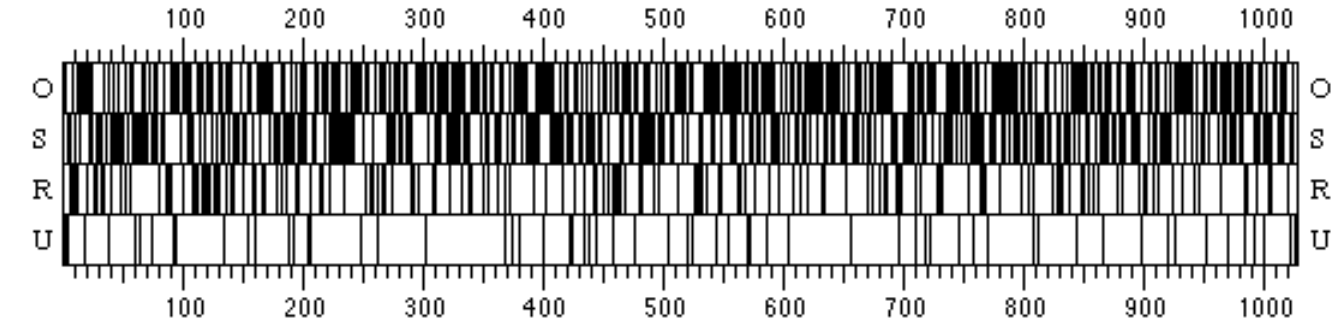
# Prokaryotic Gene Prediction

- Genes: region between *start codon* ATG and *stop codon* (TAA, TAG, or TGA). Absence of introns.
- Codon Bias
- Locate Promoter region
- Ribosome Binding site
- Terminator site

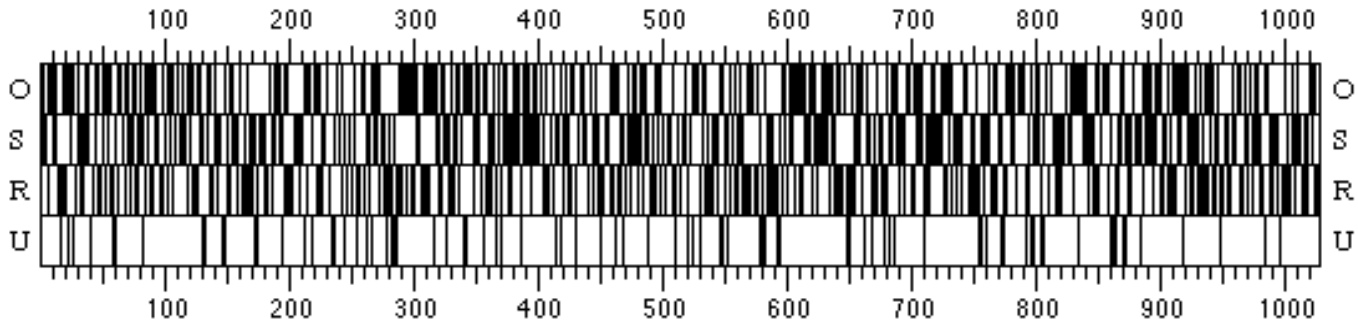
# Codon Bias

- Some codons preferred over others.

O = optimal
S = suboptimal
R = rare
U = unfavorable




Frame Shift 1

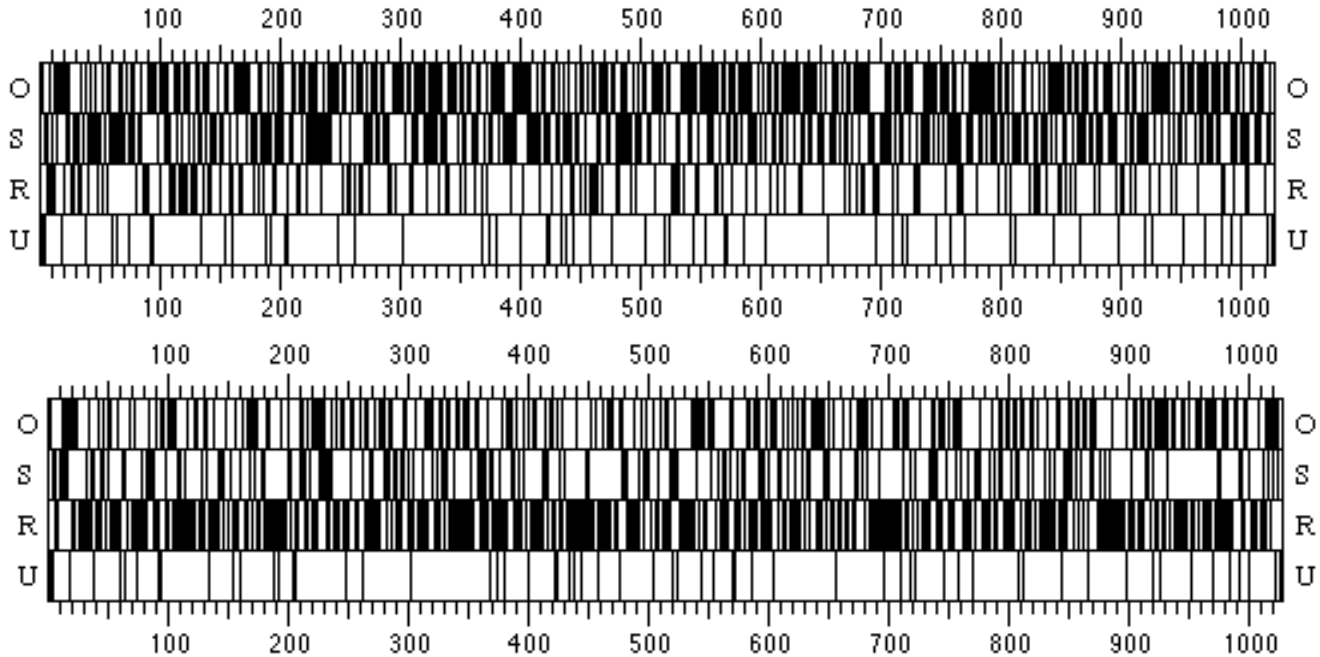



Frame Shift 2

# Codon Bias

- Codon biases specific to organisms

O = optimal
S = suboptimal
R = rare
U = unfavorable



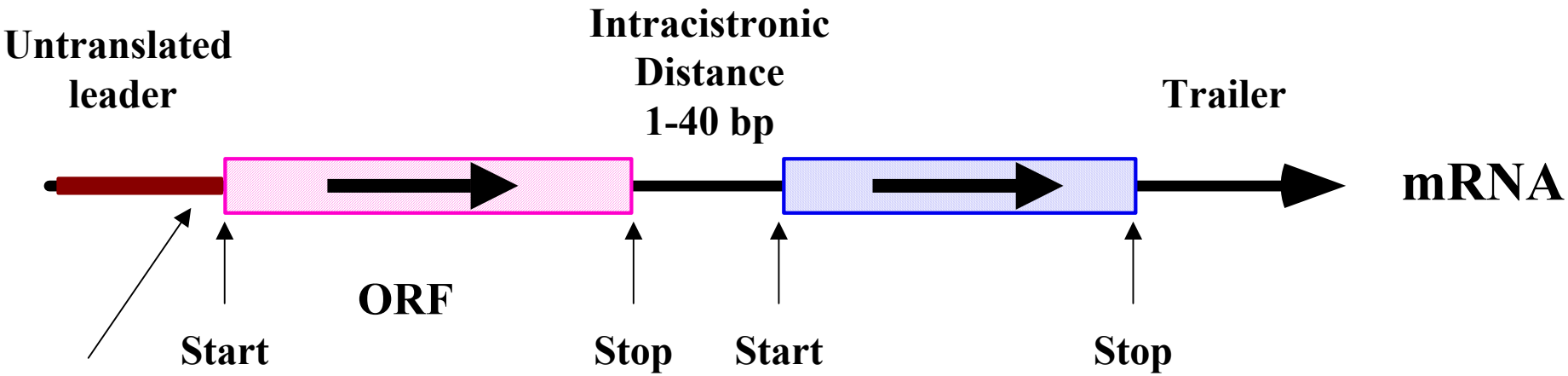

Same Frames; Different labeling of codon types (i.e., from yeast)

# Messenger RNA or mRNA

| Initation Codon | AUG | Methionine |
|---|---|---|

| Termination Codons | |
|---|---|
| UAA | Ochre |
| UAG | Amber |
| UGA | Opal |

| Others: | | |
|---|---|---|
| | GUG | Valine |
| | UUG | Leucine |
| | AUU | Isoleucine |



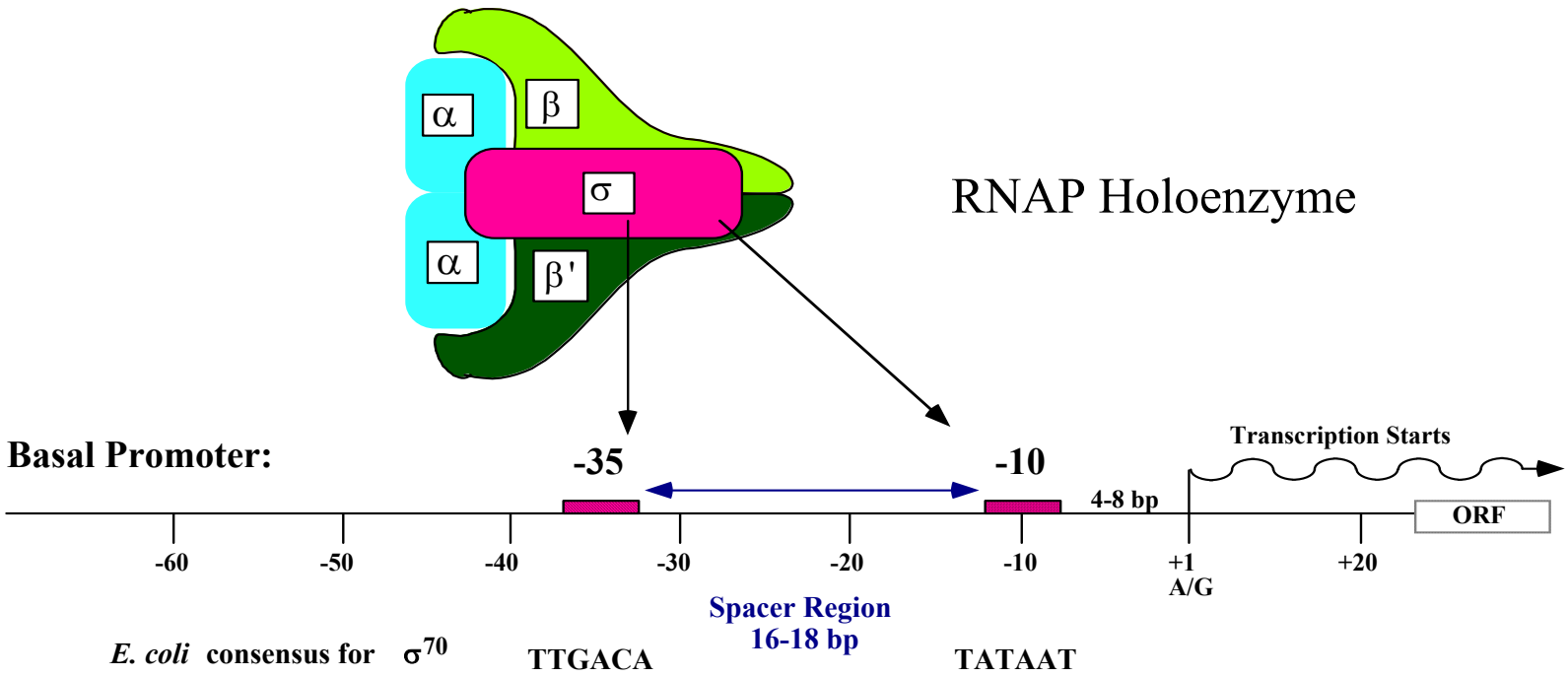

**RBS**
**Ribosome Binding Site**
**Shine-Dalgarno Sequence**

**7 bp upstream of start**
**5'--AGGAGG--3'**

**Coding region**
**Open Reading Frame (ORF)**

**Reading frame** is one of three possible ways of reading a nucleotide sequence as a series of triplets.

# Transcriptional machinery: RNA Polymerase and DNA



# Eukaryotic Gene Prediction

- Complicated by introns & alternative splicing
- Exons/introns have different GC content.
- Many other measures distinguish exons/introns
- Software:
  - GENEPARSER Snyder & Stormo (NN)
  - GENIE Kulp, Haussler, Reese, Eckman (HMM)
  - GENSCAN Burge, Karlin (Decision Trees)
  - XGRAIL Xu, Einstein, Mural, Shah, Uberbacher (NN)
  - PROCRUSTES Gelfand (Formal Languages)
  - MZEF Zhang

# Introns/Exons in *C. elegans*

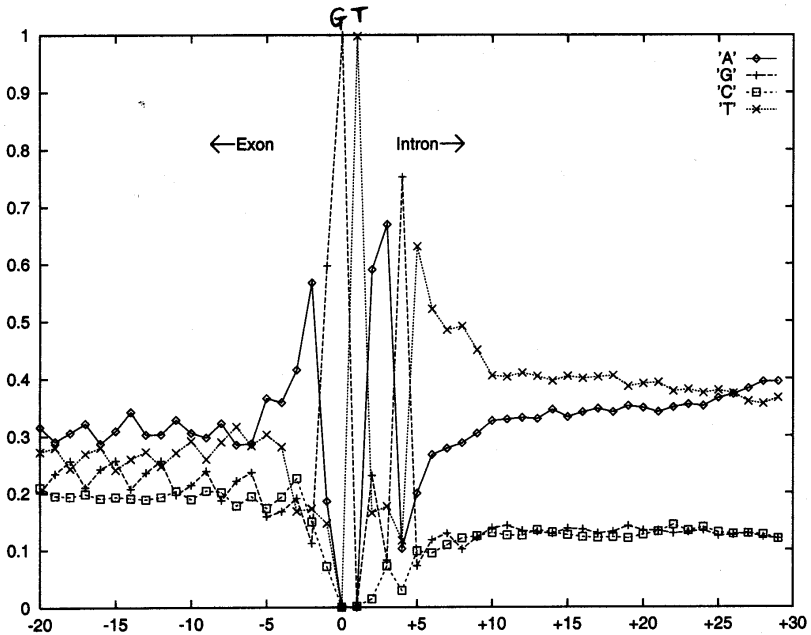



Figure 2: Profile of the same 5' collection but around a larger window.

A/T

G/C

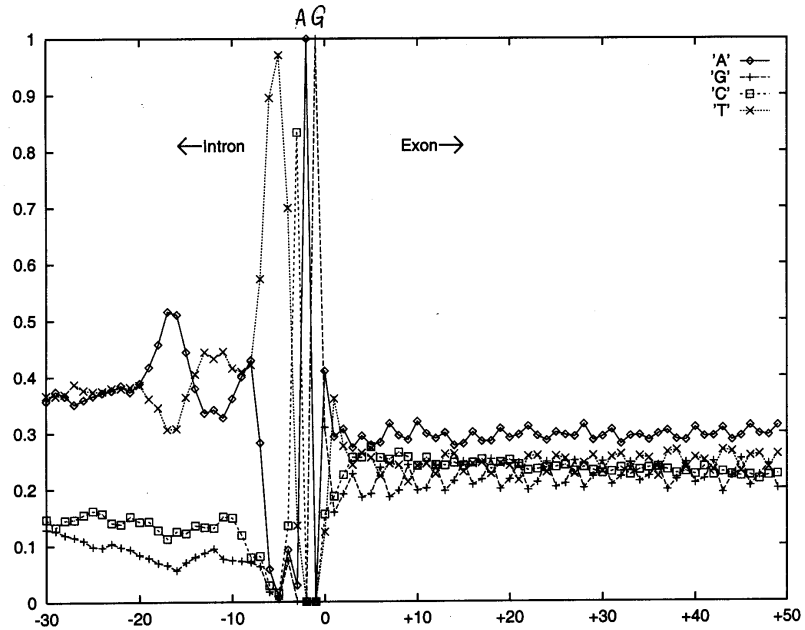
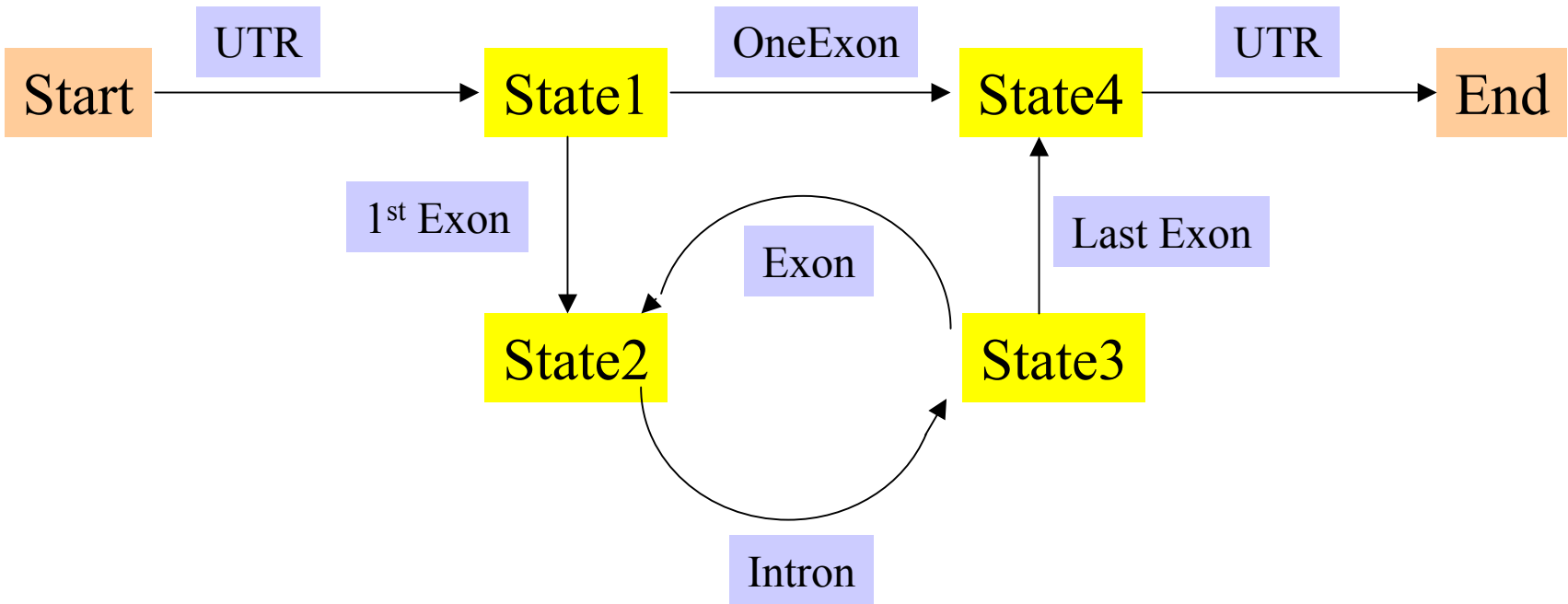



Figure 4: Profile of 8,192 sequences of length 80 around the 3' site. The first position in the exon is labeled 0.

- 8192 Introns in *C. elegans* : **[GT…AG]**
- Vary in lengths from 30 to over 600; Complexity varies

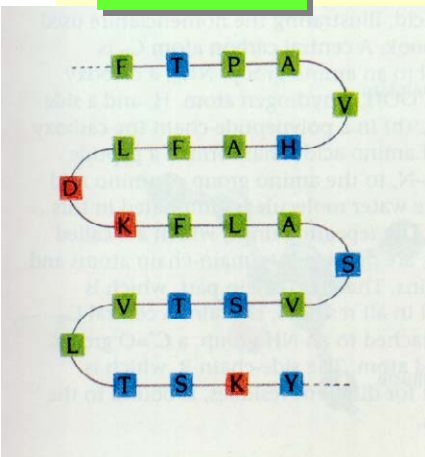
# HMM structure for Gene Finding



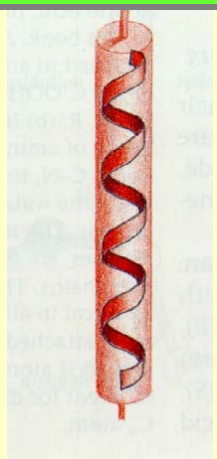
# Protein Structures

- Sequences of amino acid residues
- 20 different amino acids

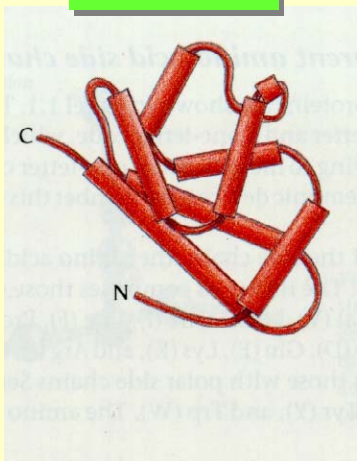
Primary



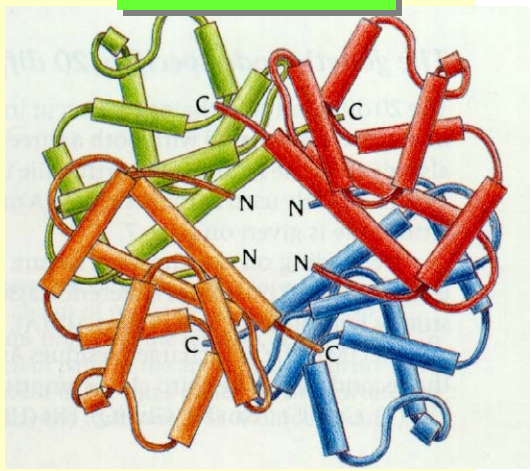

Secondary



Tertiary



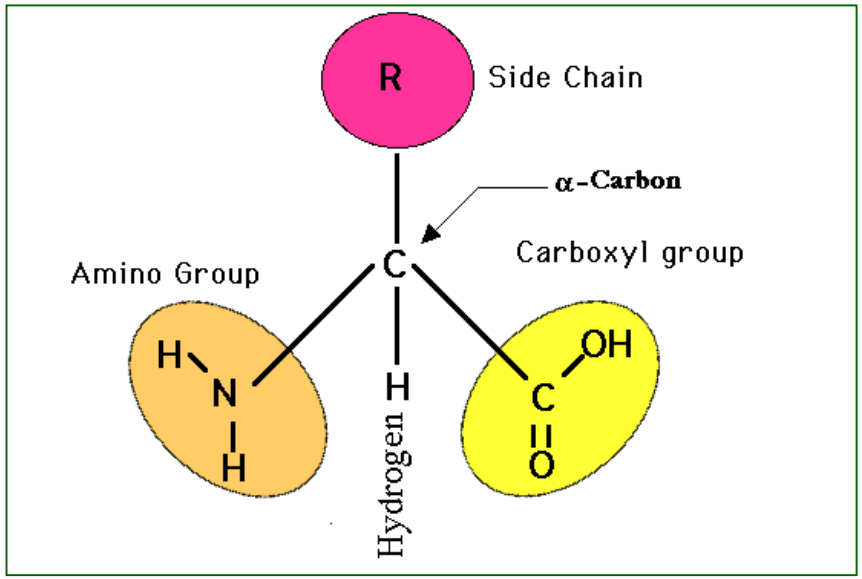

Quaternary



# Amino Acid Types

- **Hydrophobic** `I,L,M,V,A,F,P`
- **Charged**
  - **Basic** `K,H,R`
  - **Acidic** `E,D`
- **Polar** `S,T,Y,H,C,N,Q,W`
- **Small** `A,S,T`
- **Very Small** `A,G`
- **Aromatic** `F,Y,W`

All 3 figures are cartoons of an amino acid residue.



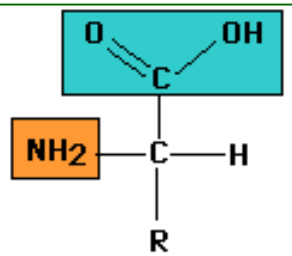
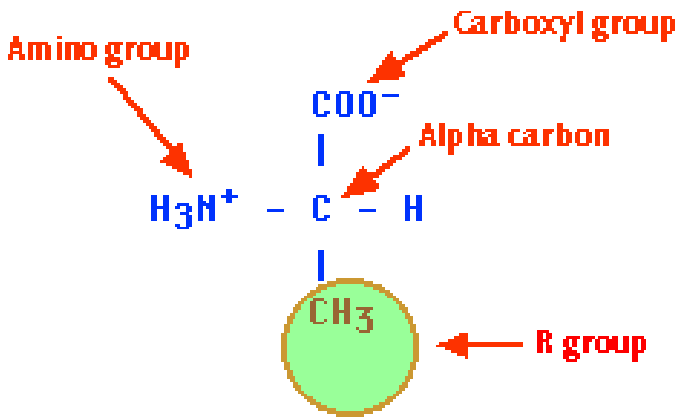
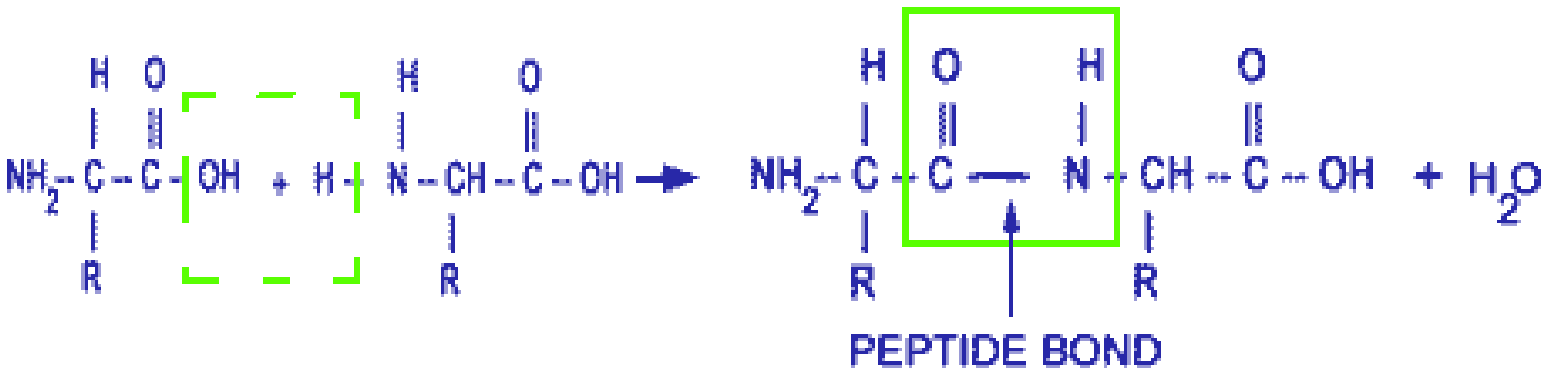
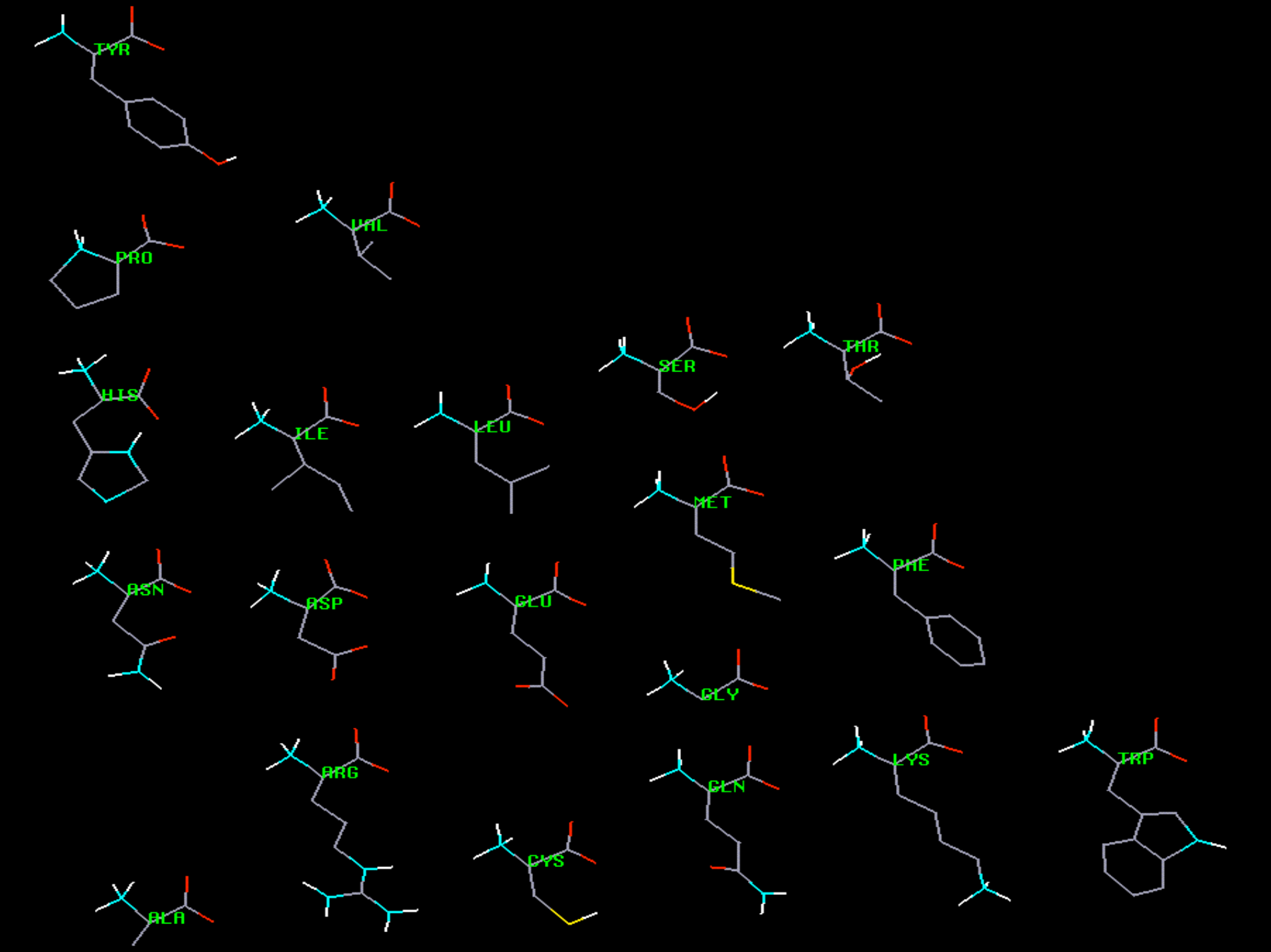




Fig. General formula for an amino acid molecule. "R" represents the variable groups that are attached to this basic molecule to make up the 20 common amino acids



# Peptide bonds in chains of residues



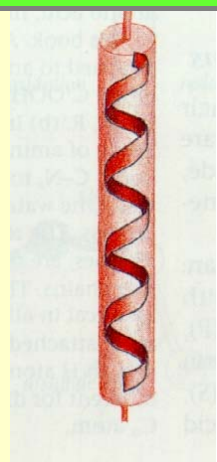

TYR
PRO
VAL
HIS
SER
THR
ILE
LEU
MET
ASN
ASP
GLU
PHE
GLY
ARG
GLN
LYS
TRP
CYS
ALA

# Proteins

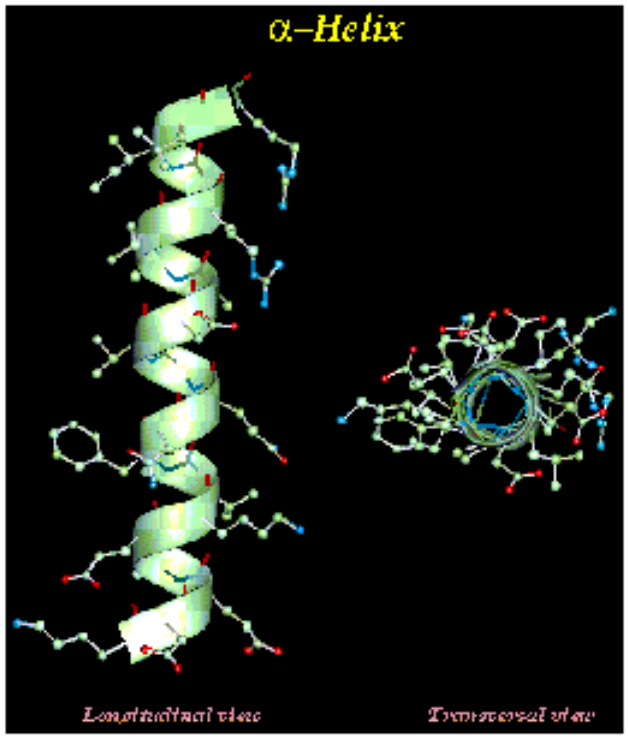
- **Primary structure** is the sequence of amino acid residues of the protein, e.g., Flavodoxin:
  **AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADA**…
- Different regions of the sequence form local regular **secondary structures**, such as
  - Alpha helix, beta strands, etc.

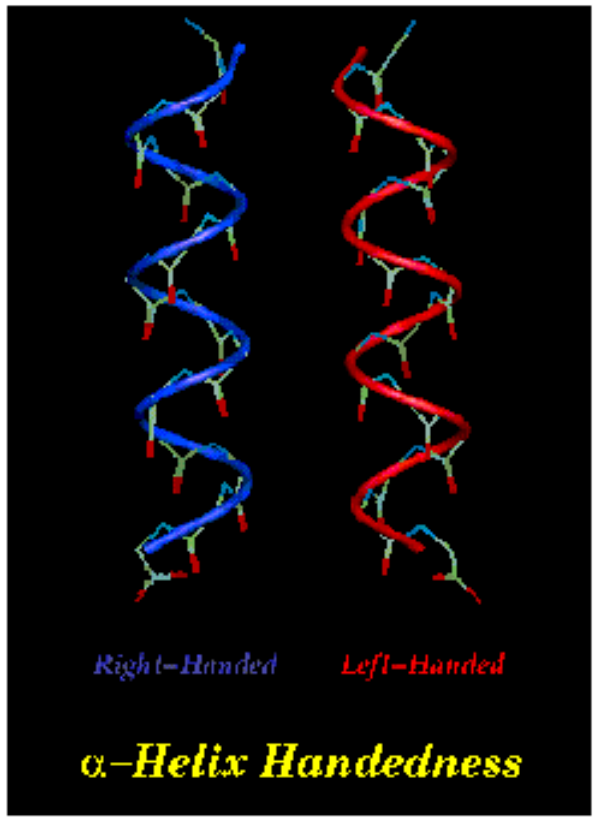
  **AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADA**…

Secondary



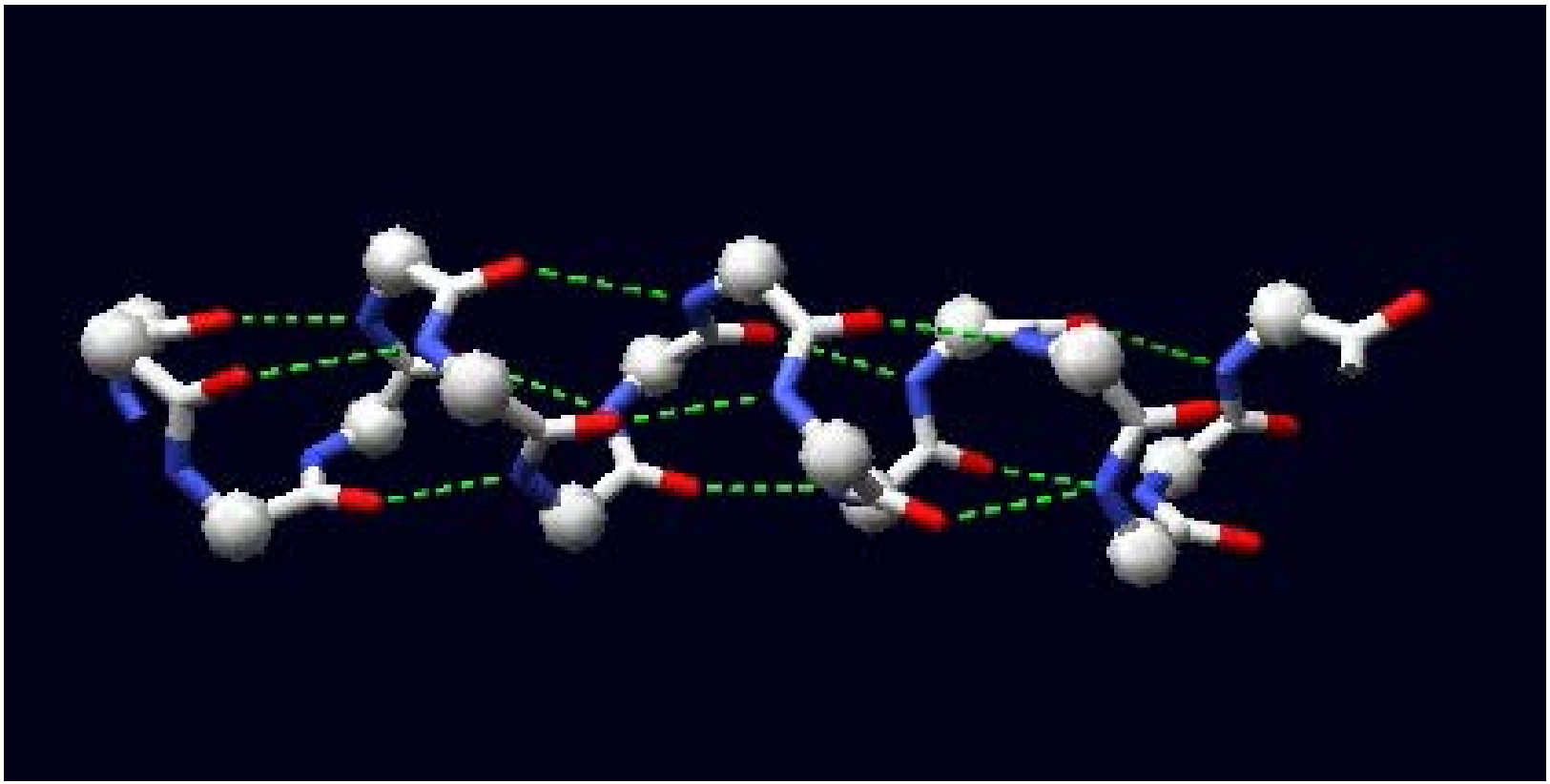
# Alpha helices



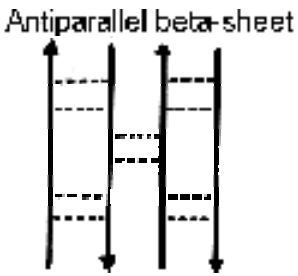



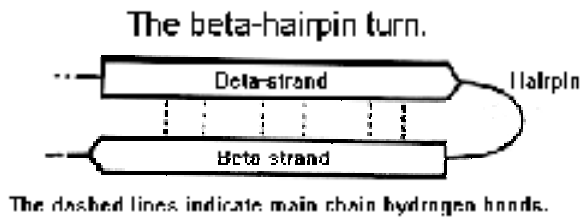

 16

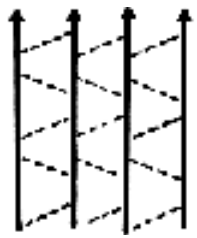
# Alpha Helix

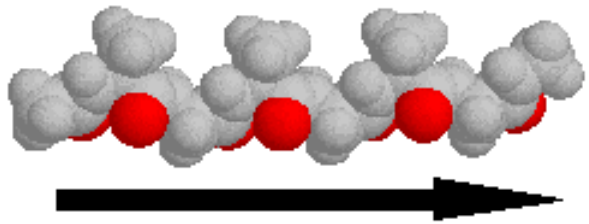


# Beta sheet



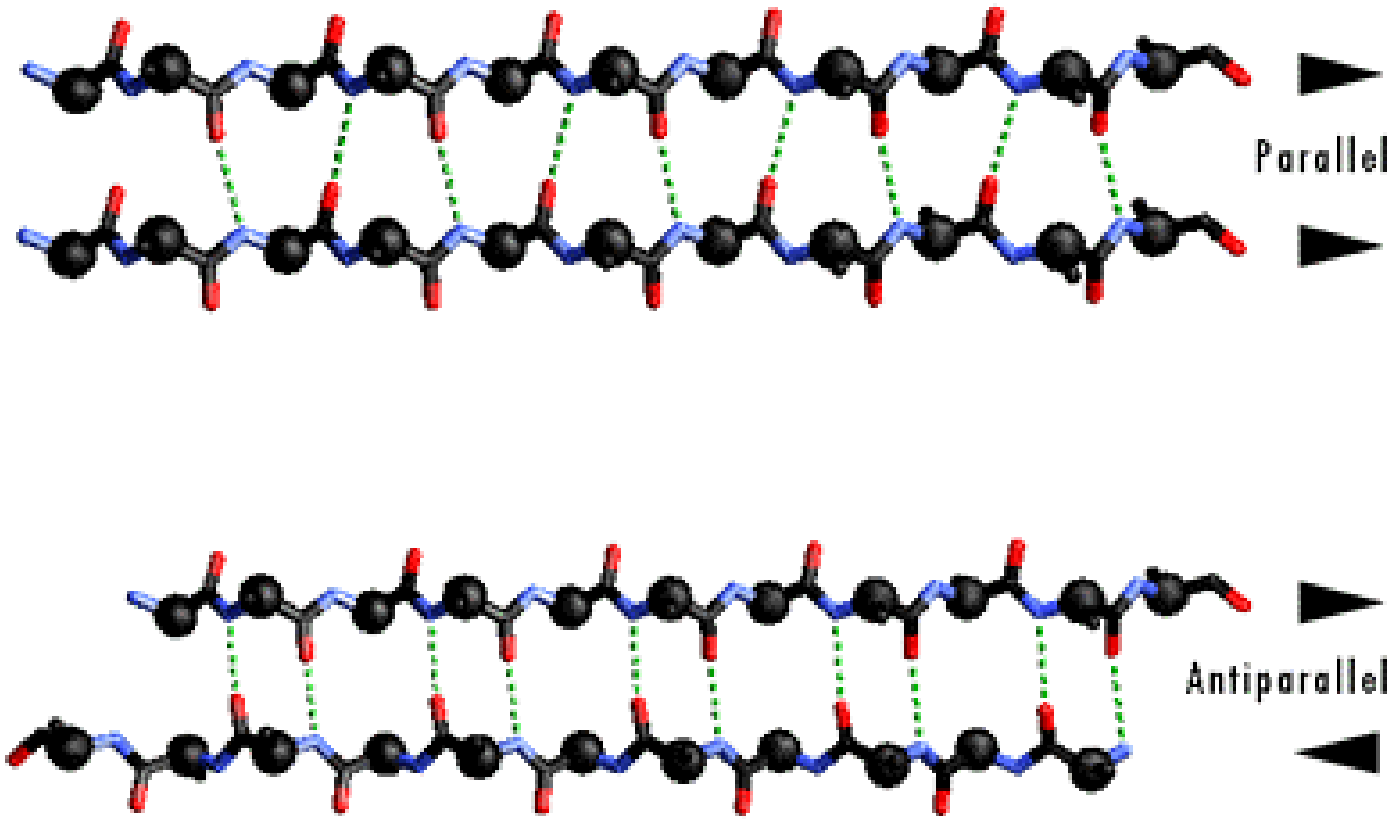



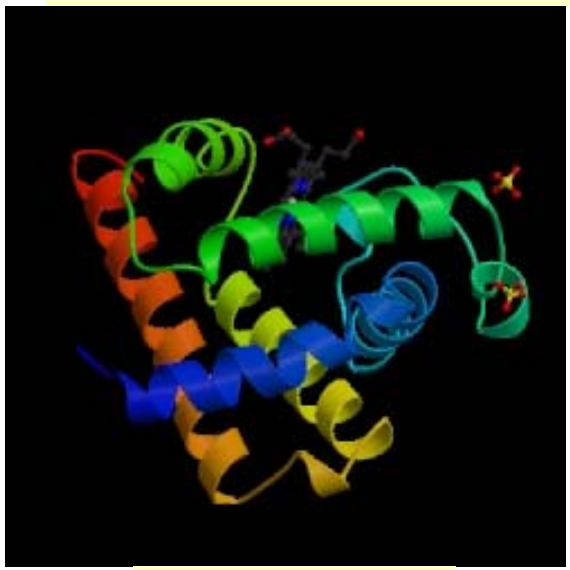
 17

# Beta Strand

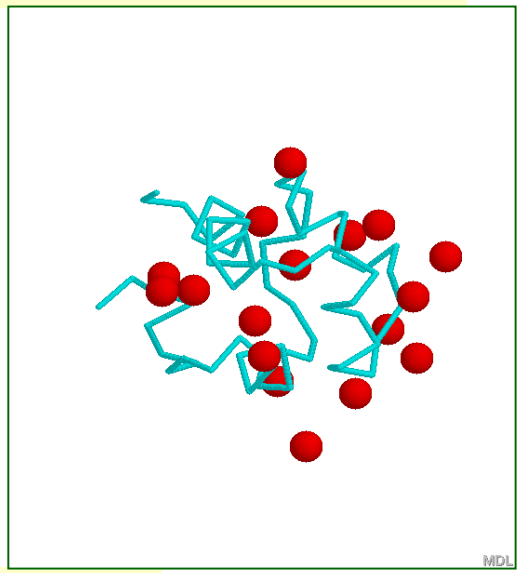


# Proteins

- **Tertiary structures** are formed by packing secondary structural elements into a globular structure.



Myoglobin

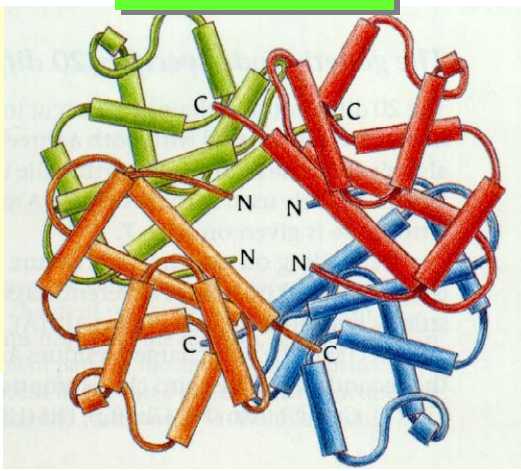


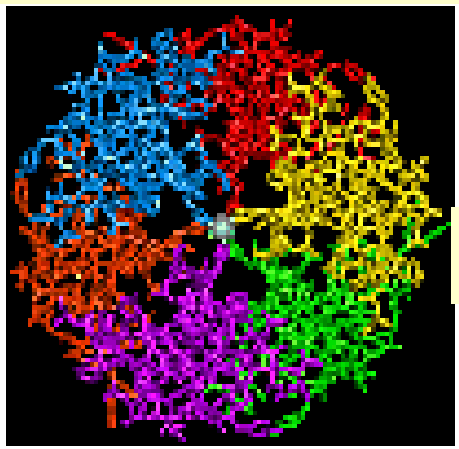

Lambda Cro

# Quaternary Structures in Proteins

- The final structure may contain more than one "chain" arranged in a **quaternary structure**.

Quaternary



Insulin Hexamer