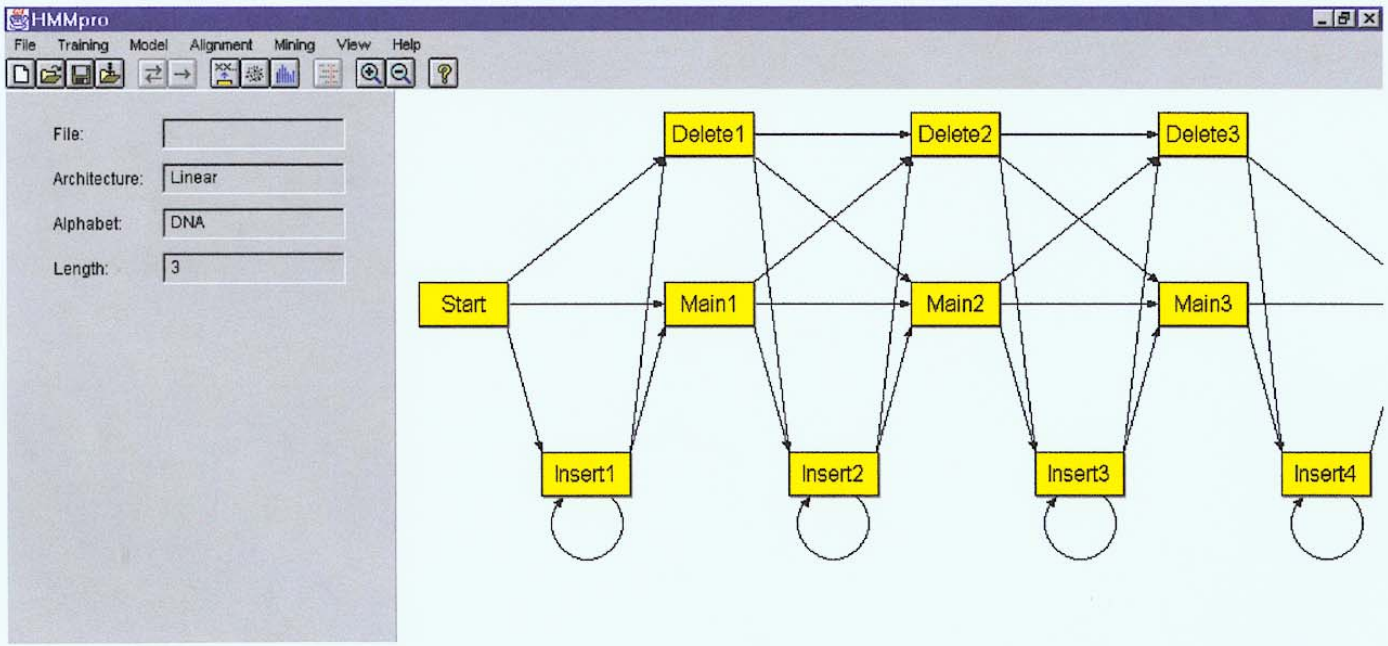
Standard HMM architectures

Linear Architecture



Entropy

- Entropy measures the variability observed in given data.

$$E = -\sum_{c} p_c \log p_c$$

- Entropy is useful in analyzing multiple alignments & profiles.
- Maximum uncertainty $\Rightarrow$ highest entropy.

G-Protein Couple Receptors

- Transmembrane proteins with 7 α-helices and 6 loops; many subfamilies
- Highly variable: 200-1200 aa in length, some have only 20% identity.
- [Baldi & Chauvin, ’94] HMM for GPCRs
- HMM constructed with 430 match states (avg length of sequences) ; Training: with 142 sequences, 12 iterations

GPCR - Analysis

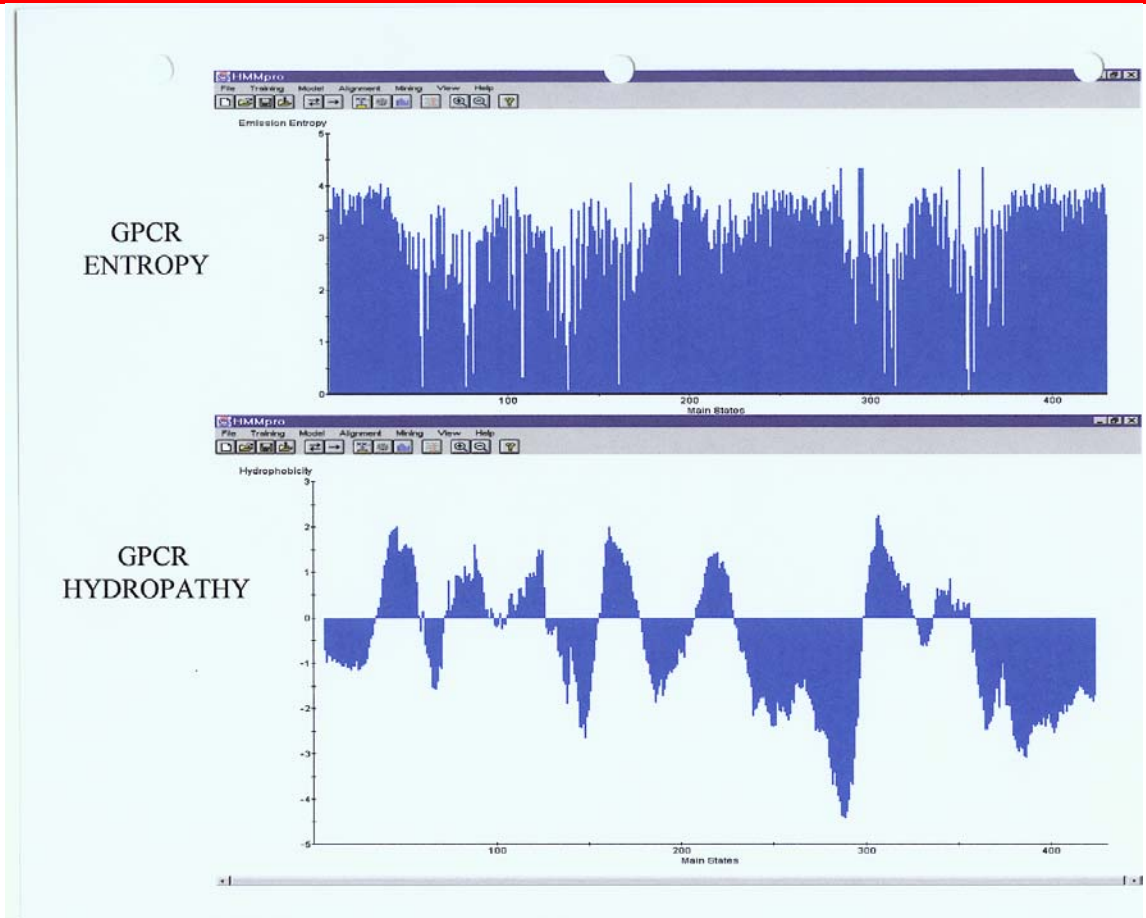
- Compute main state entropy values

$$H_i = -\sum_a e_{ia} \log e_{ia}$$

- For every sequence from test set (142) & random set (1600) & all SWISS-PROT proteins
  - Compute the negative log of probability of the most probable path π

$$Score(S) = -\log\left(P(\pi \mid S, M)\right)$$

GPCR Analysis



Entropy

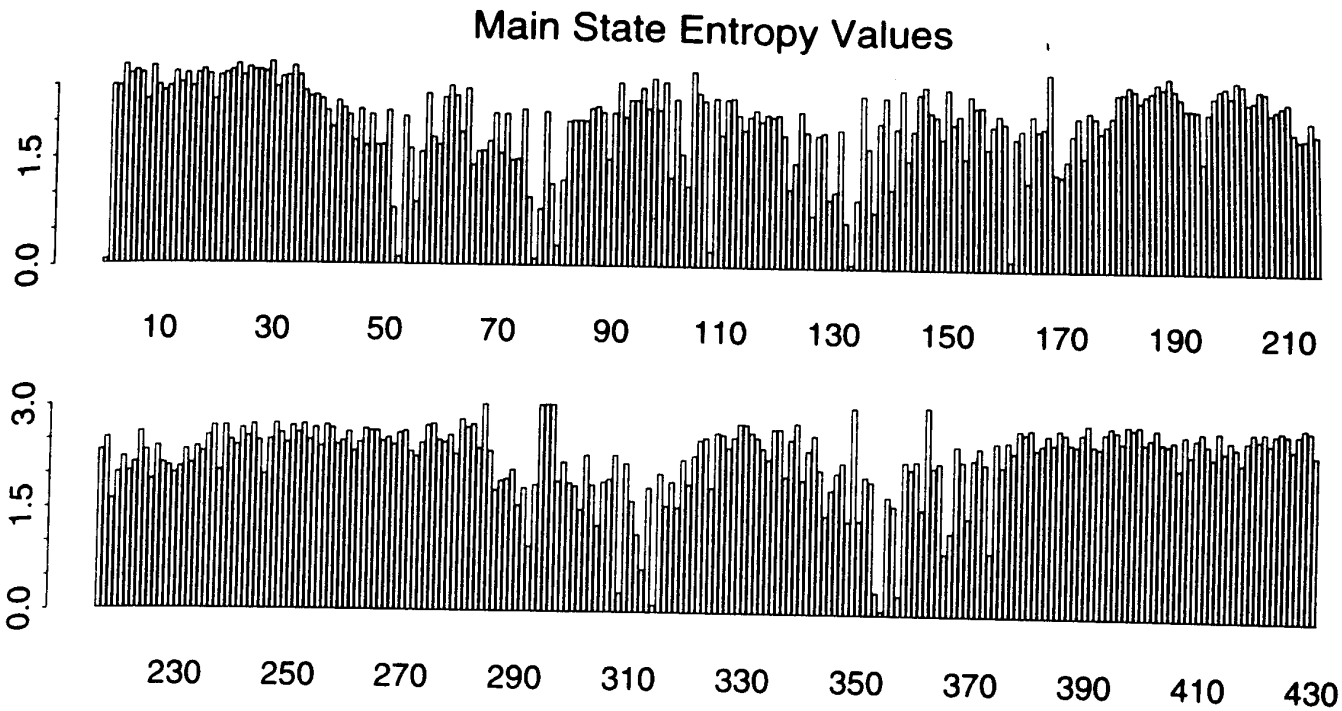


Figure 8.1: Entropy Profile of the Emission Probability Distributions Associated with the Main States of the HMM After 12 Cycles of Training.

GPCR Analysis (Cont'd)

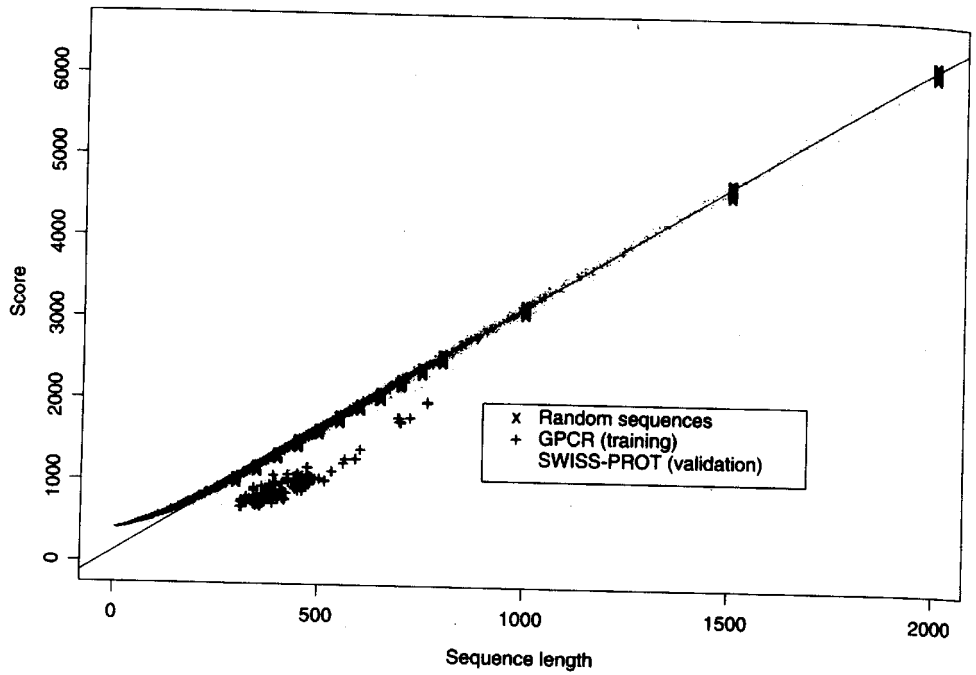


Figure 8.2: Scores (Negative Log-likelihoods of Optimal Viterbi Paths). Represented sequences consist of 142 GPCR training sequences, all sequences from the SWISS-PROT database of length less than or equal to 2000, and 220 randomly generated sequences with same average composition as the GPCRs of length 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800 (20 at each length). The regression line was obtained from the 220 random sequences. The horizontal distances in the histogram correspond to normalized scores (6).

Applications of HMM for GPCR

- Bacteriorhodopsin
  - Transmembrane protein with 7 domains
  - But it is not a GPCR
  - Compute score and discover that it is close to the regression line. Hence not a GPCR.
- Thyrotropin receptor precursors
  - Subfamily of GPCRs
  - All have long initial loop on INSERT STATE 20.
  - Also clustering possible based on distance to regression line.

HMMs – Advantages

- Sound statistical foundations
- Efficient learning algorithms
- Consistent treatment for insert/delete penalties for alignments in the form of locally learnable probabilities
- Capable of handling inputs of variable length
- Can be built in a modular & hierarchical fashion; can be combined into libraries.
- Wide variety of applications: Multiple Alignment, Data mining & classification, Structural Analysis, Pattern discovery, Gene prediction.

HMMs – Disadvantages

- Large # of parameters.
- Cannot express dependencies & correlations between hidden states.

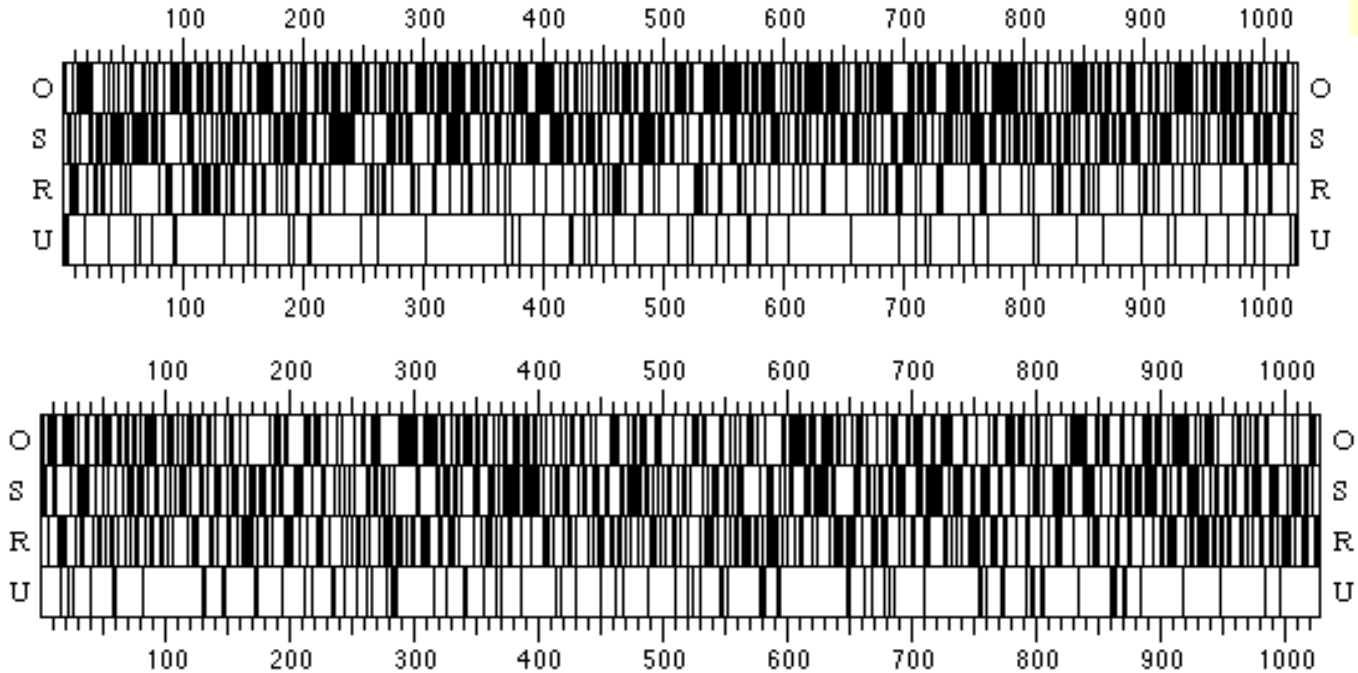
Prokaryotic Gene Prediction

- 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



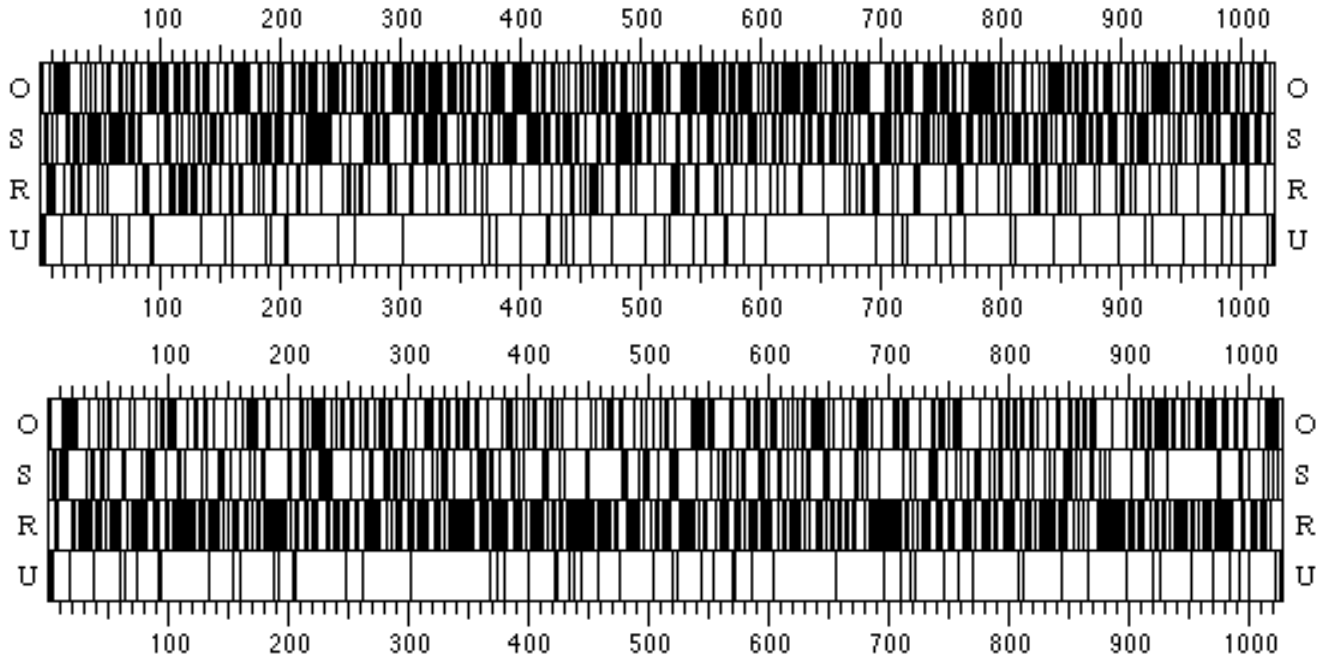
Frame Shift 1

Frame Shift 2

Codon Bias

- Codon biases specific to organisms

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



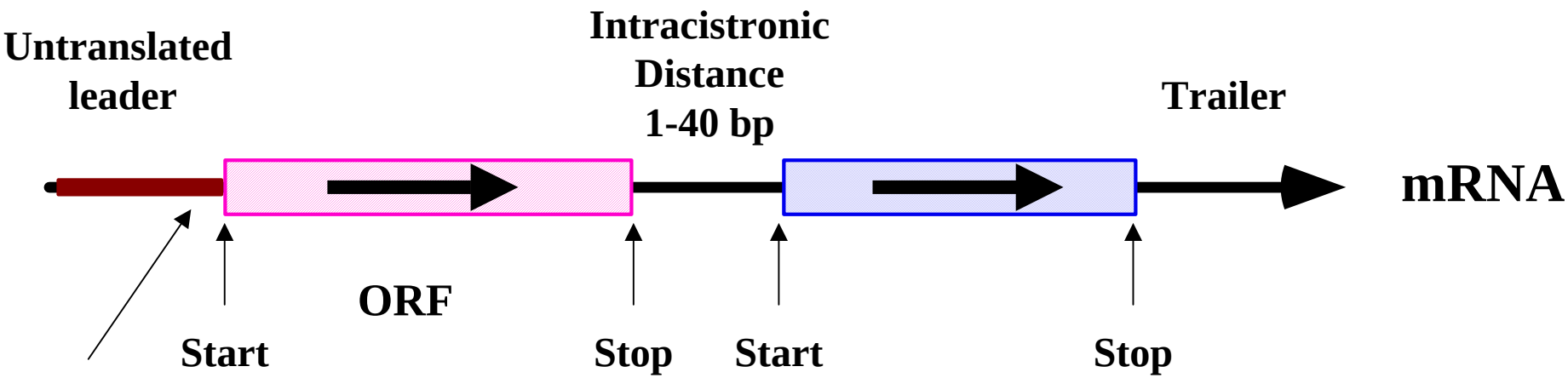
Same Frames; Different labeling of codon types (i.e., from yeast)

Messenger RNA or mRNA

Initation Codon	AUG	Methionine

Termination Codons	
UAA	Ochre
UAG	Amber
UGA	Opal

Others:		
	GUG	Valine
	UUG	Leucine
	AUU	Isoleucine



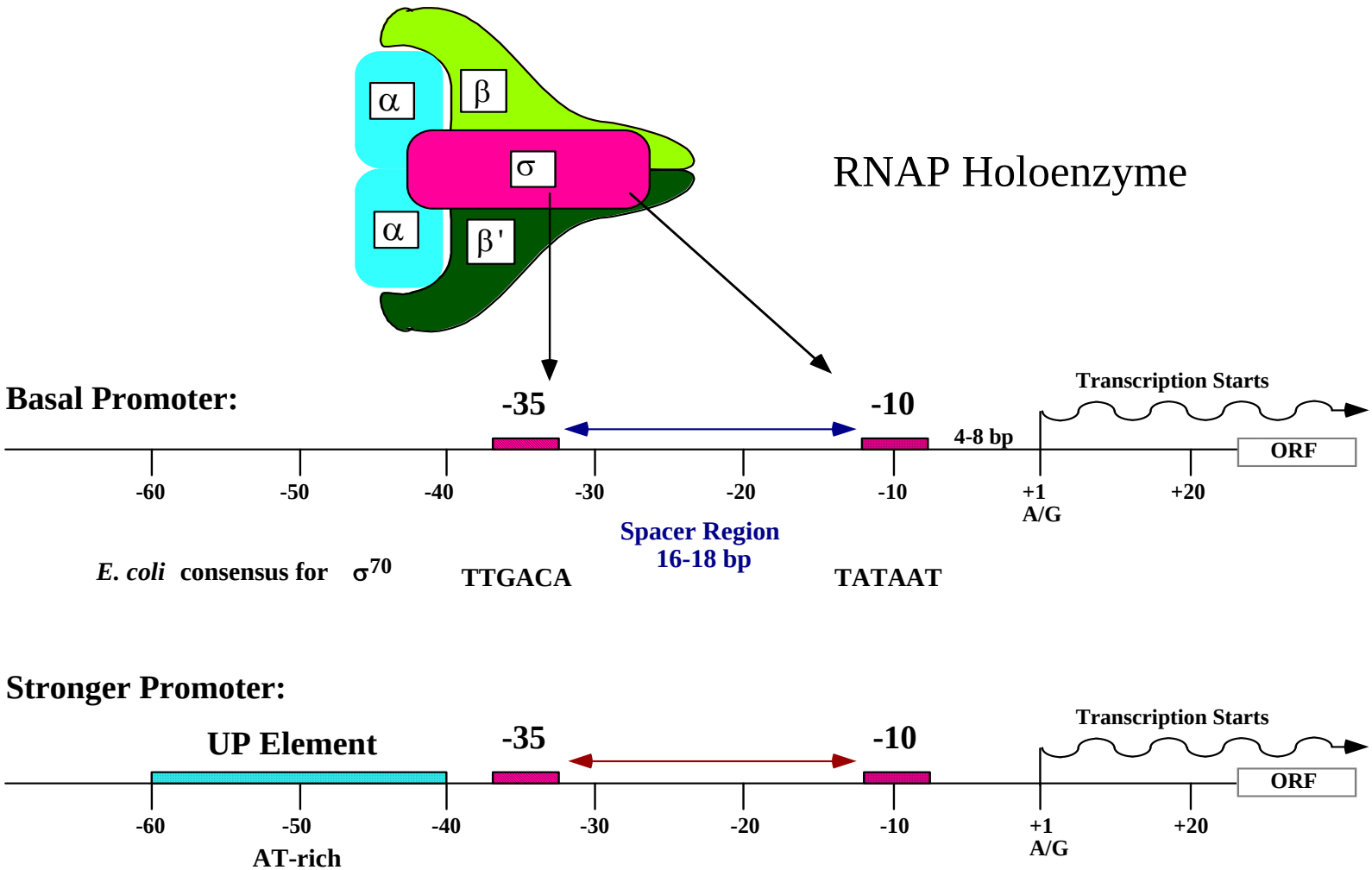
RBS
Ribosome Binding Site
Shine-Dalgarno Sequence

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*

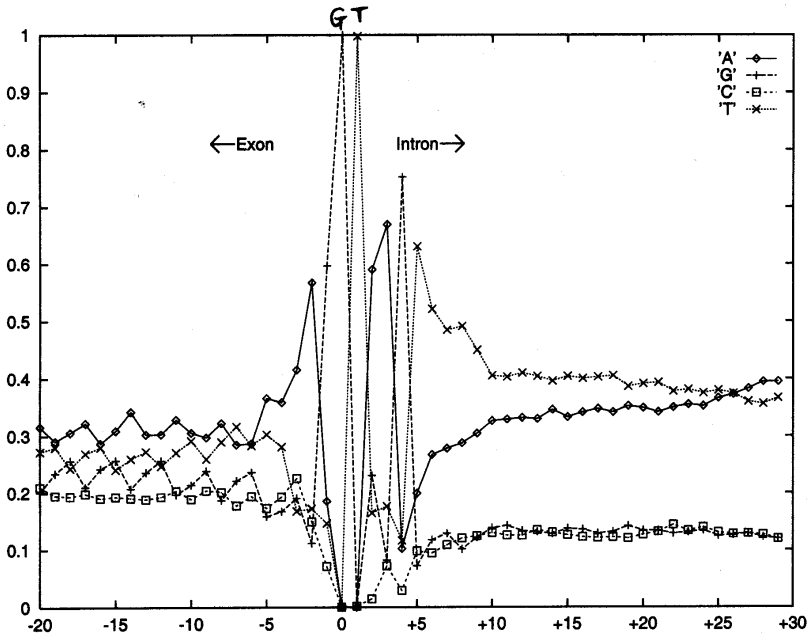


Figure 2: Profile of the same 5' collection but around a larger window.

A/T

G/C

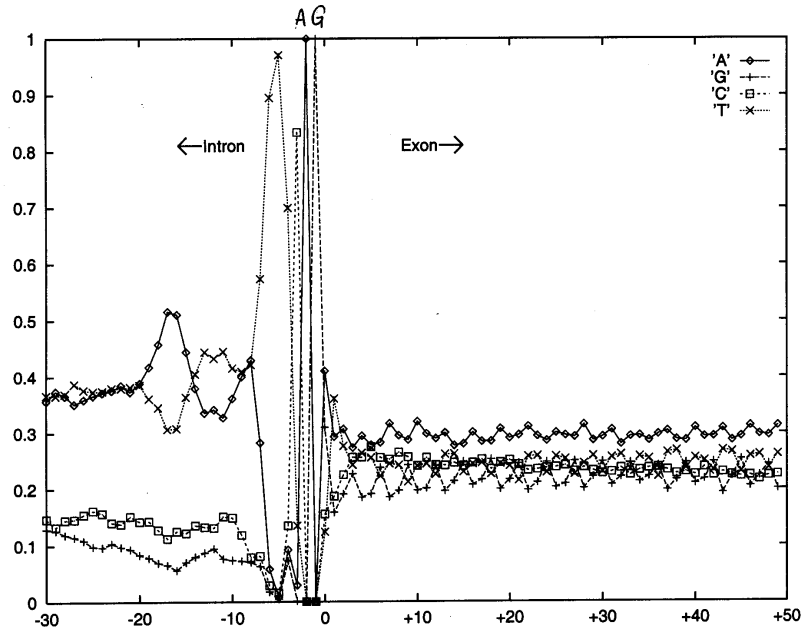
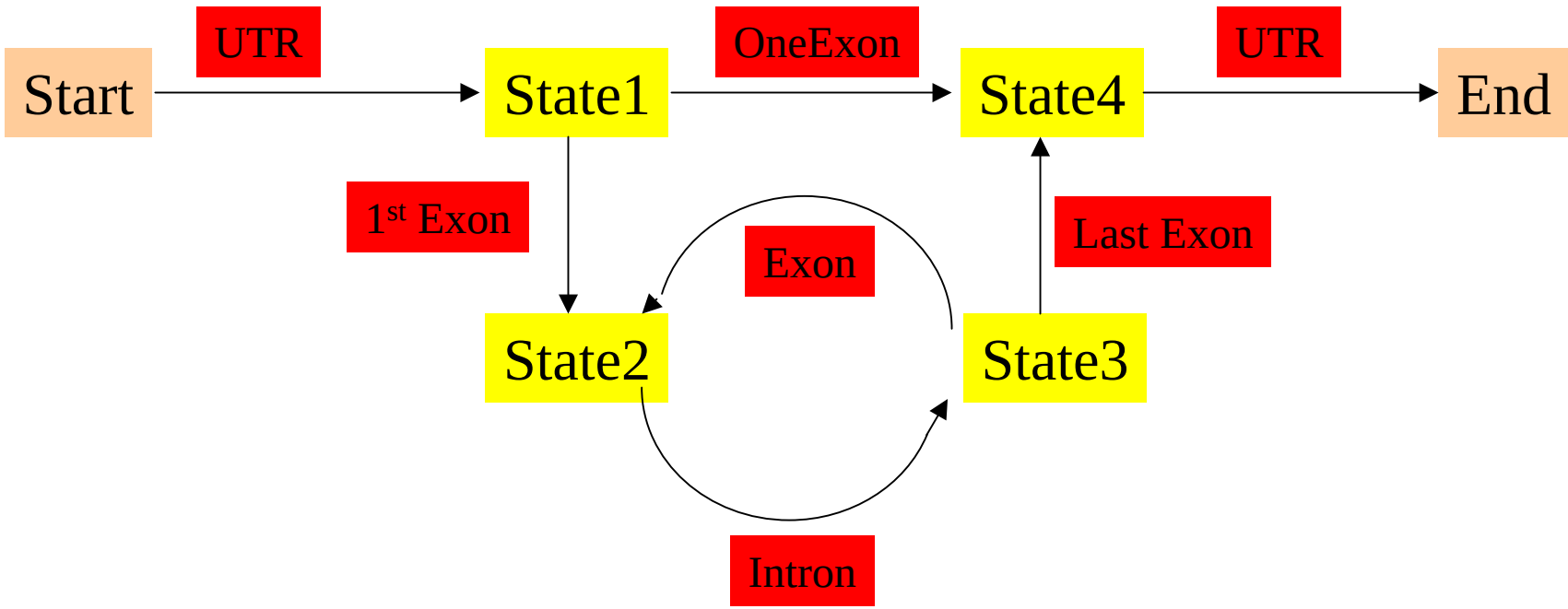


Figure 4: Profile of 8,192 sequences of length 80 around the 3' site. The first position in the exon is labeled 0.

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

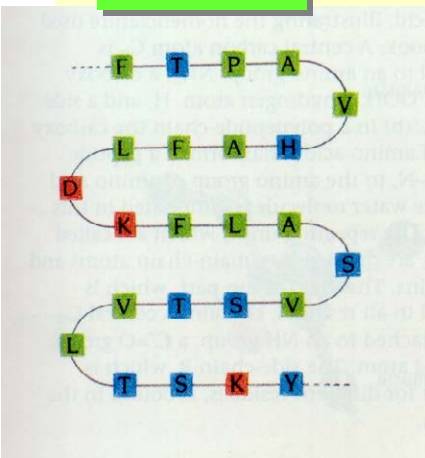
HMM structure for Gene Finding



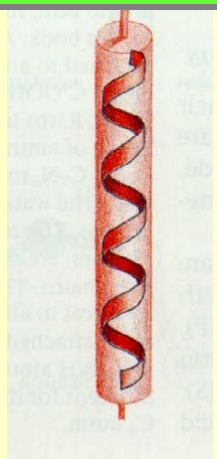
Protein Structures

- Sequences of amino acid residues
- 20 different amino acids

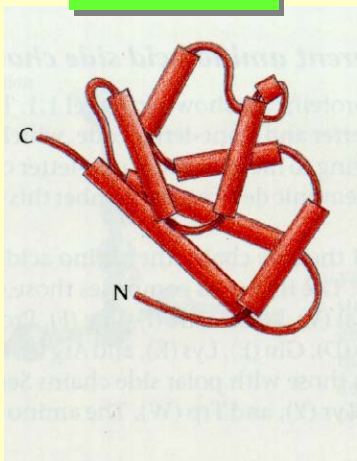
Primary



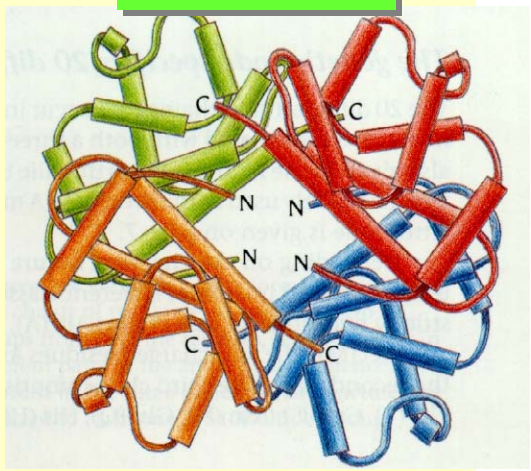
Secondary



Tertiary



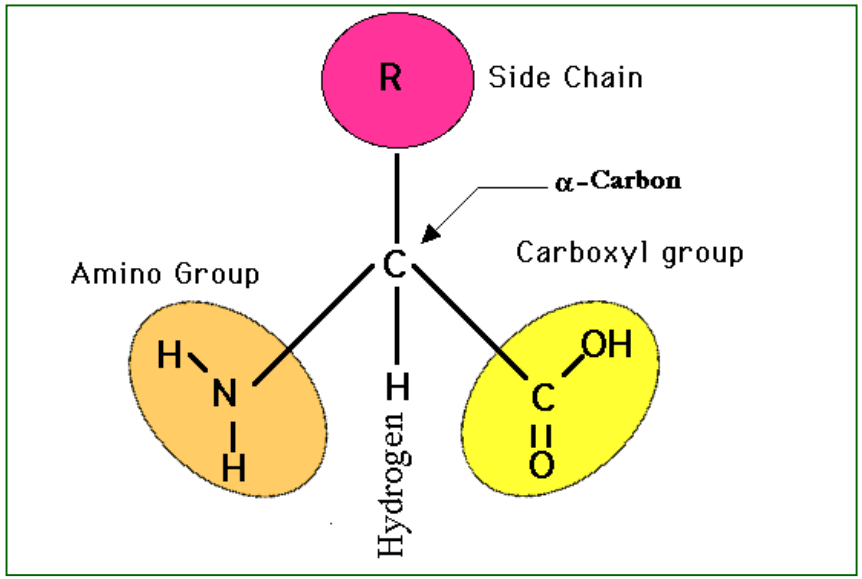
Quaternary



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.



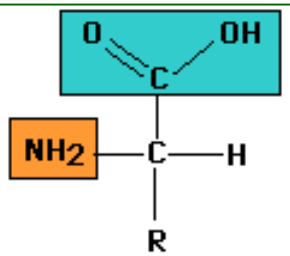
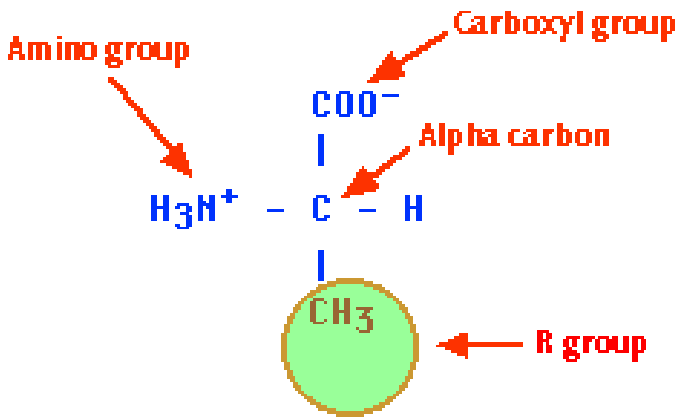


Fig. General formula for an amino acid molecule. "R" represents the variable groups that are attached to this basic molecule to make up the 20 common amino acids



Angles ϕ and ψ in the polypeptide chain

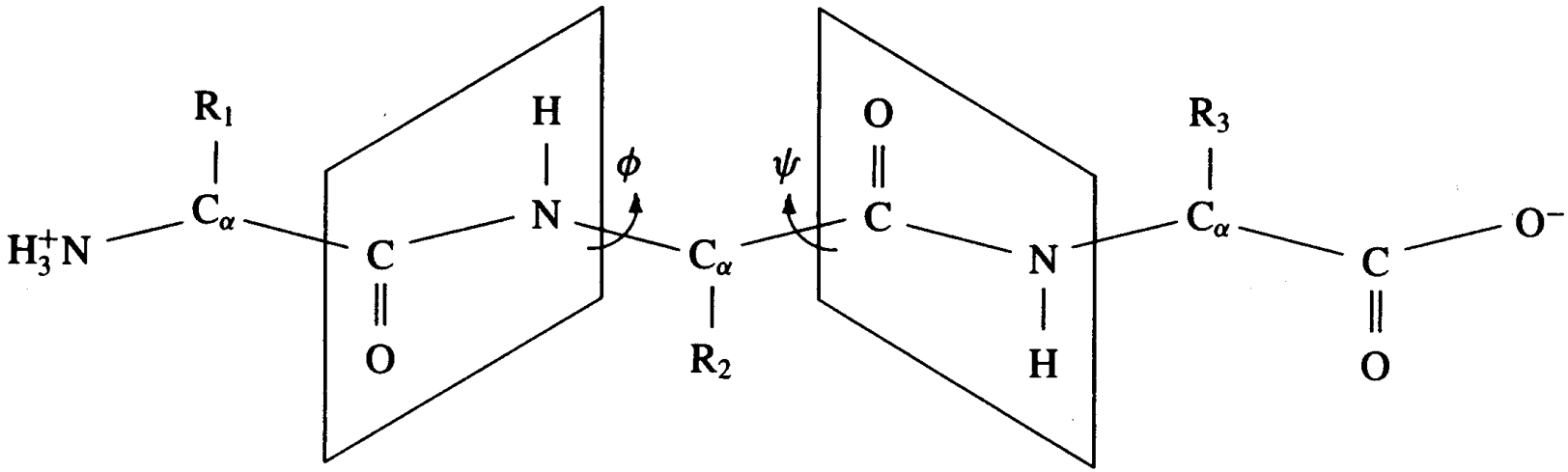
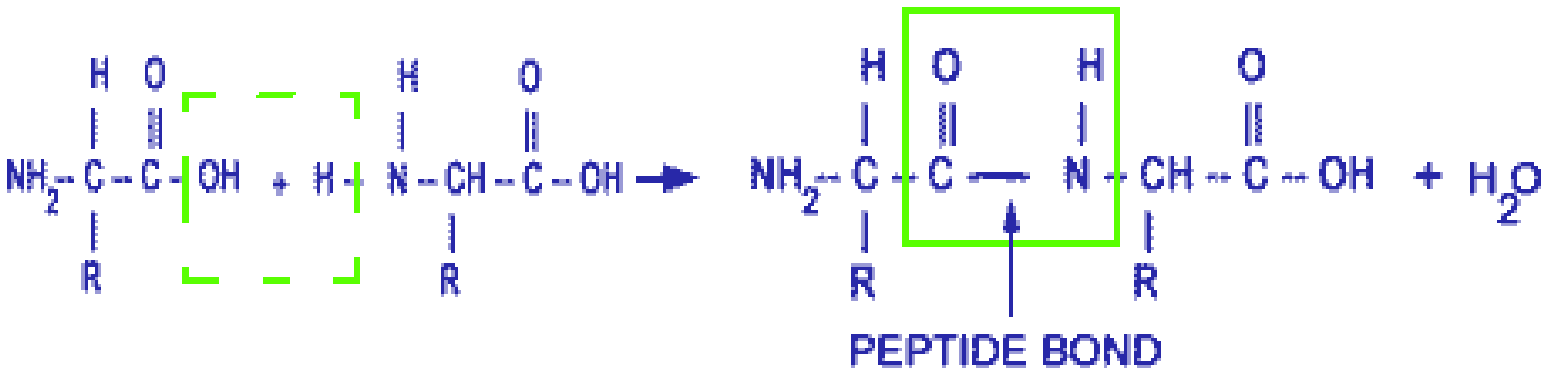
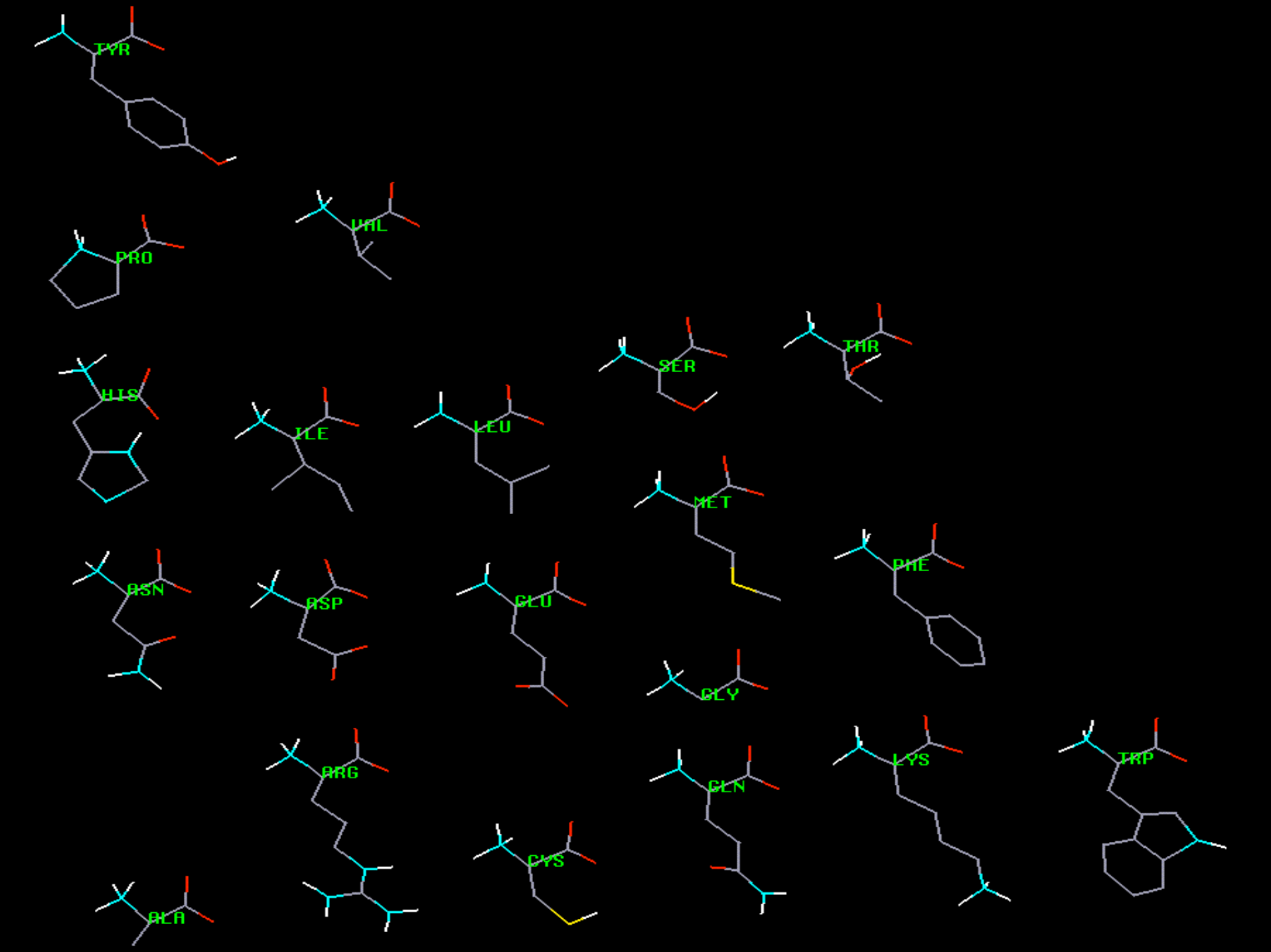


FIGURE 1.2

A polypeptide chain. The R_i *side chains identify the component amino acids. Atoms inside each quadrilateral are on the same plane, which can rotate according to angles* ϕ *and* ψ.

Peptide bonds in chains of residues



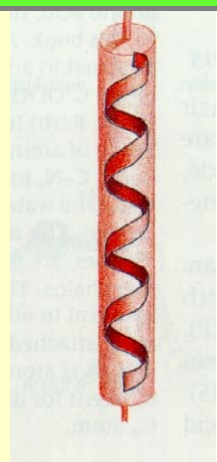

TYR
PRO
VAL
HIS
SER
THR
ILE
LEU
MET
ASN
ASP
GLU
PHE
GLY
ARG
GLN
LYS
TRP
CYS
ALA

Proteins

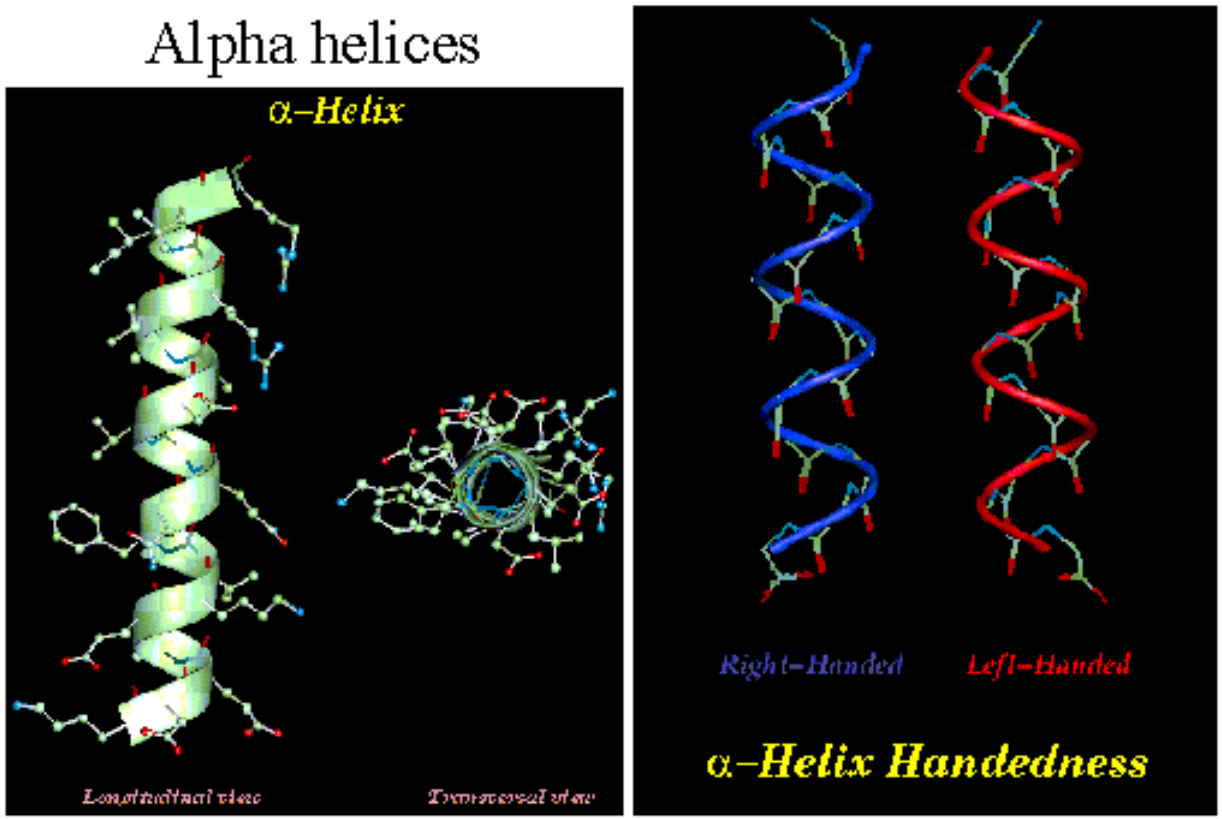
- **Primary structure** is the sequence of amino acid residues of the protein, e.g., Flavodoxin:
  AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADA…
- Different regions of the sequence form local regular **secondary structures**, such as
  - Alpha helix, beta strands, etc.

  AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADA…

Secondary



Alpha helices



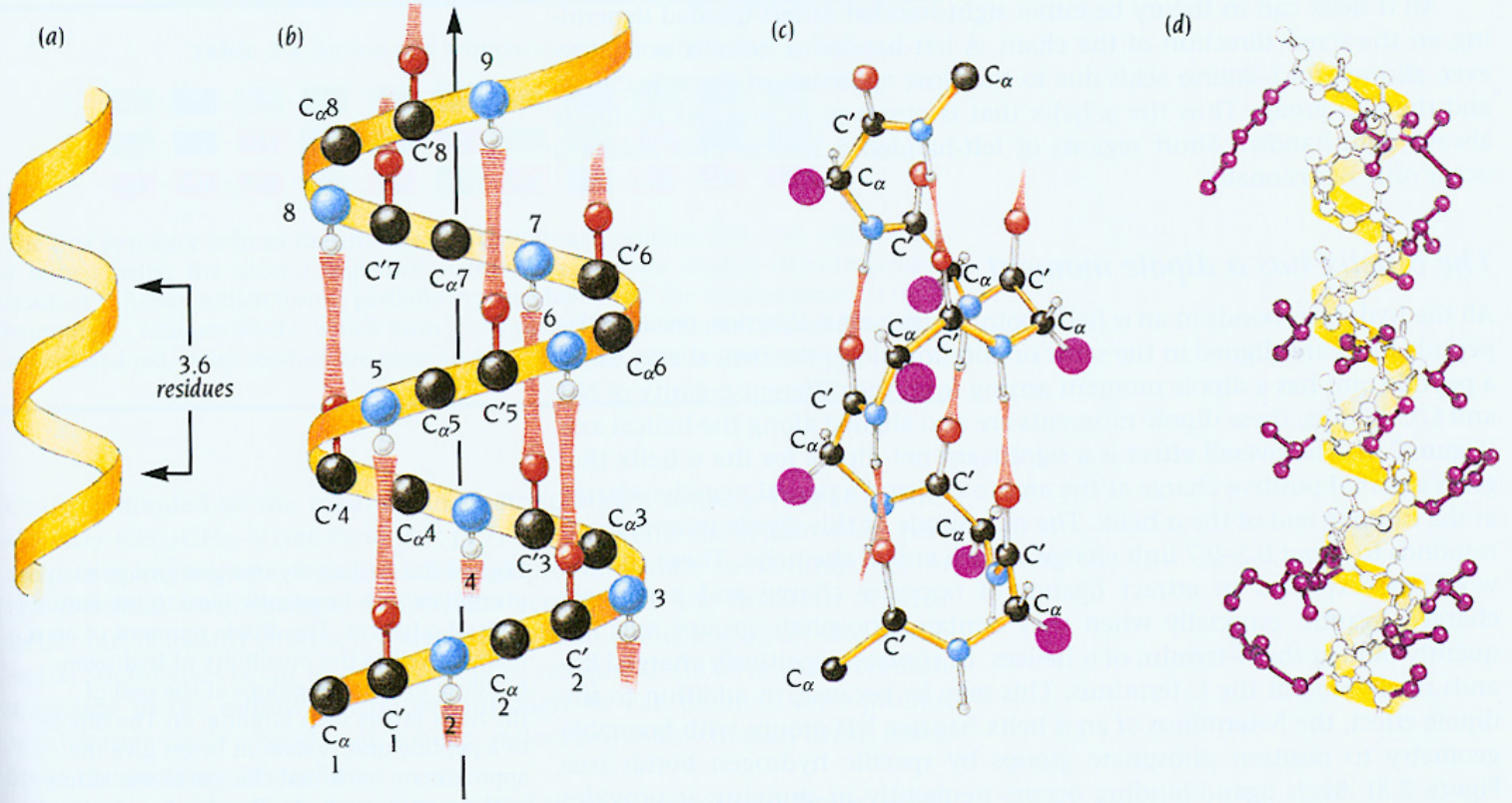
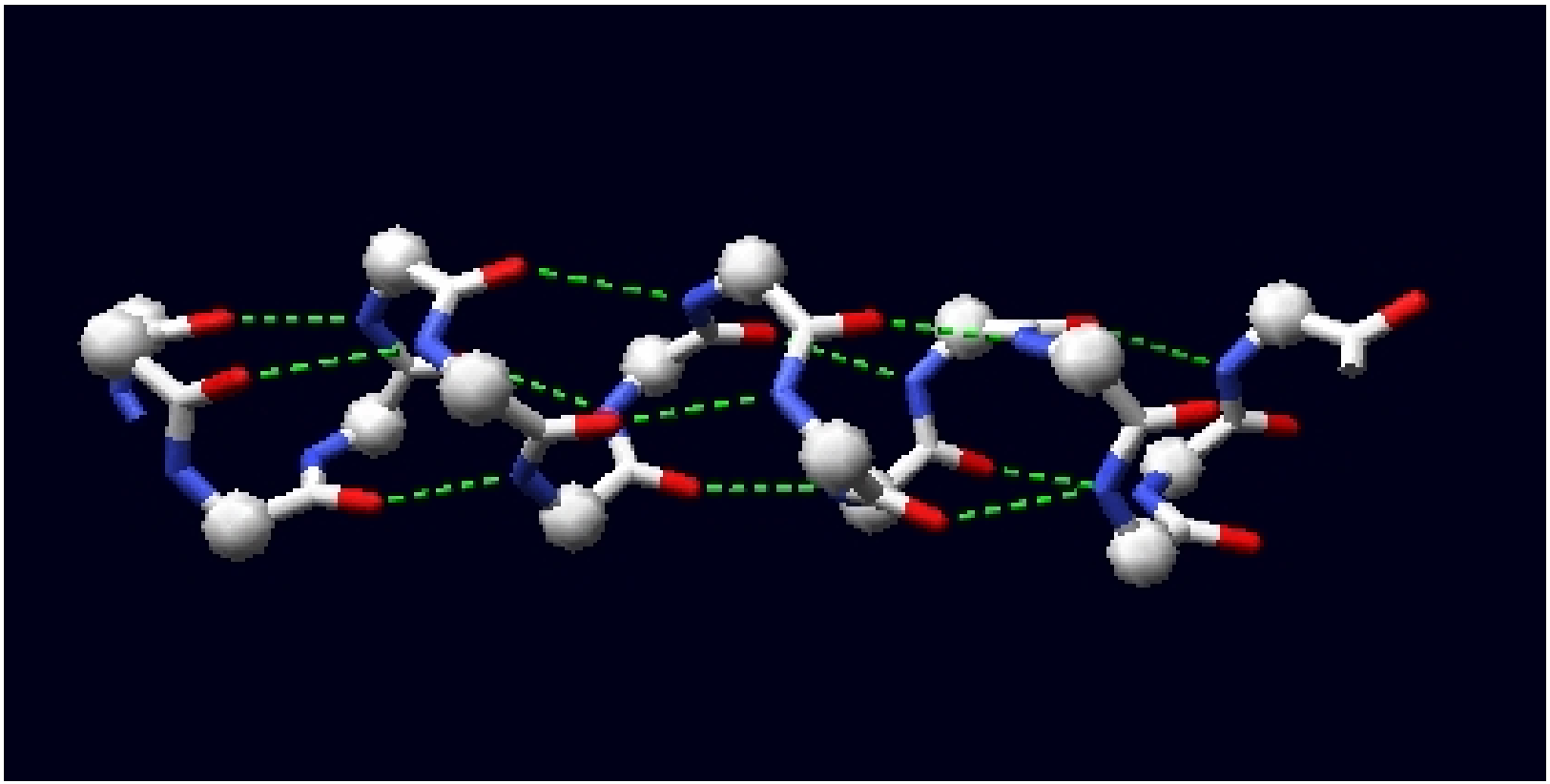
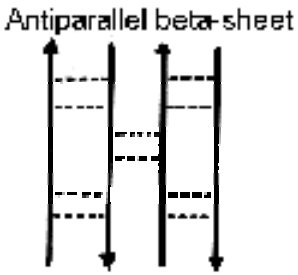


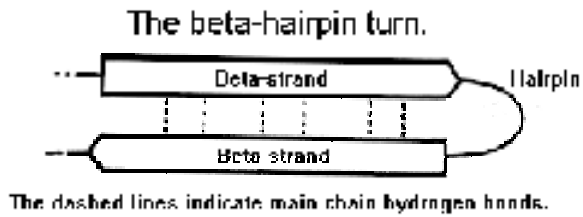
Figure 2.2 The α helix is one of the major elements of secondary structure in proteins. Main-chain N and O atoms are hydrogen-bonded to each other within α helices. (a) Idealized diagram of the path of the main chain in an α helix. Alpha helices are frequently illustrated in this way. There are 3.6 residues per turn in an α helix, which corresponds to 5.4 Å (1.5 Å per residue). (b) The same as (a) but with approximate positions for main-chain atoms and hydrogen bonds included. The arrow denotes the direction from the N-terminus to the C-terminus. (c) Schematic diagram of an α helix. Oxygen atoms are red, and N atoms are blue. Hydrogen bonds between O and N are red and striated. The side chains are represented as purple circles. (d) A ball-and-stick model of one α helix in myoglobin. The path of the main chain is outlined in yellow; side chains are purple. Main-chain atoms are not colored. (e) One turn of an α helix viewed down the helical axis. The purple side chains project out from the α helix.

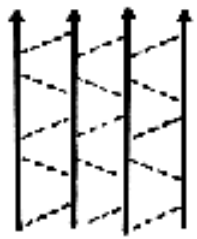
Alpha Helix

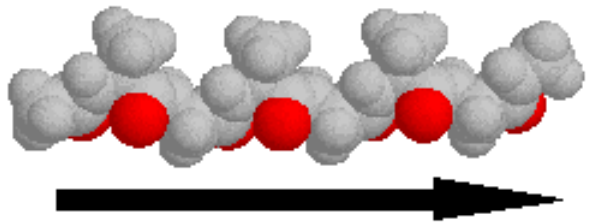


Beta sheet



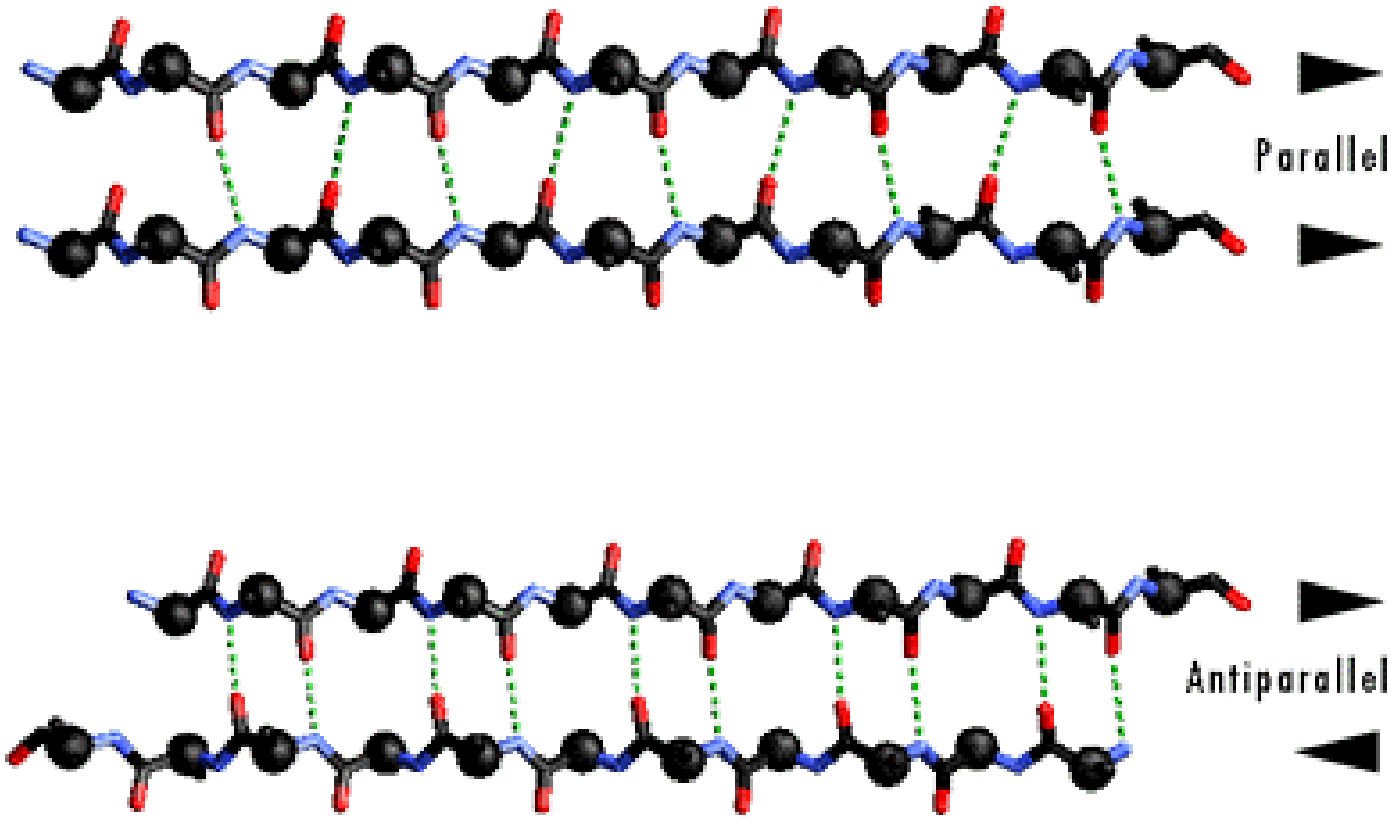






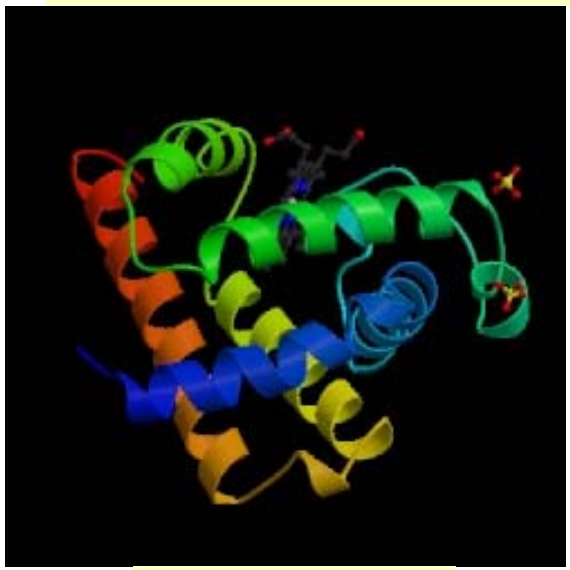
 17

Beta Strand

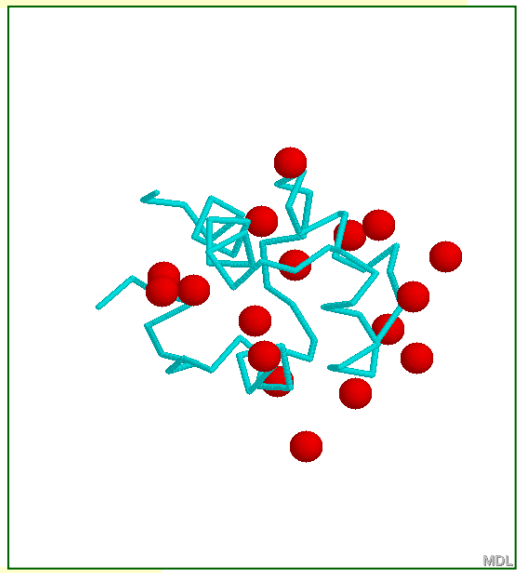


Proteins

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



Myoglobin

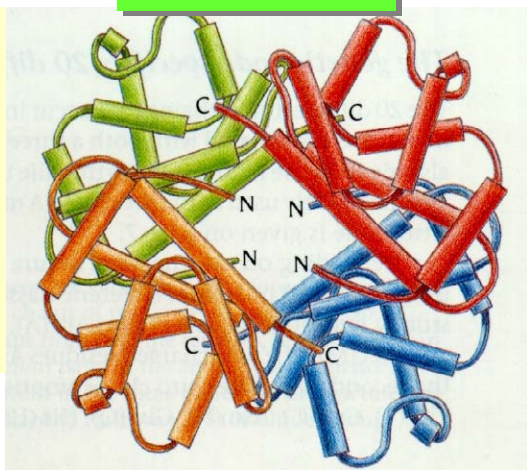


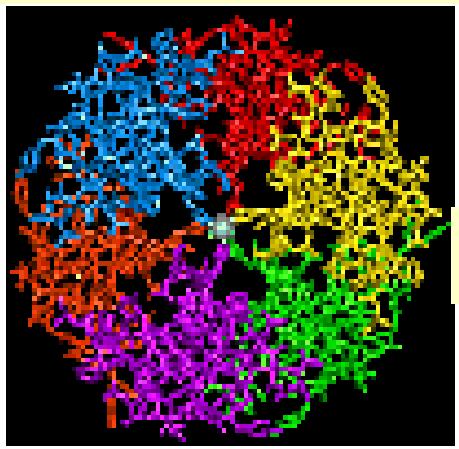
Lambda Cro

Quaternary Structures in Proteins

Quaternary

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





Insulin Hexamer

More on Secondary Structures

- **α-helix**
  - Main chain with peptide bonds
  - Side chains project outward from helix
  - Stability provided by H-bonds between CO and NH groups of residues 4 locations away.
- **β-strand**
  - Stability provided by H-bonds with one or more β-strands, forming β-sheets. Needs a β-turn.

Secondary Structure Prediction Software

Figure 11.3 Comparison of secondary structure predictions by various methods. The sequence of flavodoxin, an α/β protein, was used as the query and is shown on the first line of the alignment. For each prediction, H denotes an α helix, E a β strand, T a β turn; all other positions are assumed to be random coil. Correctly assigned residues are shown in inverse type. The methods used are listed along the left side of the alignment and are described in the text. At the bottom of the figure is the secondary structure assignment given in the PDB file for flavodoxin (1OFV, Smith et al., 1983).

Active Sites

Active sites in proteins are usually hydrophobic pockets/crevices/troughs that involve sidechain atoms.

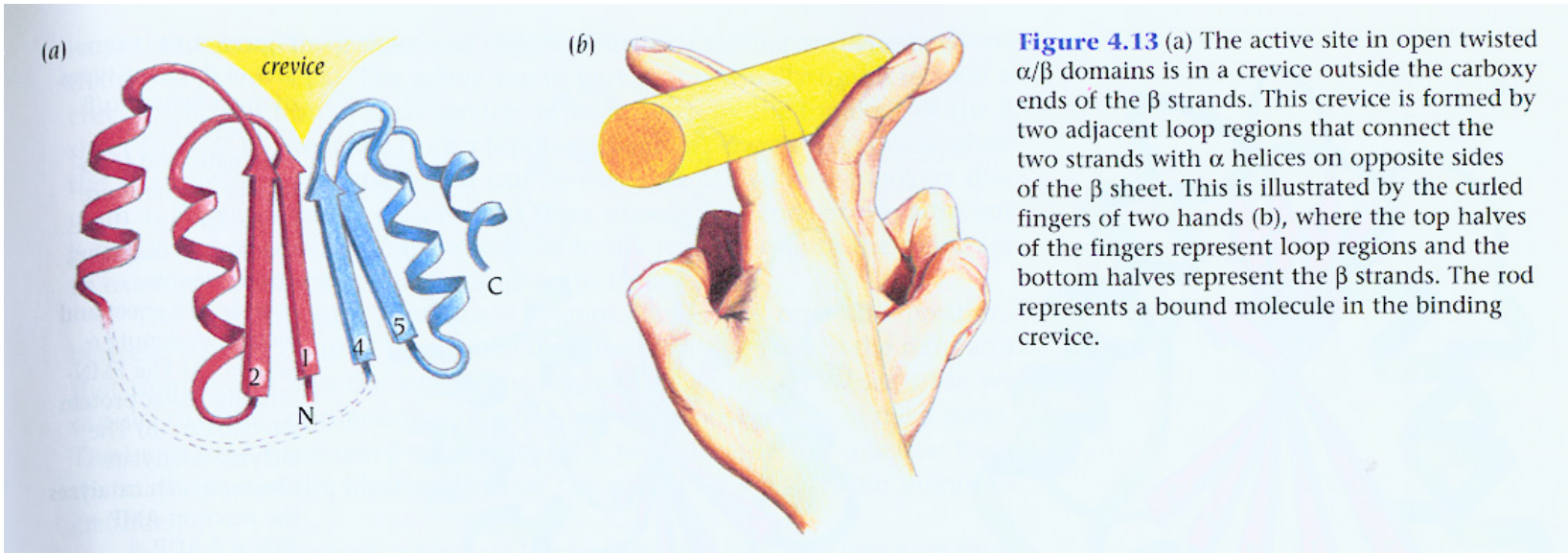
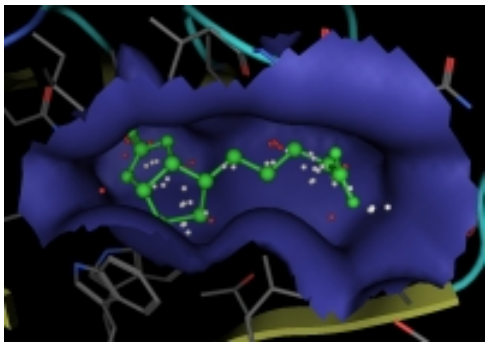
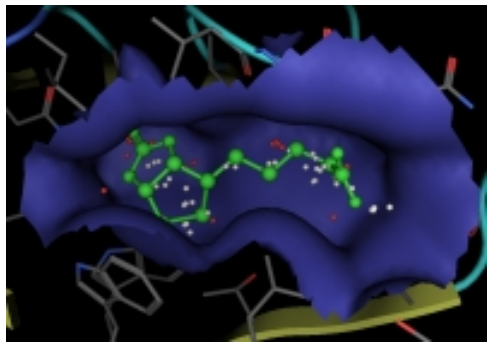
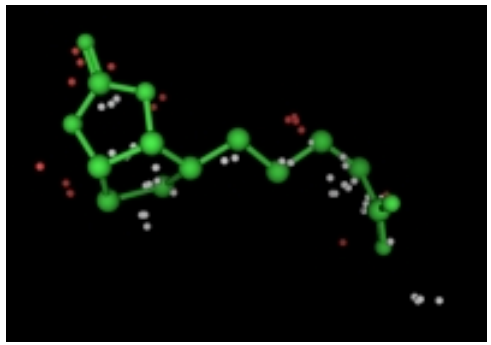


Figure 4.13 (a) The active site in open twisted α/β domains is in a crevice outside the carboxy ends of the β strands. This crevice is formed by two adjacent loop regions that connect the two strands with α helices on opposite sides of the β sheet. This is illustrated by the curled fingers of two hands (b), where the top halves of the fingers represent loop regions and the bottom halves represent the β strands. The rod represents a bound molecule in the binding crevice.

Active Sites

Left PDB 3RTD (streptavidin) and the first site located by the MOE Site Finder. **Middle** 3RTD with complexed ligand (biotin). **Right** Biotin ligand overlaid with calculated alpha spheres of the first site.

Shift-And Method (Baeza-Yates & Gonnet)

Idea: *Build a bit-matrix M* such that

M[I,J] = 1 ⇔ P[1..I] = T[J-I+1 .. J]

Thus, M[I,J] = 1 ⇔ (M[I-1, J-1] = 1) & (P[I] = T[J])

M	T	A	G	T	A	G	A	A	G	A	A	C
A	0	1	0	0	1	0	1	1	0	1	1	0
G		0	1	0	0	1	0	0	1	0	0	0
A			0	0	0	0	1	0	0	1	0	0
A				0	0	0	0	1	0	0	1	0
C					0	0	0	0	0	0	0	1

Shift-And (Cont'd)

Idea: *Operate on column bit-vectors*

Step 1: Build a bit-matrix U such that for each e $\in \Sigma$

U[I,e] = 1 $\Leftrightarrow$ P[I] = e

U	A	C	G	T
A	1	0	0	0
G	0	0	1	0
A	1	0	0	0
A	1	0	0	0
C	0	1	0	0

Step 2: M[J] = RightShift(M[J-1]) && U[T[J]]

Shift-And (Cont'd)

Step 2: M[J] = RightShift(M[J-1]) && U[T[J]]

M	T	A	G	T	A	G	A	A	G	A	A	C
A	0	1	0	0	1	0	1	1	0	1	1	0
G		0	1	0	0	1	0	0	1	0	0	0
A			0	0	0	0	1	0	0	1	0	0
A				0	0	0	0	1	0	0	1	0
C					0	0	0	0	0	0	0	1

U	A	C	G	T
A	1	0	0	0
G	0	0	1	0
A	1	0	0	0
A	1	0	0	0
C	0	1	0	0

Shift-And (Generalizations)

Generalization 1: *Wild Cards*: match all characters.

U	A	C	G	T
A	1	0	0	0
G	0	0	1	0
A	1	0	0	0
*	1	1	1	1
C	0	1	0	0

Generalization 2: *k Mismatches*: Compute M_0, M_1, …, M_k

$M_s[J]$ = RightShift($M_{s-1}[J-1]$ **AND** U[T[J]])
OR $M_{s-1}[J]$
OR $M_{s-1}[J-1]$

String Matching Methods: Overview

Methods:

- Naïve Method **O(mn)** *time*
- Rabin Karp Method **O(mn)** *time;* Fast on average.
- FSA-based method **O(n+mA)** *time*
- Knuth-Morris-Pratt algorithm **O(n+m)** *time*
- Boyer-Moore **O(mn)** time; Very fast on average.
- Suffix Tree method; **O(m+n)** *time*
- Shift-And method; Fast on average; Bit operations.