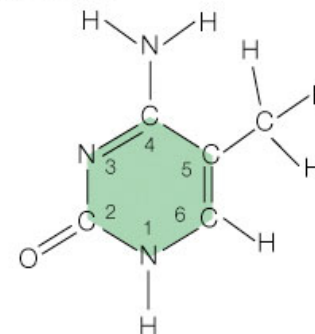
Cytosine (C)
(DNA/RNA)



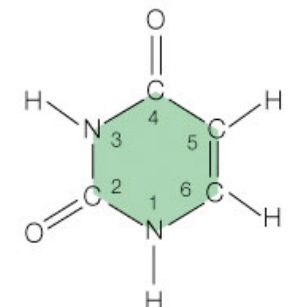
Thymine (T)
(DNA)



5-Methylcytosine



Uracil (U)
(RNA)



Regions of DNA called Genes direct protein synthesis



Unnumbered 3 p50
© 2013 John Wiley & Sons, Inc. All rights reserved.

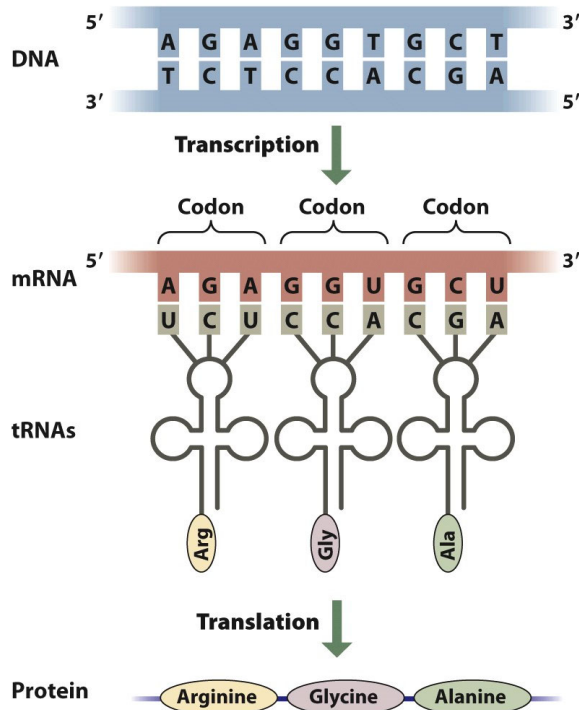


Figure 3-12
© 2013 John Wiley & Sons, Inc. All rights reserved.

Translation occurs in a special organelle called the ribosome

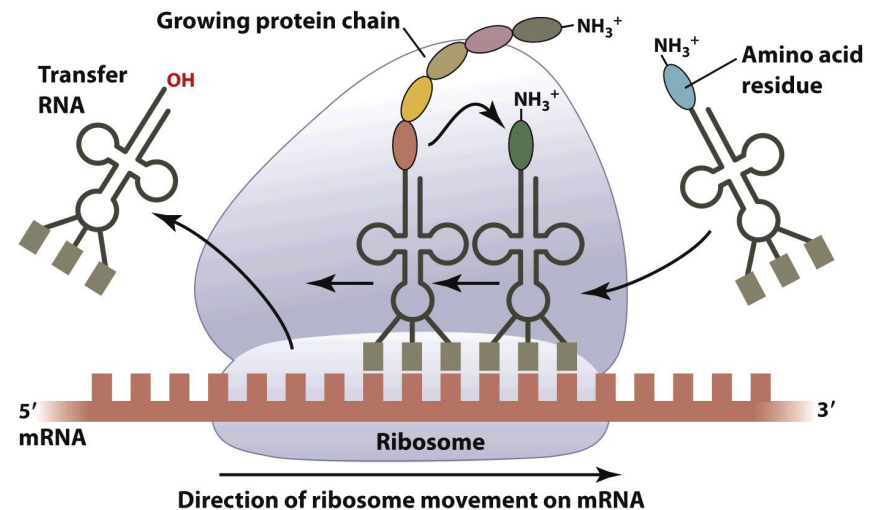
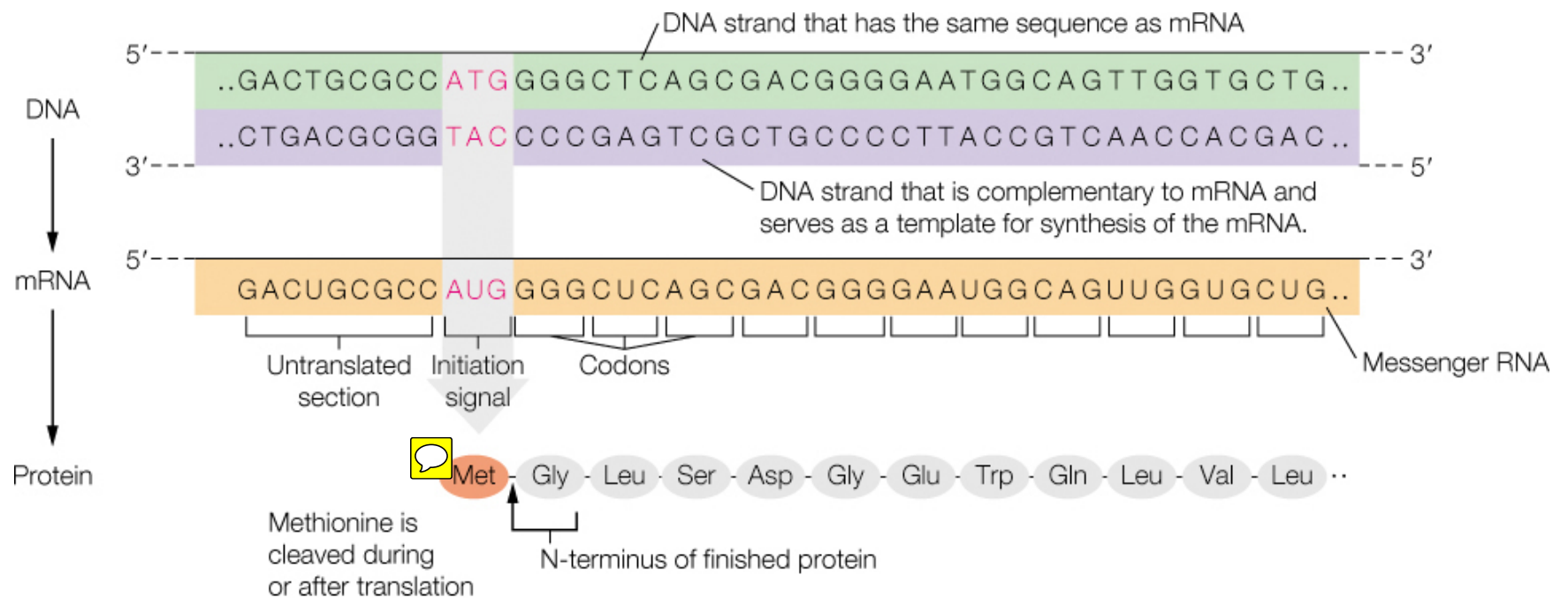


Figure 3-13
© 2013 John Wiley & Sons, Inc. All rights reserved.

E.g. Making myoglobin

- The myoglobin gene is transcribed into an mRNA which is translated into protein



Decoding DNA/RNA via the Genetic Code

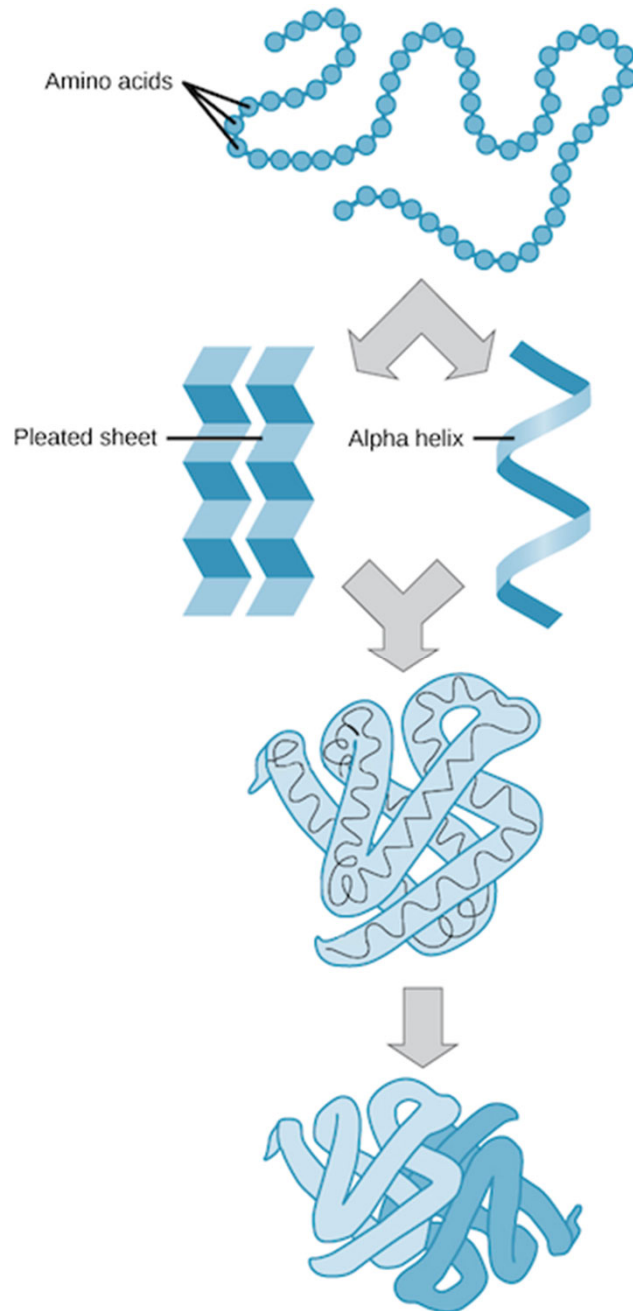
- AUG is the start codon in eukaryotes
 - codes for Met (M)
- There are three stop codons
 - Don't code for anything
- The other 60 codons make the remaining 19 amino acids

		Second position				
		U	C	A	G	
First position (5' end)	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G

Introduction:

Protein structure recap

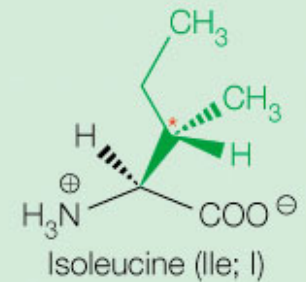
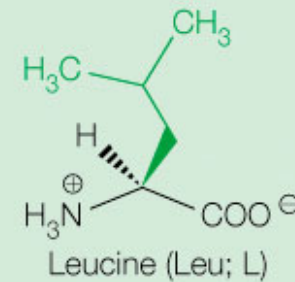
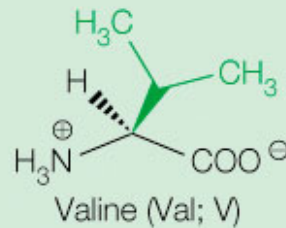
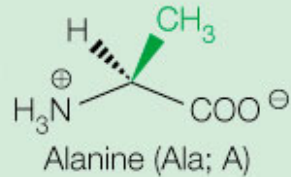
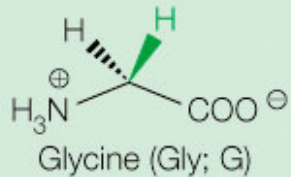
Protein structure – definition of terms



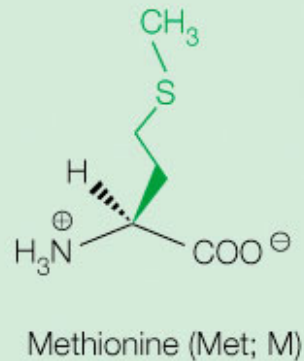
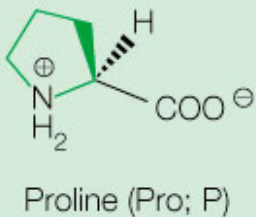
1. Primary structure
 - Sequence of a chain of amino acids
2. Secondary structure
 - Structural elements in proteins that are primarily formed through peptide backbone interactions
 - H-bonding causes amino acids to fold into a repeating pattern
3. Tertiary structure
 - The overall three-dimensional structure of a protein
4. Quaternary Structure
 - Arrangement of subunits within a multisubunit protein

Classification of Naturally Occurring Amino Acids (1 of 2)

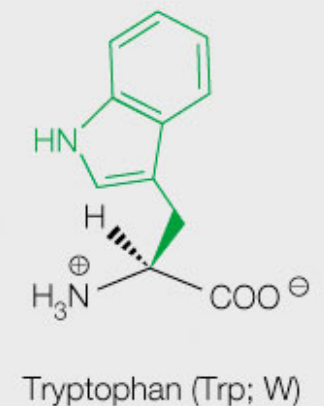
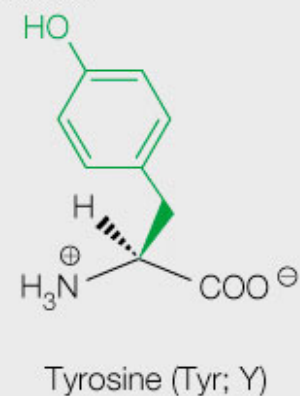
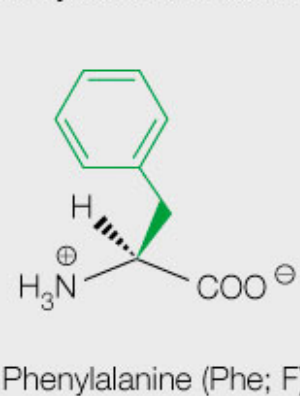
Nonpolar Aliphatic Amino Acids



Nonpolar Amino Acids



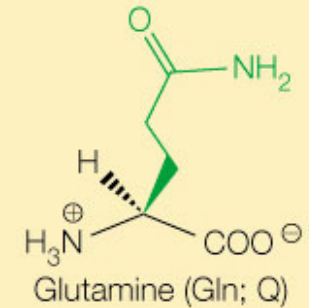
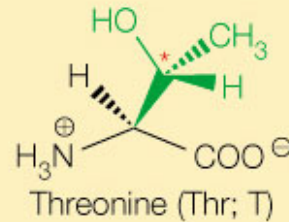
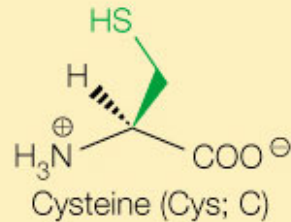
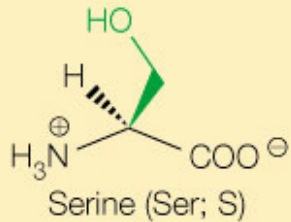
Nonpolar Aromatic Amino Acids



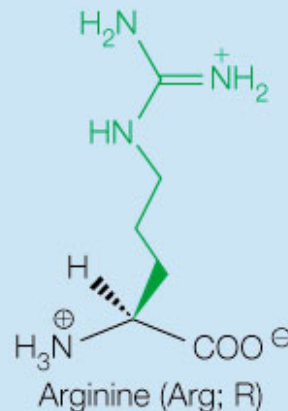
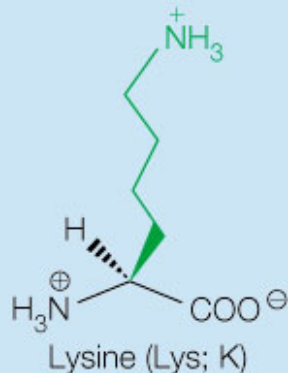
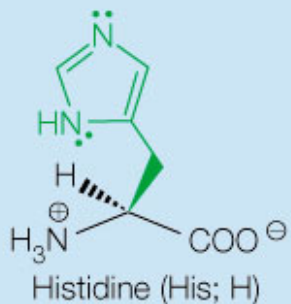
Continue

Classification of Naturally Occurring Amino Acids (2 of 2)

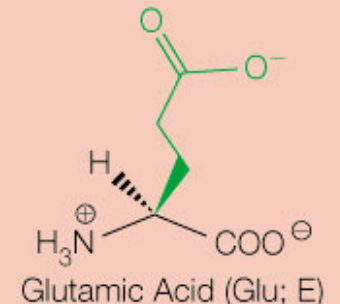
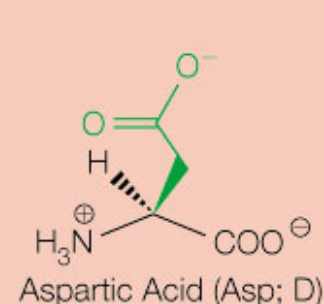
Polar Amino Acids



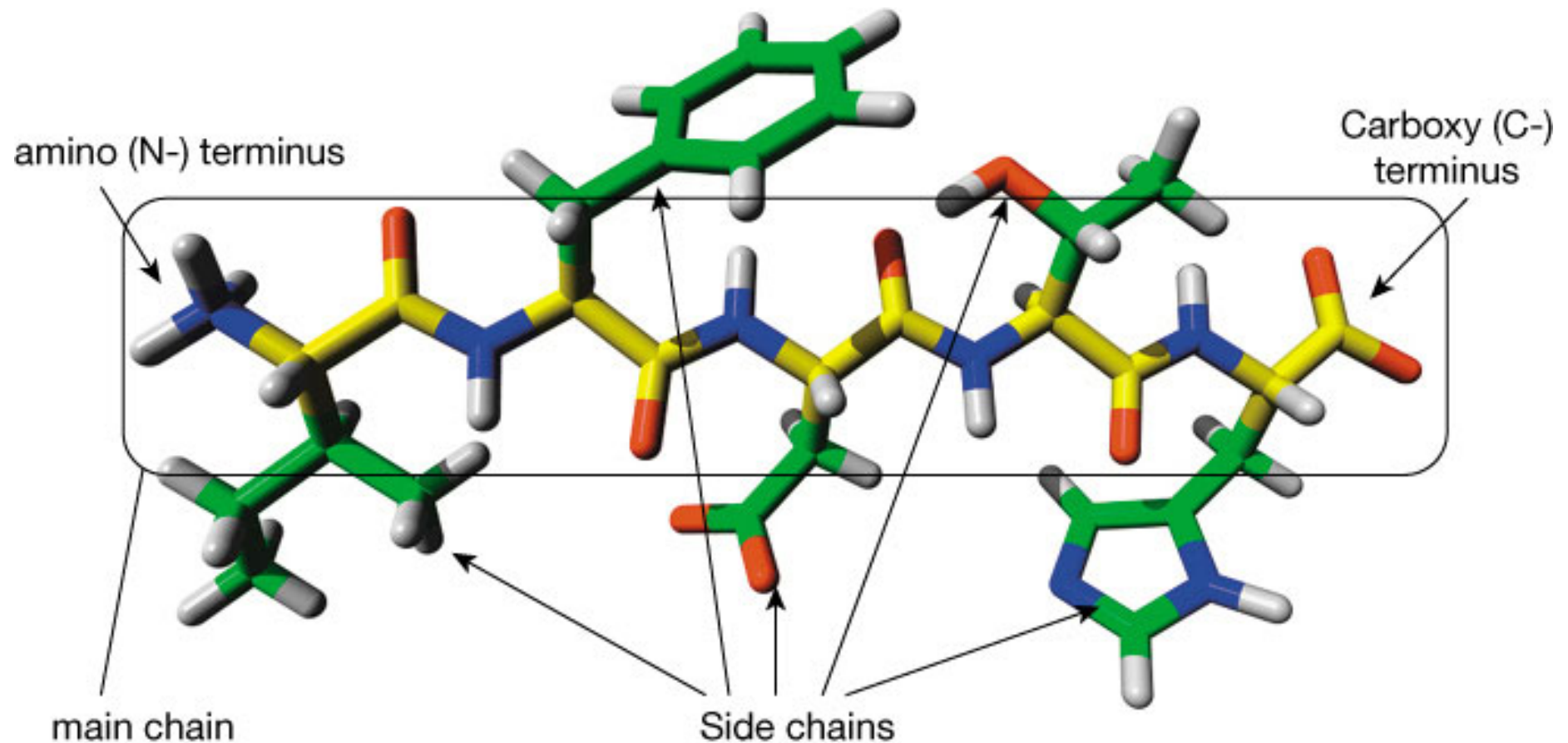
Positively Charged Polar Amino Acids



Negatively Charged Polar Amino Acids



Important Peptide Regions



Secondary structures in proteins

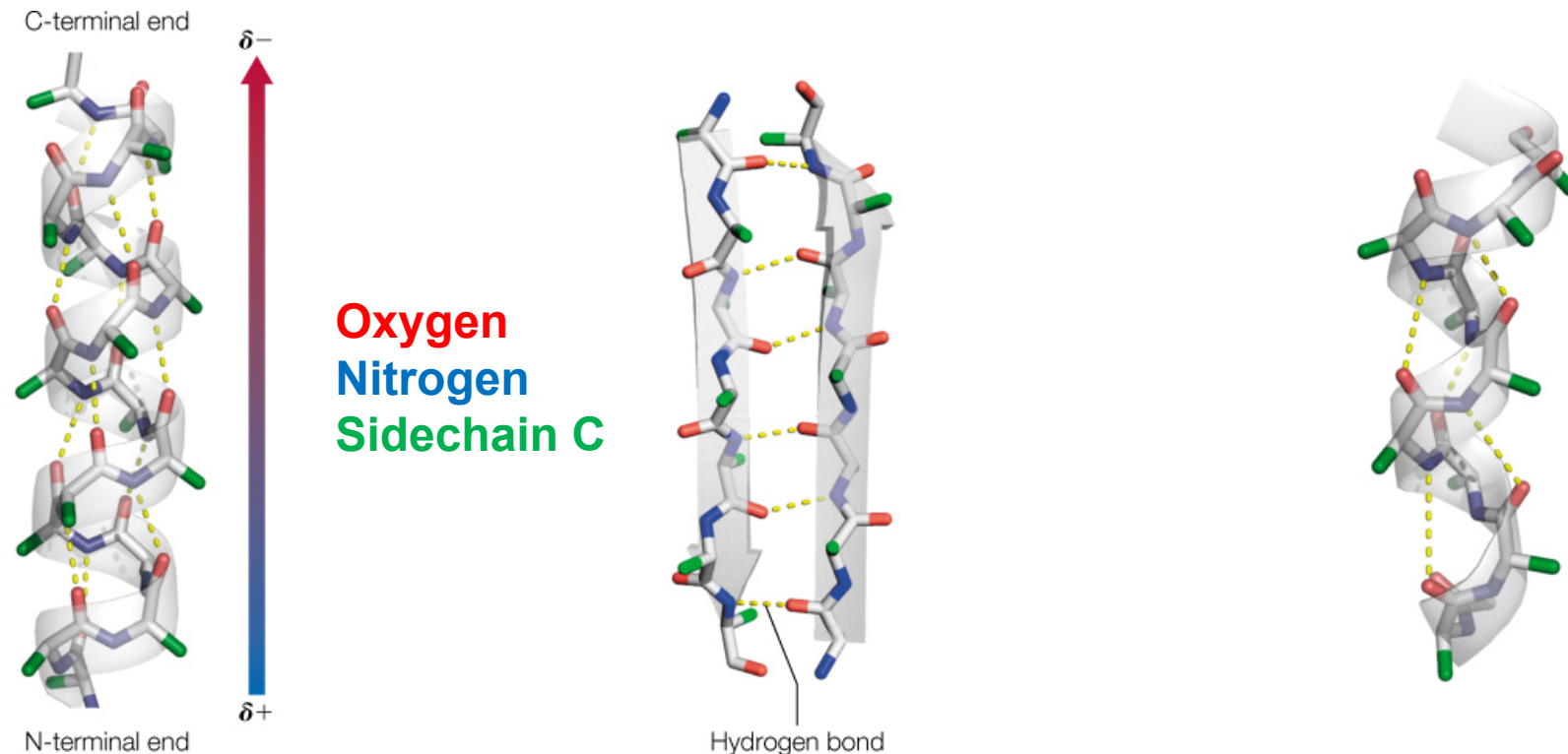
Figure 6.4

The right-handed α helix,

β sheet,

and

3_{10} helix.



(a) In the α helix, the hydrogen bonds are within a contiguous stretch of amino acids and are almost parallel to the helix axis. This orientation of the amide bonds in the helix gives rise to a helical macrodipole moment shown by the arrow (see Figure 2.5). The N-terminal end of the helix has partial (+) charge character, and the C-terminal end has partial (−) charge character.

(b) In the β sheet, the hydrogen bonds are between adjacent strands (only two strands are shown here), which are not necessarily contiguous in the primary sequence. In this structure, the hydrogen bonds are nearly perpendicular to the chains. Note that in the cartoon rendering a strand is shown by a flat arrow, where the head of the arrow points to the C-terminus of the strand.

(c) The 3_{10} helix is found in proteins but is less common than the α helix. Note that, compared to the α helix, the 3_{10} helix forms a tighter spiral.

Tertiary structure of globular proteins: Representations of 3D Structures

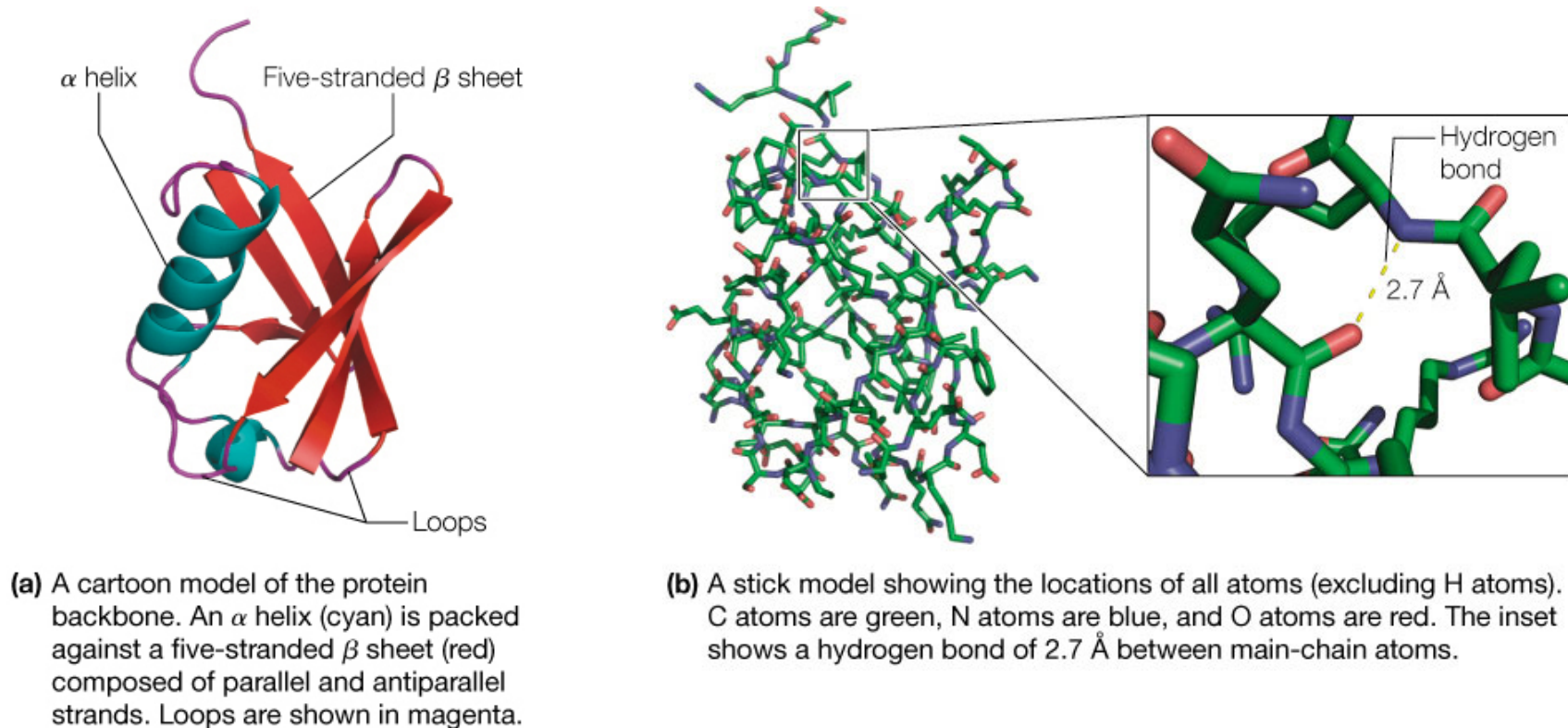
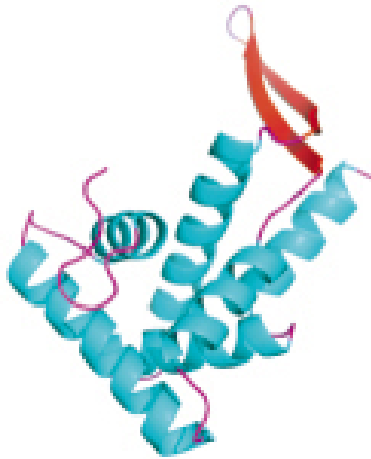


FIGURE 6.15 The structure of human ubiquitin.

Classification of Protein Structure by Domains

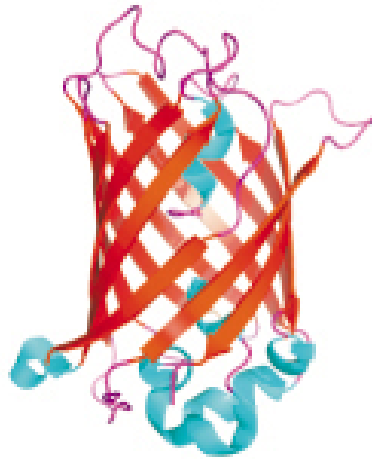
Domain
Class:

“Mainly α ”



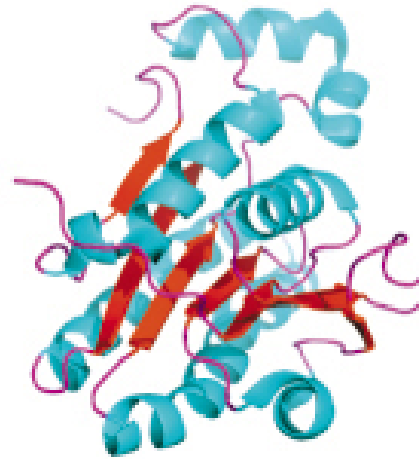
1RSS

“Mainly β ”



2AWK

“ $\alpha + \beta$ ”



1UZM

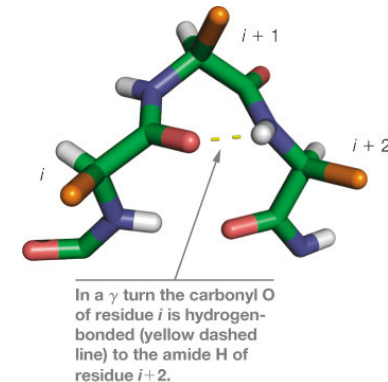
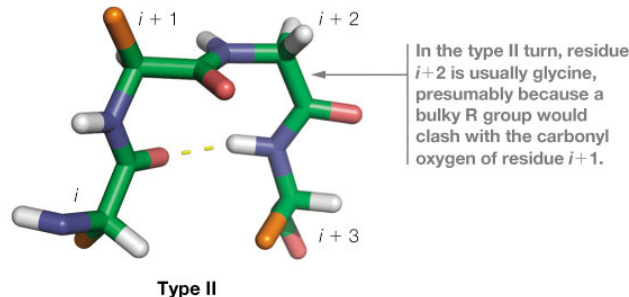
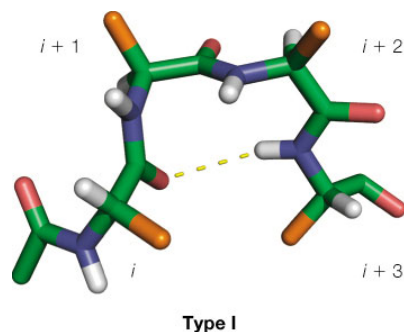
“Few secondary structures”



1JFW

Common Features of Folded Globular Proteins

- Globular proteins have a nonpolar (hydrophobic) interior and a more hydrophilic exterior
- β -sheets are usually twisted or wrapped into barrel structures
- The polypeptide chain can turn corners, for example, β -turns (type I and II left) or γ -turns (right)



Oxygen
Nitrogen
Carbon

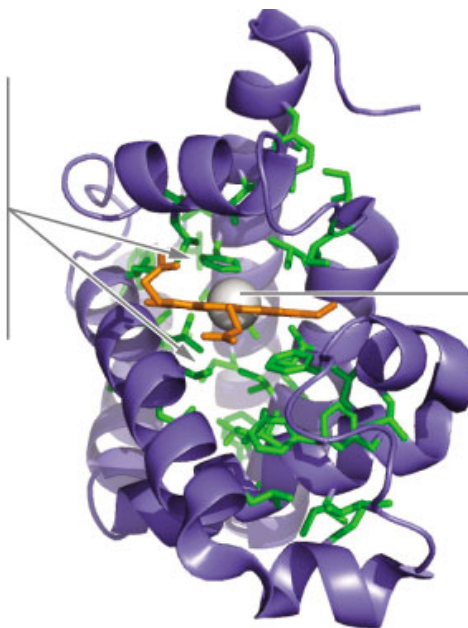
Common Features of Folded Globular Proteins

Distribution of hydrophobic and hydrophilic residues in myoglobin

Hydrophobic residues are **green**, hydrophilic ones are **magenta**, and ambivalent ones are **black**

VLSEGEWQLV LHWAKVEAD VAGHGQDILI RLFKSHPETL EKFDRLFHLK
TEAEMKASED LKKHGVTVLT ALGAILKKKG HHEAELKPLA QSHATKHKIP
IKYLEFISEA IIHVLHSRHP GDFGADAQGA MNKALELFRK DIAAKYKELG
YQG

Hydrophobic side chains (shown in green) cluster about the hydrophobic heme cofactor (orange with iron ion in gray) and on the inside of the molecule.



Heme group with iron

Hydrophilic side chains (red) tend to lie on the solvent-exposed surface of the protein.

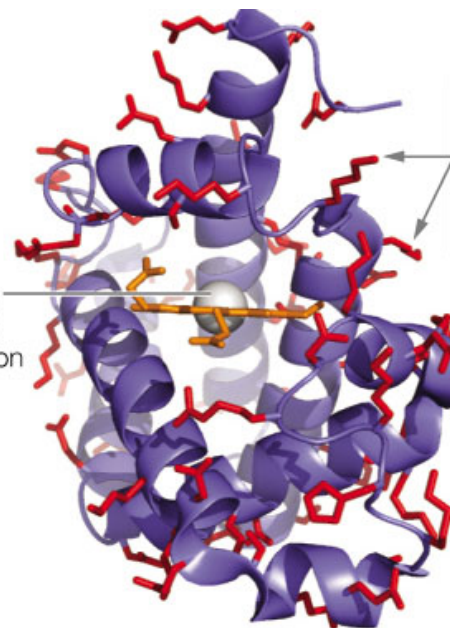
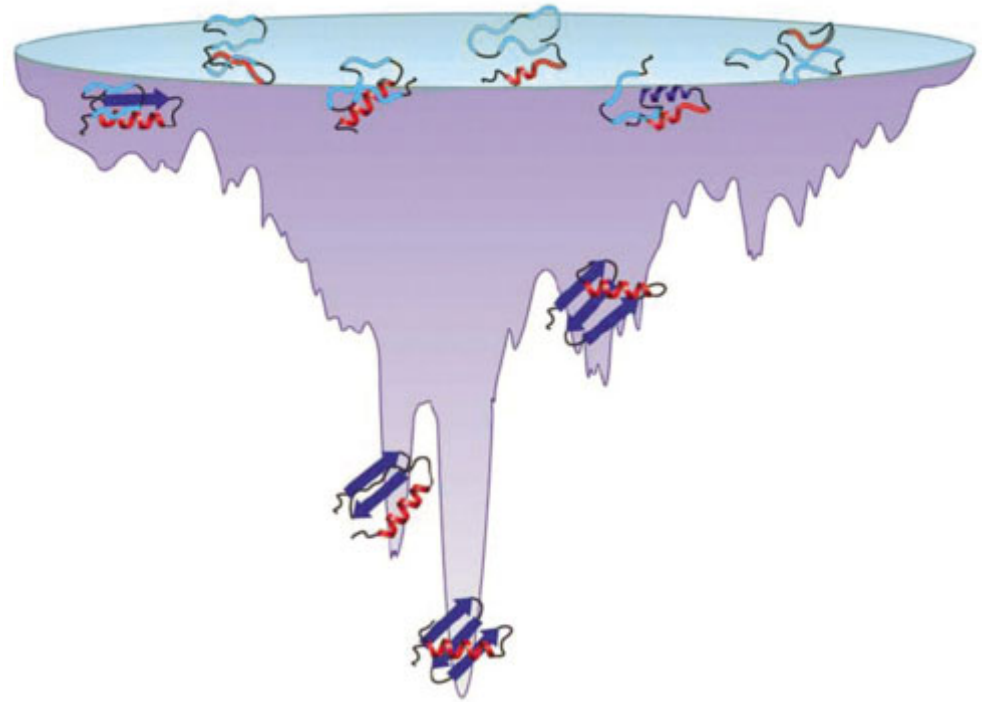


FIGURE 6.19 The distribution of hydrophilic and hydrophobic residues in globular proteins. **Hydrophobic residues** prefer to be in the interior of the protein, while **hydrophilic residues** prefer to be at the solvent-exposed surface of the protein.

Protein folding is a thermodynamic process

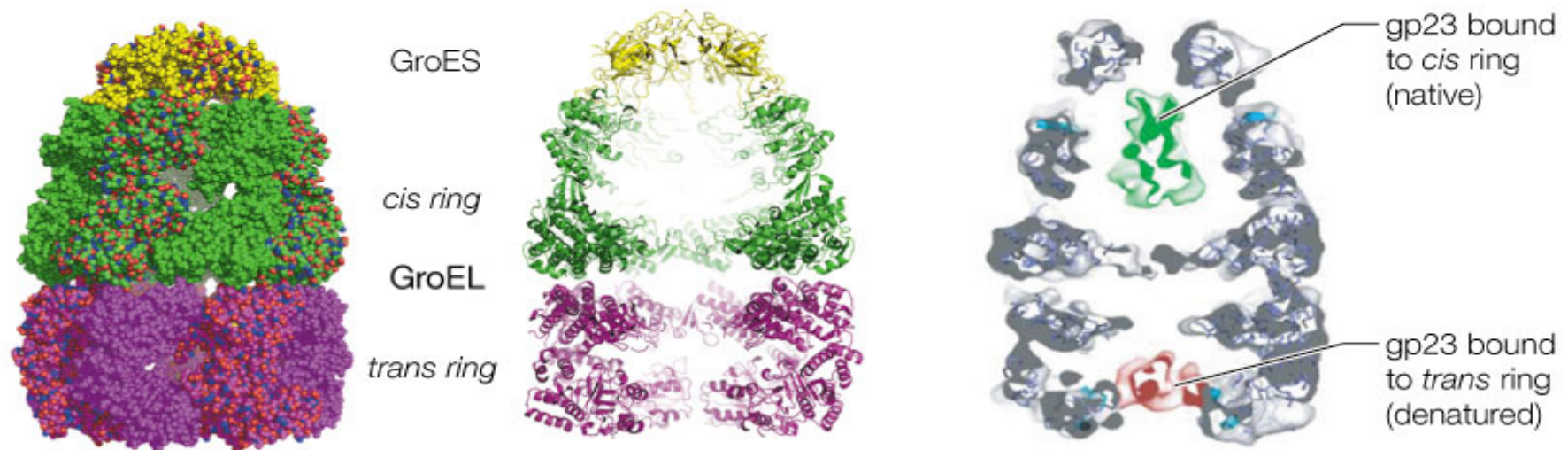
- The protein fold is critical for its function.
- As the protein tries to fold, several intermediate states are possible.
- Final folded state is the most stable one.
- Unfolded or misfolded proteins are non-functional and may aggregate and cause diseases.



(d) A more rugged energy landscape showing local minima for metastable intermediate states. This landscape describes a multistate folding process.

Chaperonins help to fold misfolded proteins

- GroEL (2x7 subunits) - GroES (7 subunits) complex:
- Hydrophobic cavity where misfolded proteins can unfold and re-fold properly



(b) Space-filling side view of the chaperonin, colored as in (a).

(c) Cartoon view showing the enclosed cavity formed by the *cis* ring of GroEL and GroES (top), as well as the compaction of the *trans* ring compared to the *cis* ring (bottom).

(d) Electron density map obtained from cryoelectron microscopy of a chaperonin complexed with bacteriophage T4 coat protein gp23. The gp23 bound to the *trans* ring of GroEL is shown in red and appears to be denatured. The gp23 bound to the *cis* ring is shown in green and appears to have a native-like conformation.

Clues to a Protein's Function from its sequence: enter BLAST

- We now have primary structures for:
 - 69,082,330 non-redundant nucleotide sequences
 - 370,455,262 non-redundant protein sequences
- It is thus possible that an unknown protein may have similarities to known proteins stored in sequence databases.
- To search for similar sequences, we use BLAST



BLAST Homepage and Selected Search Pages

Introducing the BLAST homepage and form elements/functions of selected search pages

<https://blast.ncbi.nlm.nih.gov>

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

- The Basic Local Alignment Search Tool (**BLAST**) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

BLAST can do many things: we will carry out Protein BLAST

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

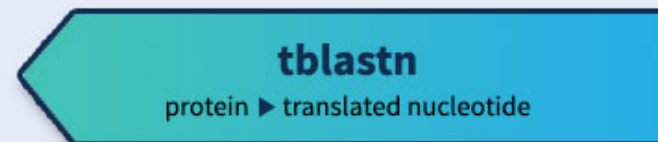
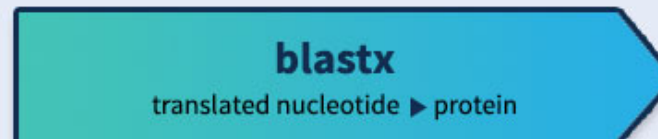
A new feature was added to Primer-BLAST.

We now offer the ability for user to run primer-blast from NCBI assembly page..

Tue, 23 Feb 2021 12:00:00 EST

[More BLAST news...](#)

Web BLAST

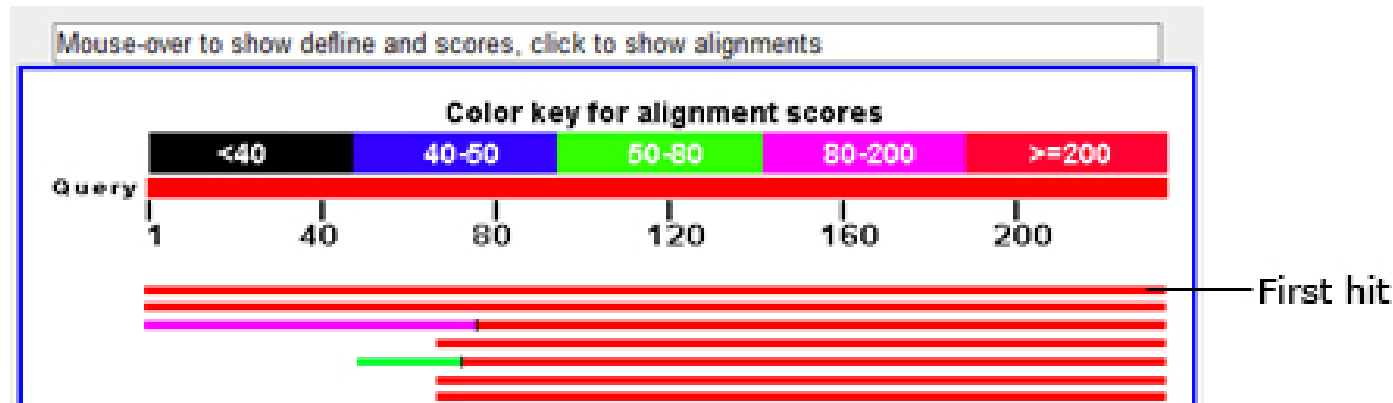


BLAST Genomes

[Human](#)[Mouse](#)[Rat](#)[Microbes](#)

BLAST results

- A graphical view of similar sequences



- A list of the top matches (aka hits):

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input type="checkbox"/>	myoglobin [synthetic construct]	synthetic construct	313	313	100%	1e-107	100.00%	155	AAX36993.1
<input type="checkbox"/>	myoglobin isoform 1 [Homo sapiens]	Homo sapiens	312	312	100%	2e-107	100.00%	154	NP_001349775.1
<input type="checkbox"/>	myoglobin isoform CRA_a [Homo sapiens]	Homo sapiens	313	313	100%	2e-107	100.00%	166	EAW60065.1
<input type="checkbox"/>	myoglobin transcript variant 1 [Homo sapiens]	Homo sapiens	311	311	100%	5e-107	99.35%	154	AAX84516.1
<input type="checkbox"/>	myoglobin [Homo sapiens]	Homo sapiens	311	311	100%	6e-107	99.35%	154	AAA59595.1

- **E-value** is the measure of likeliness that sequence similarity is not by random chance – the smaller the better: <e-05 usually considered significant
 - **Percent Identity** describes how similar the query is to the aligned sequences
- Clicking on each link takes you to the alignment with that sequence.

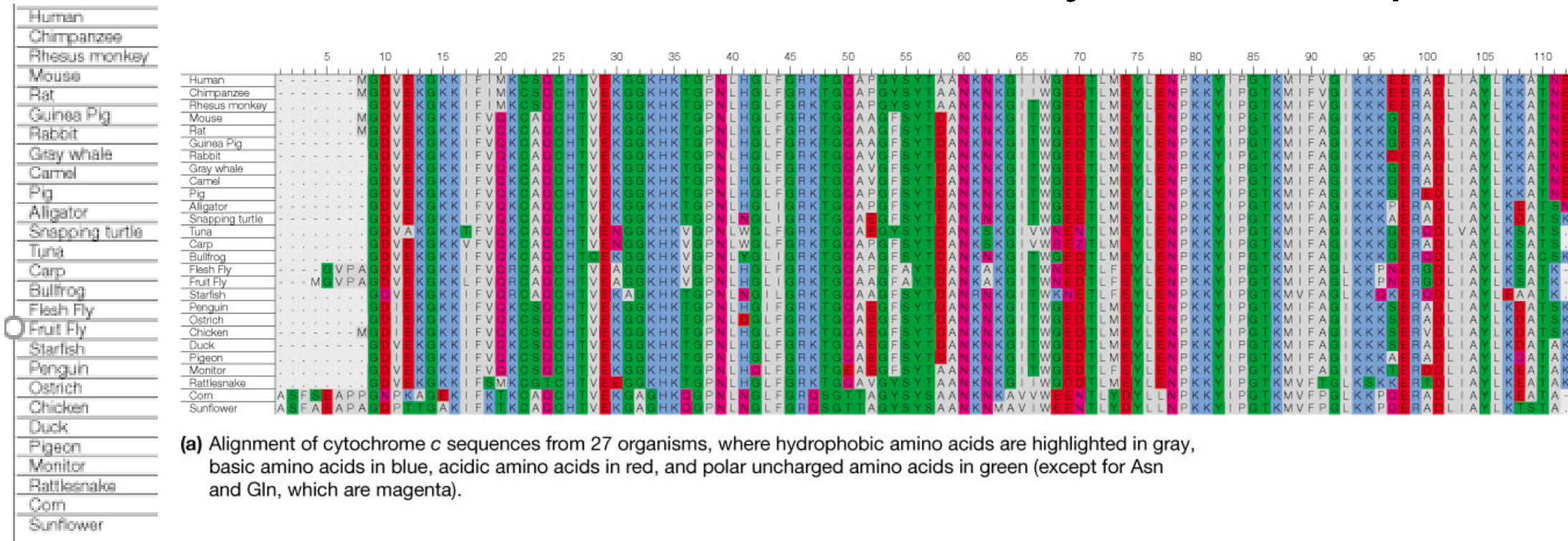
BLAST alignment between human myoglobin and human A globin sequences.

Score = 30.8 bits (68), **Expect = 6e-06**, Method: Compositional matrix adjust.
 Identities = 32/133 (25%), Positives = 48/133 (37%), Gaps = 40/133 (30%)

Human Mb	2	LSDG EWQLVL N WGK VEADIPG HGQ EVLI RL FKGHPETLEK FDK FKHLKSEDEMKASEDL	61
		LS + V WGKV A +G E L R+F P T F F	
Human α	2	LSPADKTN VKAA WGK VG AHAGE YGAE ALERMFLSFPTTKTY FP H-----	46
Human Mb	62	KKHGATV L TALGG I KKKG H HEAEIKPLAQSHATKH KI -----PV KY LEFISE CI IQV LQ SKHP	120
		L+AL I HA K ++ PV +++S C++ L + L	
Human α	47	-----ALSALSDI-----HAHKL R VD P VNF-----KLLSHCLLVTLAAHL P	82
Human Mb	121	G D FGADAQ GAMNK	133
		+F +++K	
Human α	83	A E FTPAVHA SLDK	95

- The sequence of myoglobin (Mb) appears above that for the α globin. Between them are shown the **identical amino acids** (blue) and those that are considered (chemically) **similar** (green "+"). Gaps in the alignment are shown by red text and red dashes.
- In this alignment, there is a 25% sequence identity, suggesting a high degree of structural similarity between the proteins.
- The expect score in this case is very low ($6 * 10^{-6}$), indicating that the alignment is statistically significant.

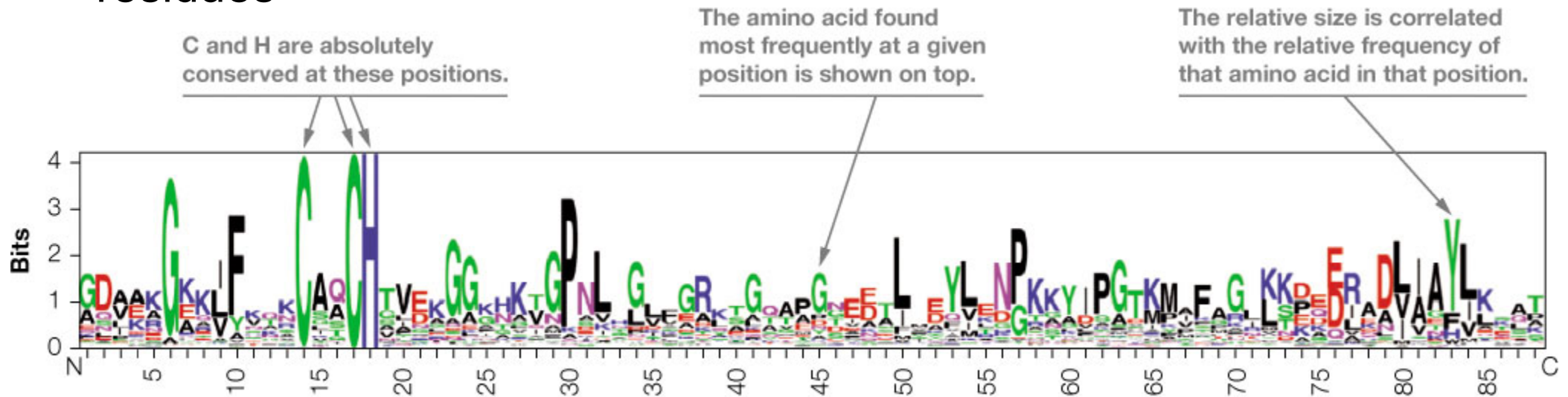
Aligning sequences from different organisms can show conservation and evolutionary relationships



- Sequence alignment of cytochrome c, a mitochondrial respiratory protein, from 27 organisms
- The more similar two homologous protein sequences are, the more closely they are related evolutionarily
- Functionally important proteins show few changes in their primary sequences.

Consensus Sequence from alignment

- A consensus sequence (here for cytochrome *c*) allows an investigator to identify the most highly conserved amino acid residues



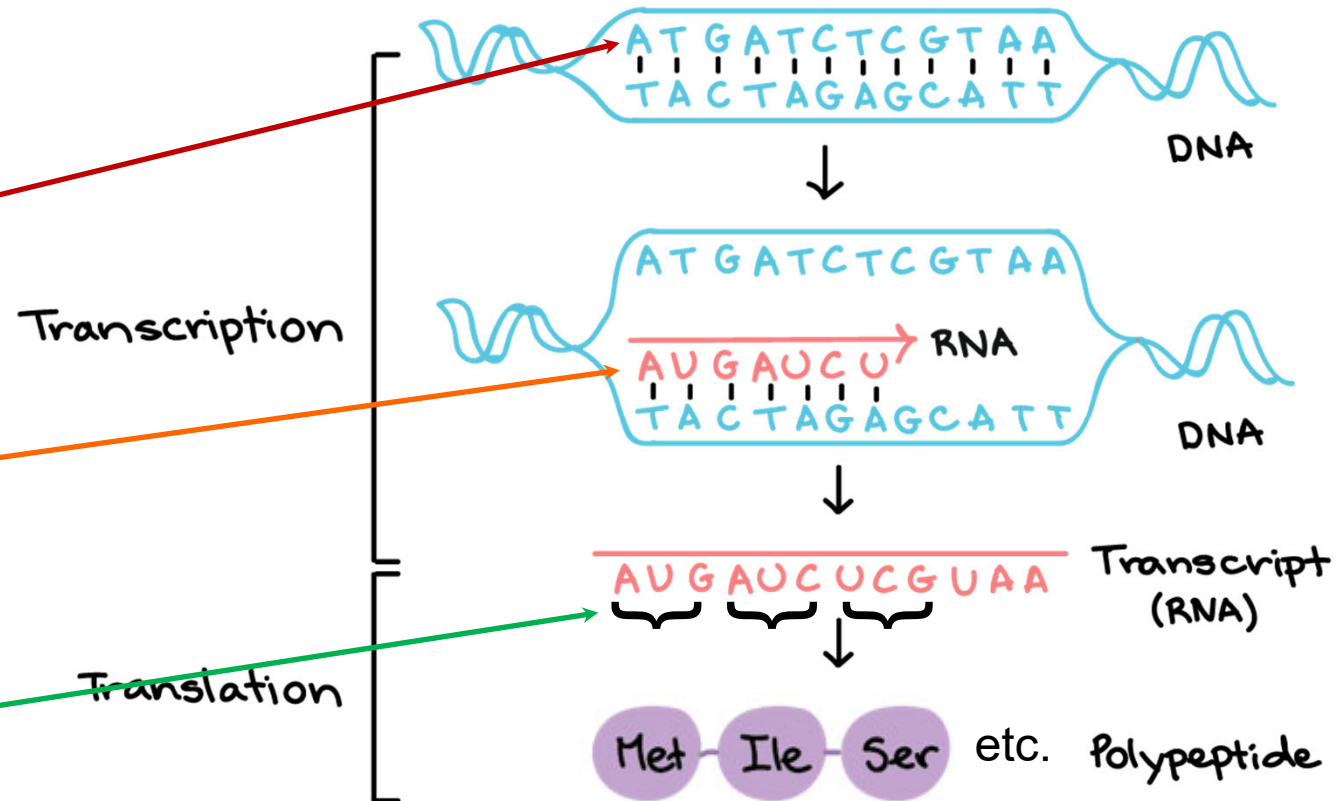
- The most important residues are absolutely conserved across organisms.
- The most highly conserved amino acids tend to be those that serve a critical structural and/or functional role in the protein.
- In enzymes, these are part of the active site!

Prac 3

- Go to iLearn and from then to Pearson Mastering site
- Look for Assignment:
“Prac 3 including Prac 3 Quiz due 3 May 2021”
- Go through the learning activities from Q1-Q6.
- Answer each question as you go
- *There are Hints if you cannot directly answer the questions....however, this may cost you some credit*


Q1: DNA to RNA to Protein:

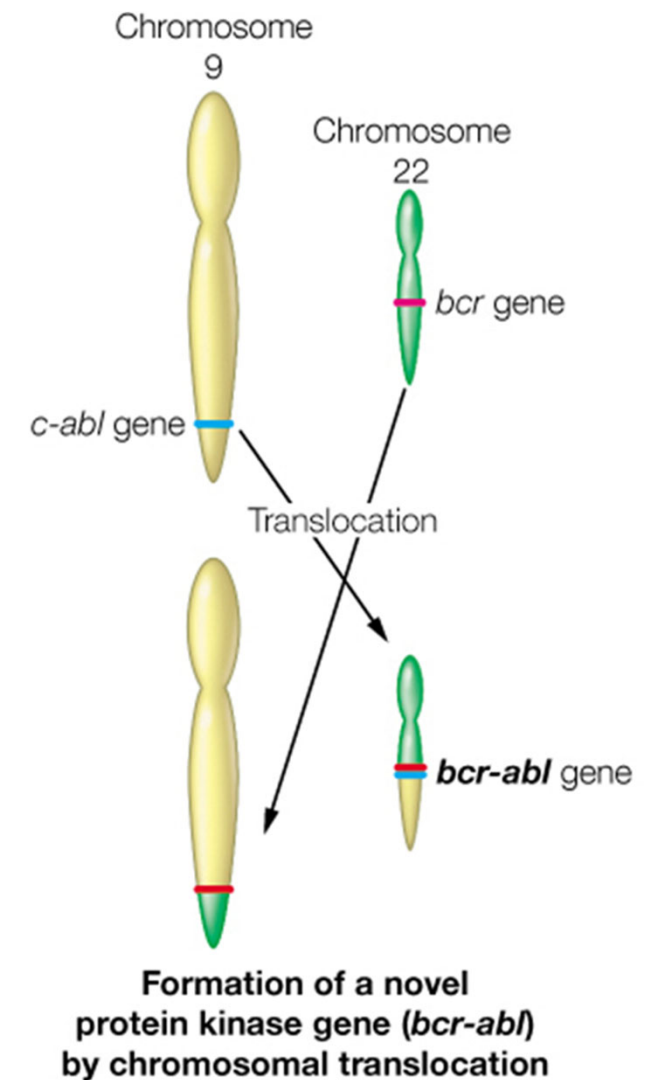
- For a given DNA sequence:
 - Identify the coding strand
 - RNA will be the same, with U replacing T
 - Read 3 RNA bases (i.e. a codon) at a time to get the corresponding amino acids



Q5: Using BLAST: What Can a Protein Sequence Reveal about Cancer?

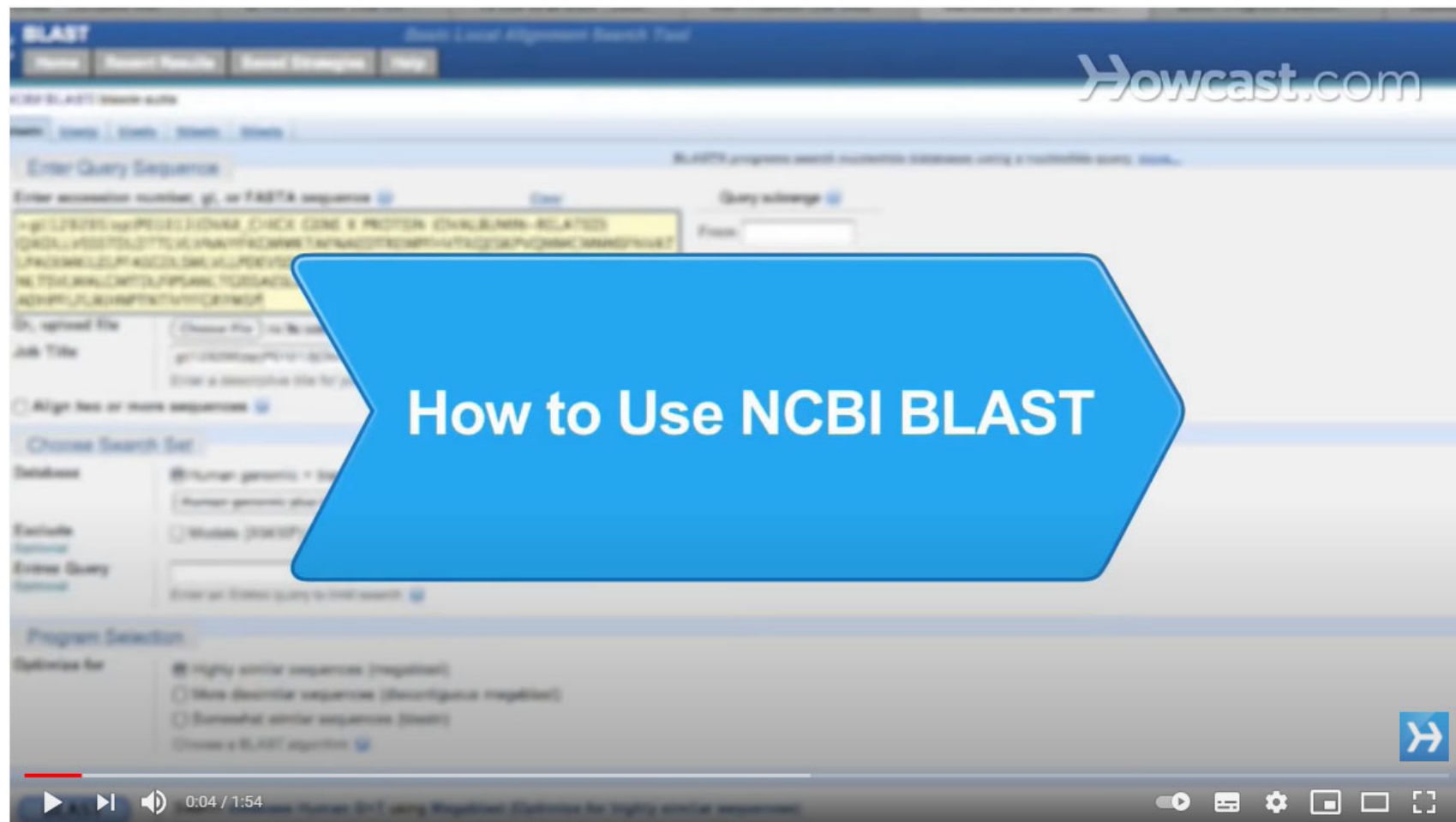
Chromosomal Translocation in Chronic Myelogenous Leukemia (CML)

- CML is due to a chromosomal translocation event
- Portions of genes on Chr 9 and Chr 22 are linked, fusing the parts of two oncogenes (bcr and c-abl)
- Gene fusion yields a novel protein tyrosine kinase (BCR-ABL) that stimulates growth-promoting pathways, as it cannot to switched off! 
- Results in leukemia.
- Running a database search using BLAST provides us the identity of this protein



How to run BLAST with a sequence

- <https://www.youtube.com/watch?v=gKRDe7-l42M>



BLAST Search Instructions

1. In the middle of the page under the Web BLAST heading, click **Protein BLAST**.
2. **Copy the amino acid sequence from above** and **paste it** in the Enter Query Sequence box at the top of the page. (BLAST ignores numbers and spaces, but you may need to delete the number "1" if you receive an error message.)
 - **Here is the amino acid sequence to copy and paste into the box!**

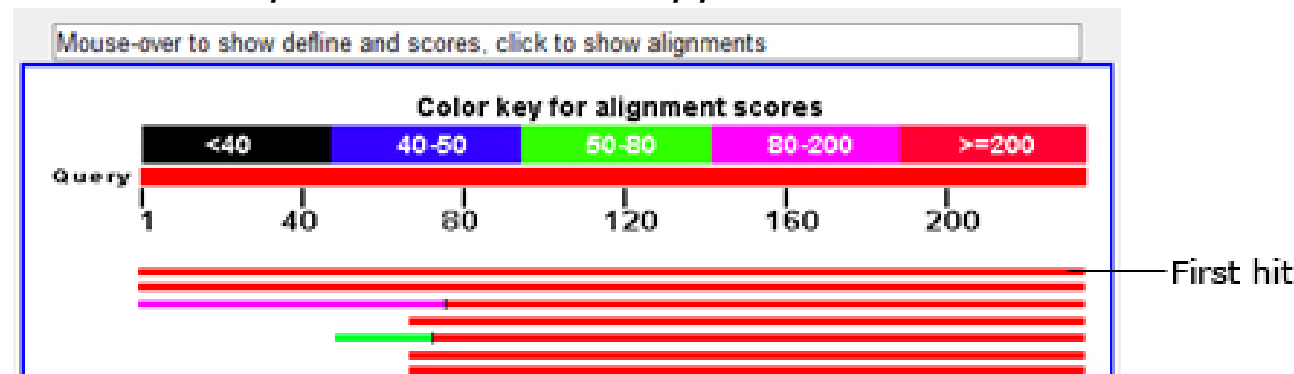
```
1   psmافرvhsr ngksytflis sdyeraewre nireqqkkcf rsfs!tsvel qm!tnscvkl
61  qtvhsiplti nkegeklrv! gynhngewce aqtkngqgwv psnyitpans !ekhswyhgp
121 vsrnaaeyll ssgingsflv resesspgqr sislryegr! yhyrintasd gklyvssesr
181 fntlaelvhh hstvadglit tlhypapk!rn kptvygvspn ydkwemertd itm!kh
```

3. In the text box to the right of Organism, type "Homo sapiens," and then click on **Homo sapiens (taxid:9606)**.
4. In the Program Selection box, choose **blastp (protein-protein BLAST)**.
5. Scroll down and click the **BLAST** button. Wait for BLAST to complete the search. Initial information (Conserved Domains) may appear quickly, but full results may take 30 seconds. *(If your search returns a screen that displays "No significant similarity found," open Hint 1 to see what might have gone wrong.)*

Follow other instructions (from 6.) on Pearson!

BLAST Search Instructions - 2

6. A new screen displays your search results. Briefly scroll down to look at the three main sections, and notice the type of information that is provided in each section. Note that in each of the three sections, similar sequences, or hits, are listed beginning with the best statistical match to your query sequence.
 - The **Graphic Summary** gives a color-based summary of the sequence alignment between your query sequence and the most similar sequences (hits). (For this exercise, you can ignore the conserved domains information at the top.)
 - The **Descriptions** section provides the unique accession number of each hit, a brief description of the sequence, and several “scores” that quantify the similarity between each hit and your query sequence.
 - The **Alignments** section shows the alignment between each hit and your query sequence, amino acid by amino acid.
7. Scroll to the top of the Graphic Summary section, but ignore the conserved domains information. Place your cursor over the top red bar that represents the first hit and click on it. Look at the information in the small text box immediately above the Graphic Summary figure. *For an explanation of what appears in the text box, see Hint 2.*



8. *Follow the other instruction*