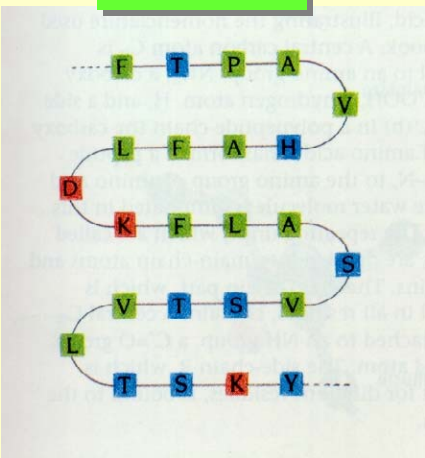
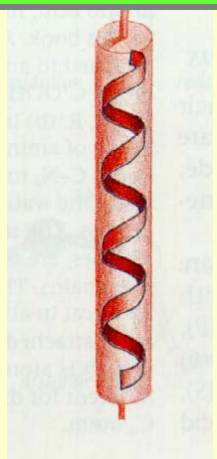
Protein Structures

- Sequences of amino acid residues
- 20 different amino acids

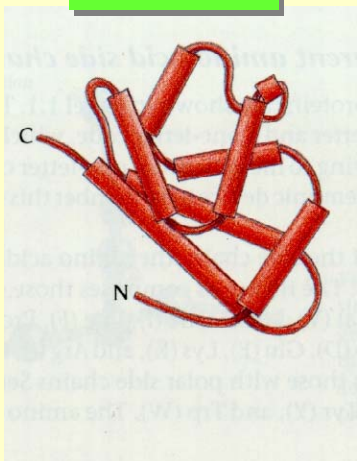
Primary



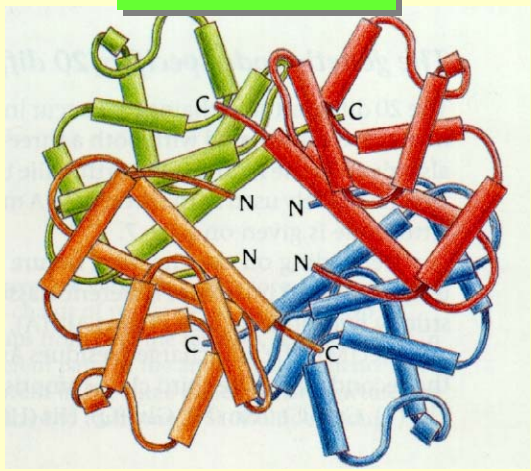
Secondary



Tertiary



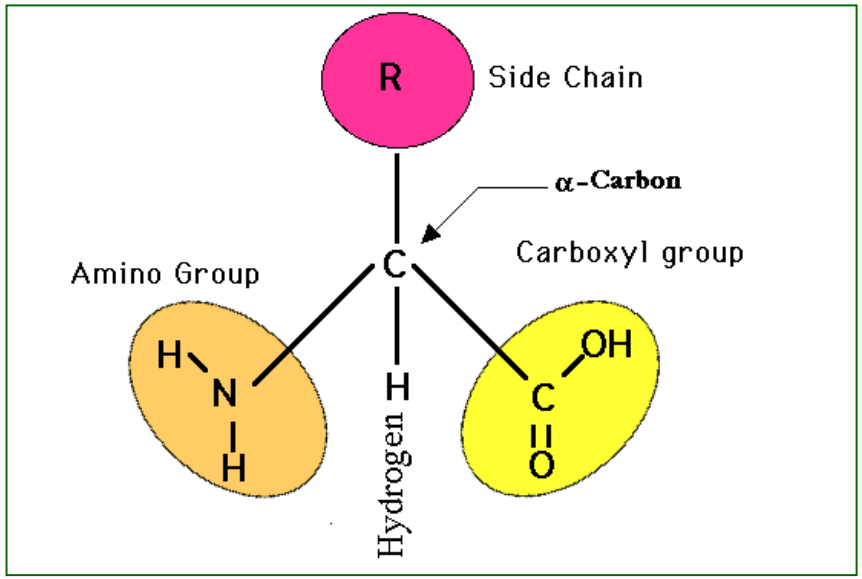
Quaternary



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.



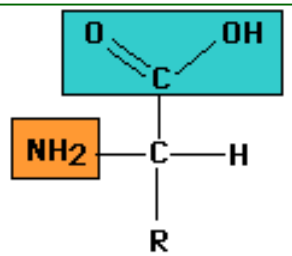
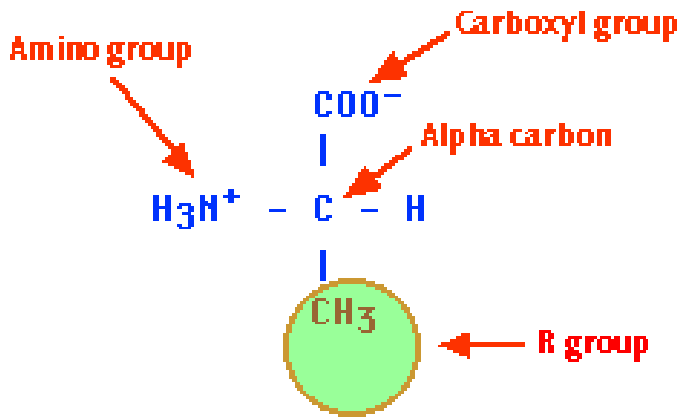


Fig. General formula for an amino acid molecule. "R" represents the variable groups that are attached to this basic molecule to make up the 20 common amino acids



Angles ϕ and ψ in the polypeptide chain

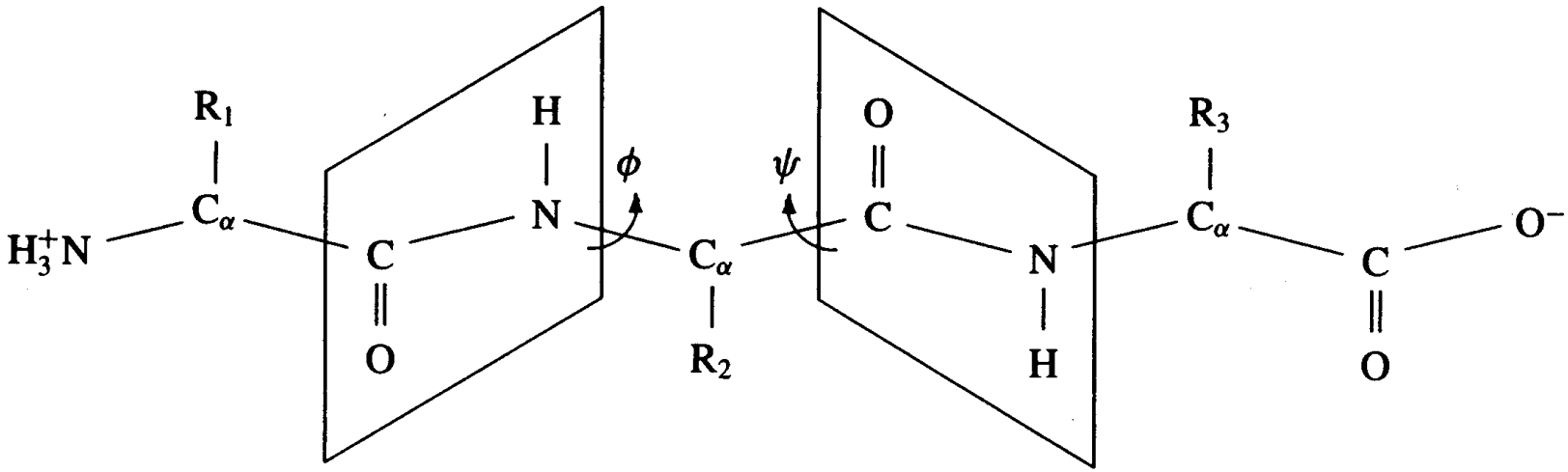
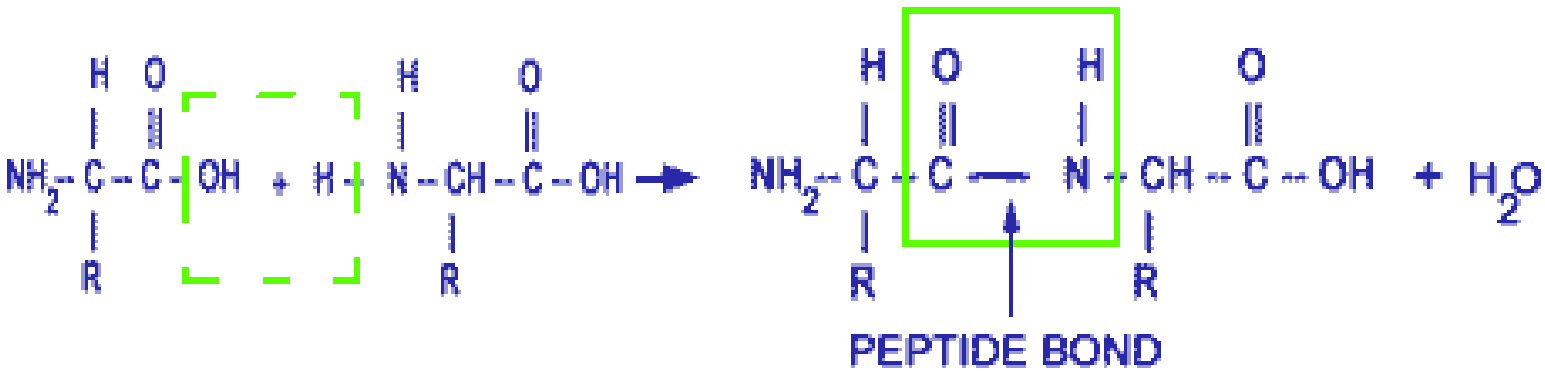
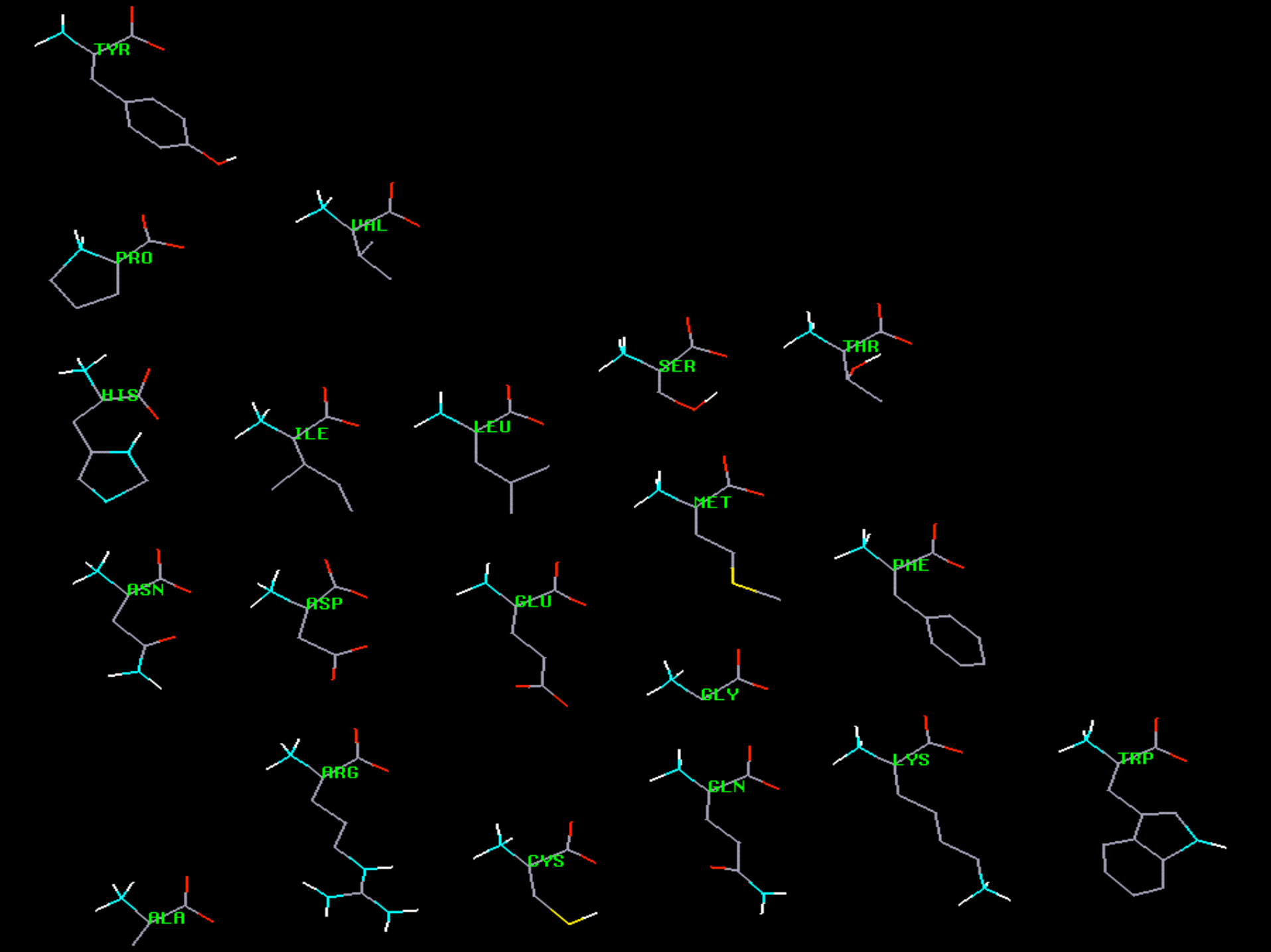


FIGURE 1.2

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

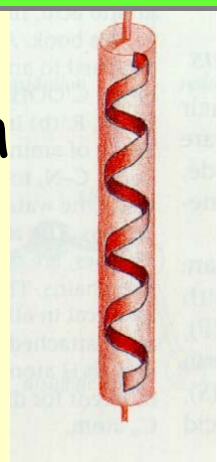
- **Primary structure** is the sequence of amino acid residues of the protein, e.g., Flavodoxin:

  AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADA…

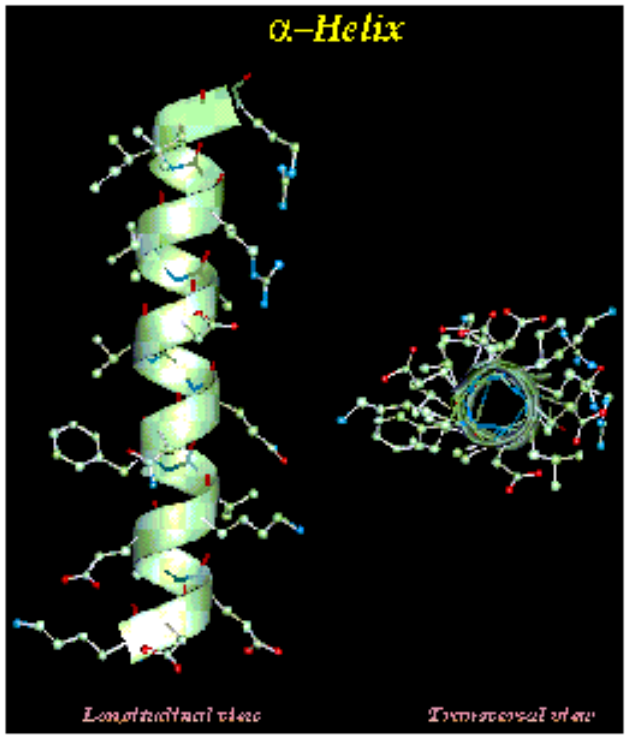
- Different regions of the sequence form local regular **secondary structures**, such
  - Alpha helix, beta strands, etc.

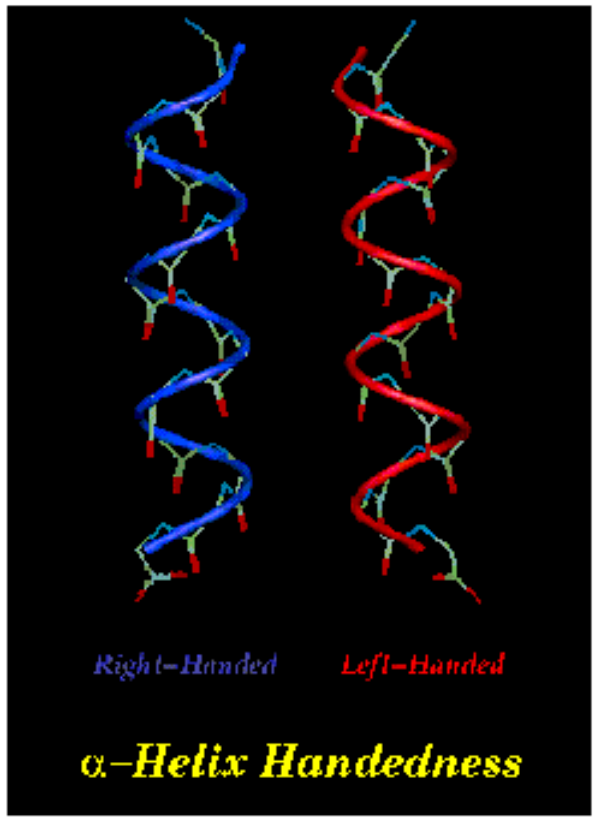
  AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADA…

Secondary



Alpha helices





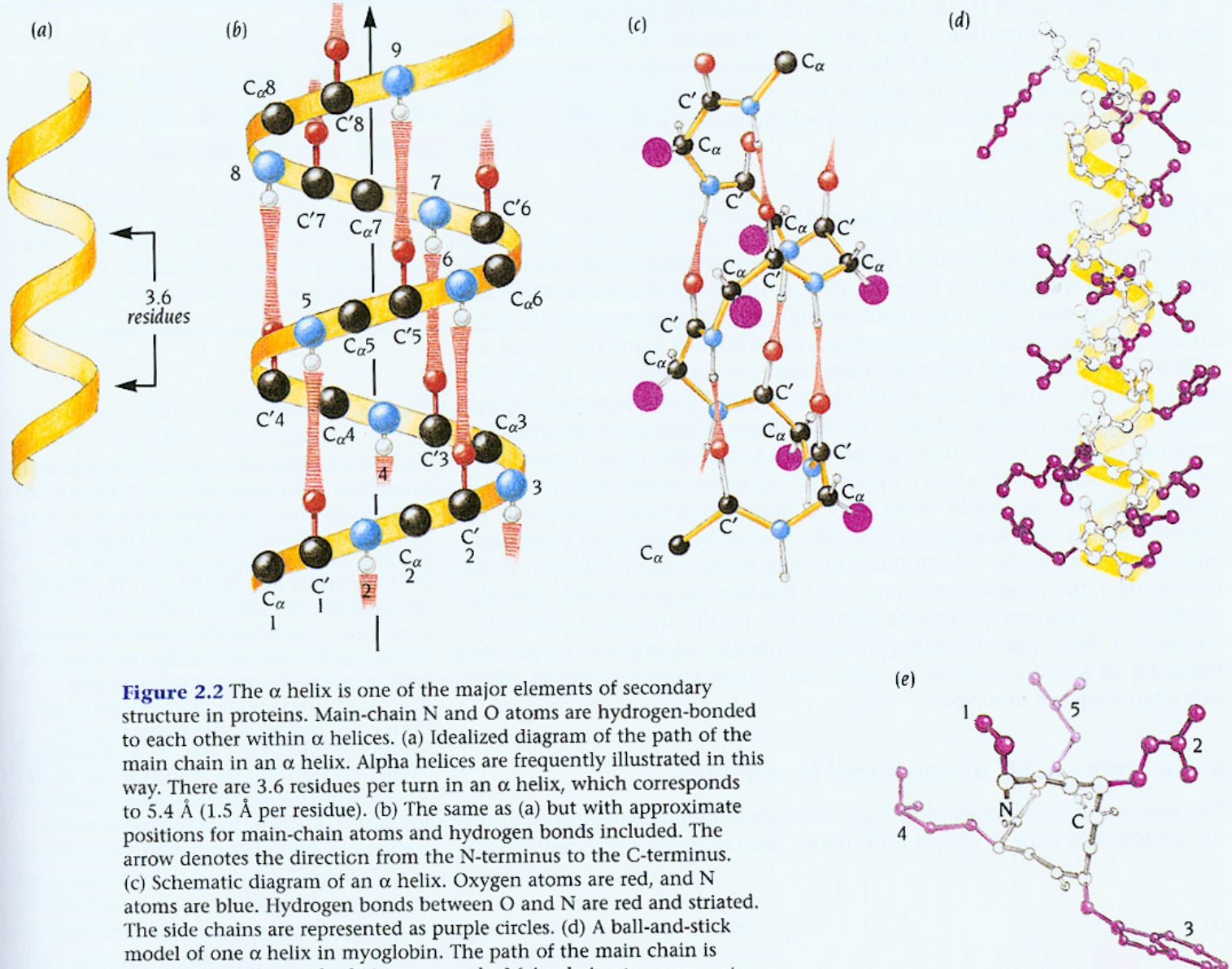
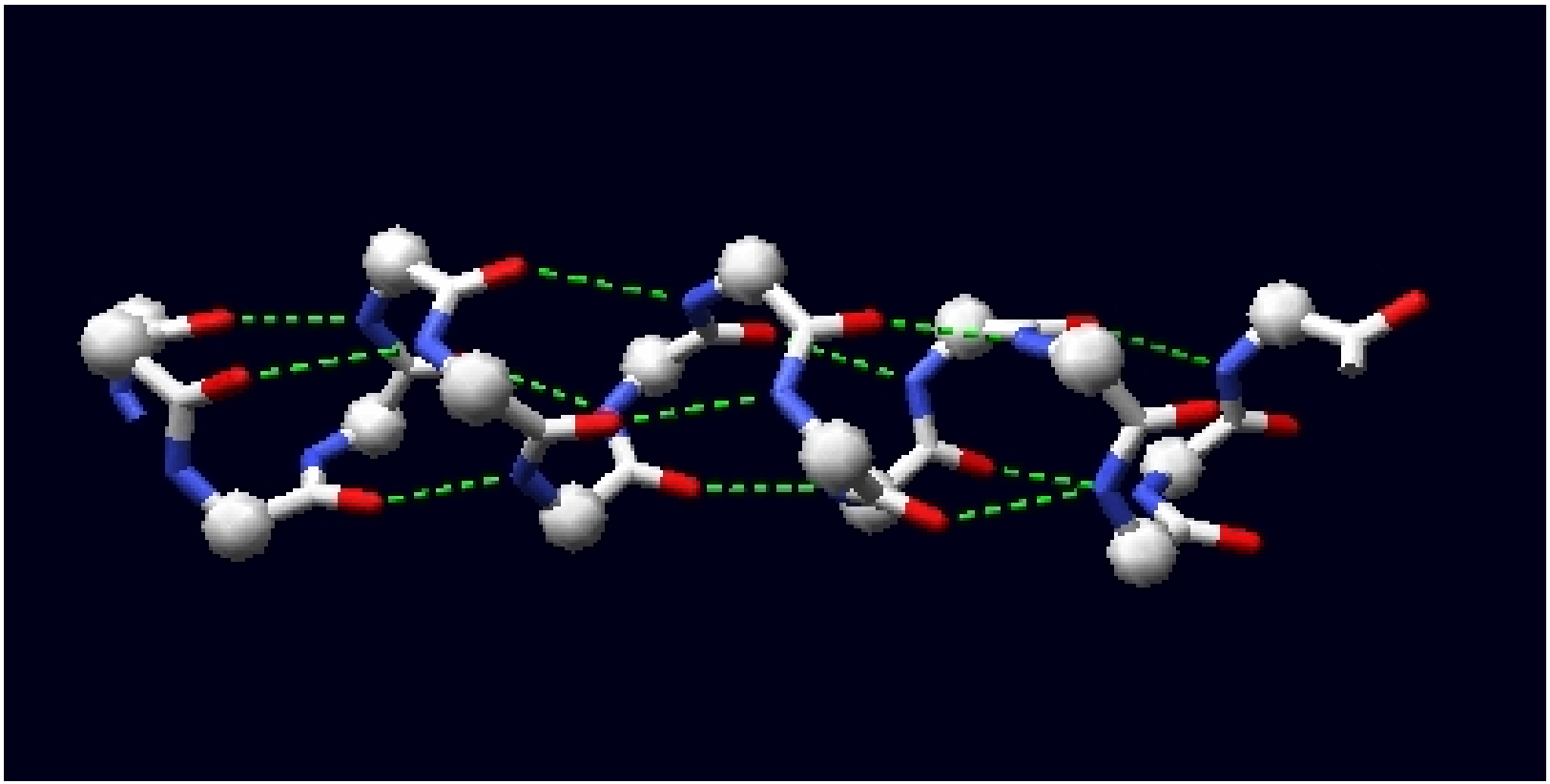
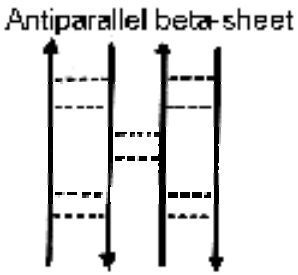


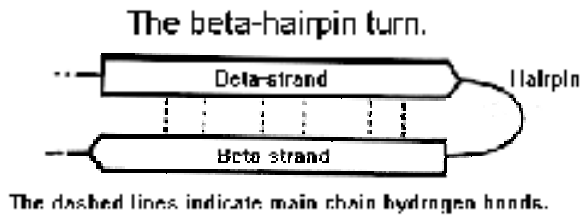
Figure 2.2 The α helix is one of the major elements of secondary structure in proteins. Main-chain N and O atoms are hydrogen-bonded to each other within α helices. (a) Idealized diagram of the path of the main chain in an α helix. Alpha helices are frequently illustrated in this way. There are 3.6 residues 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.

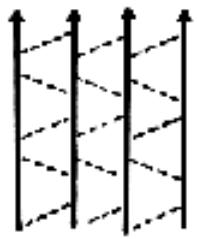
Alpha Helix

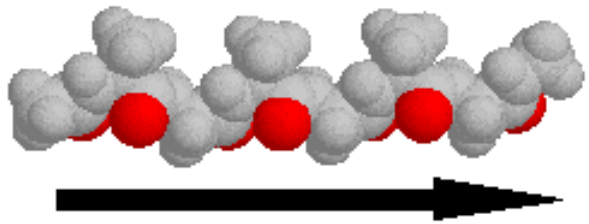


Beta sheet



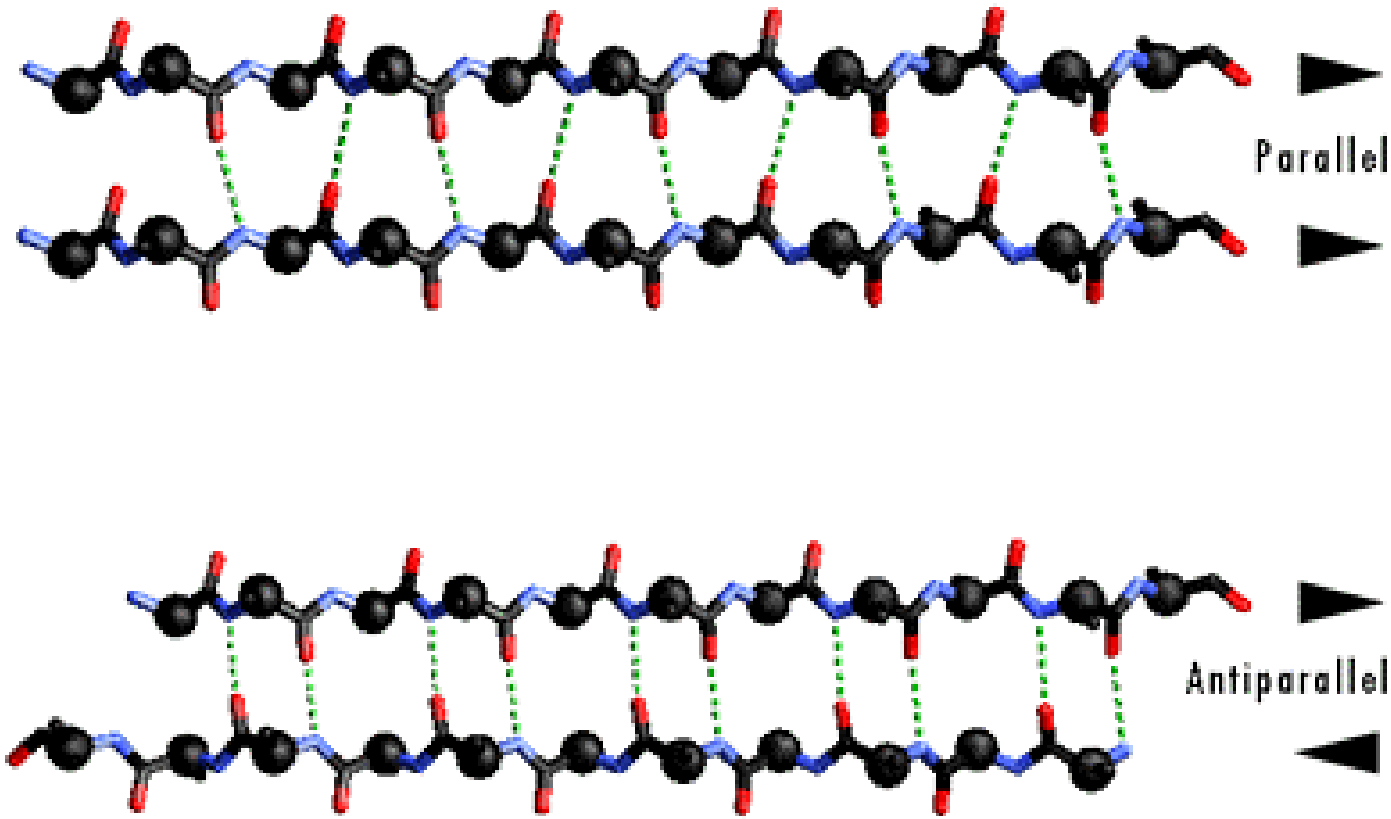






 17

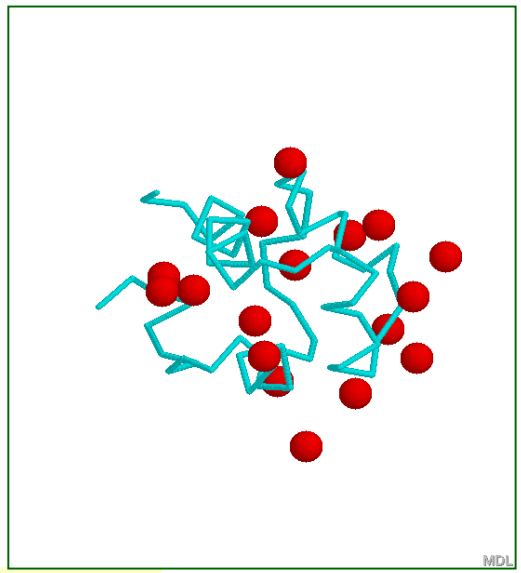
Beta Strand



Proteins

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

Myoglobin

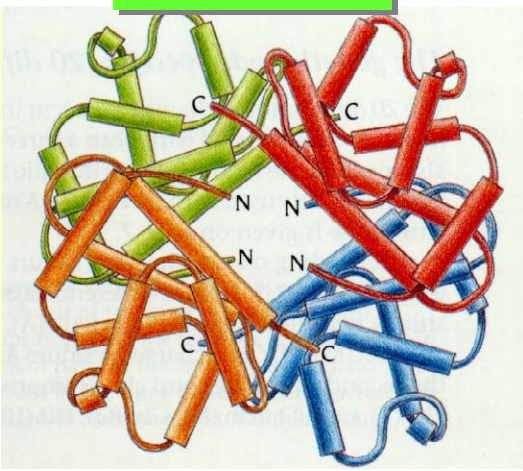


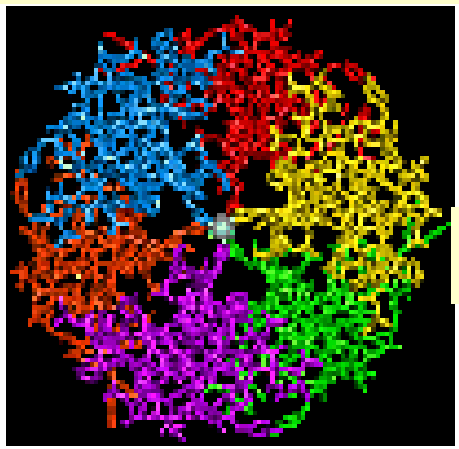
Lambda Cro

Quaternary Structures in Proteins

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

Quaternary





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).

PDB: Protein Data Bank

- Database of protein tertiary and quaternary structures and protein complexes. http://www.rcsb.org/pdb/
- Over 29,000 structures as of Feb 1, 2005.
- Structures determined by
  - NMR Spectroscopy
  - X-ray crystallography
  - Computational prediction methods
- Sample PDB file: Click here [ ]

Active Sites

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

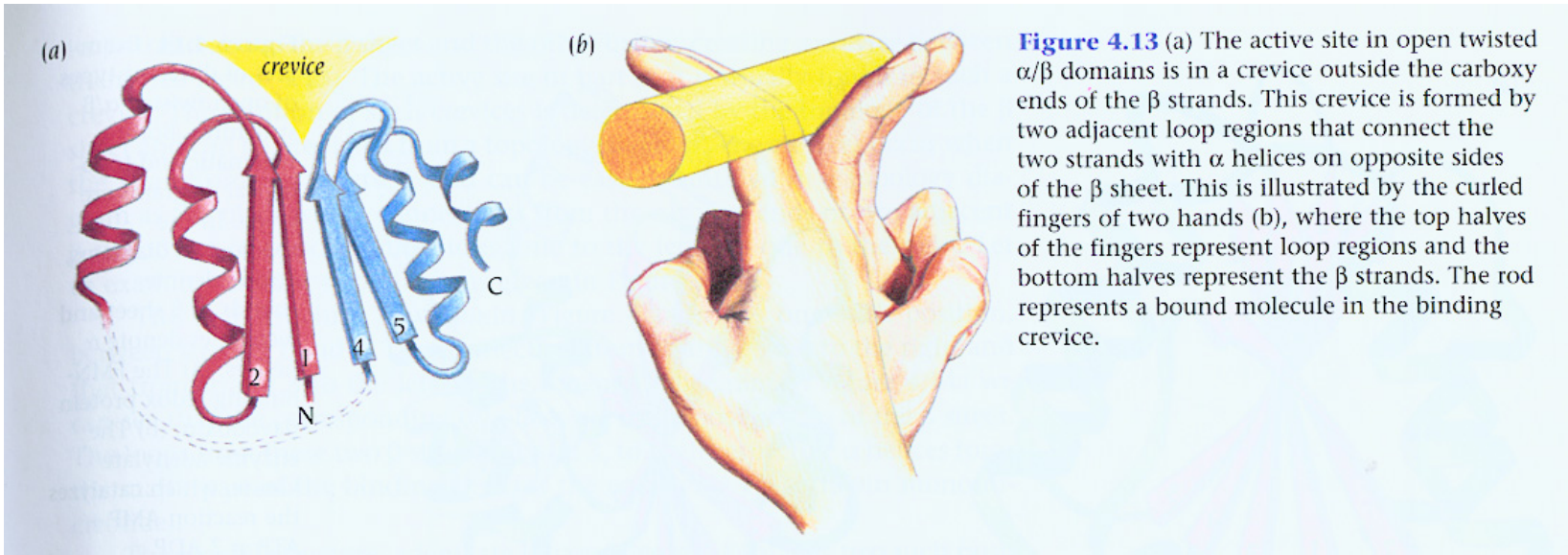
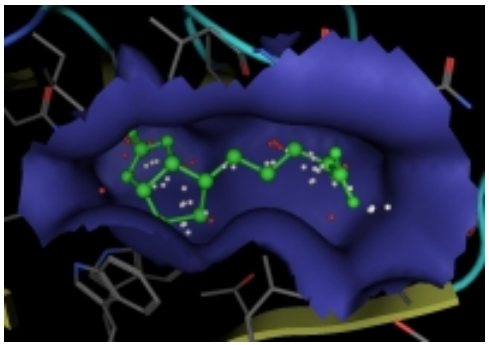
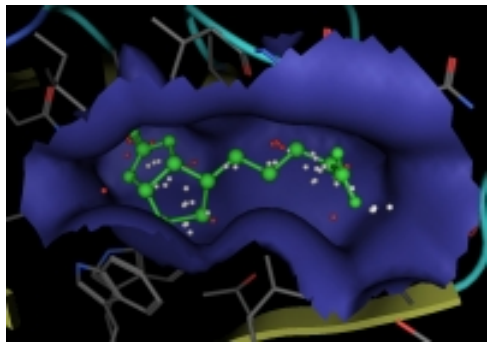
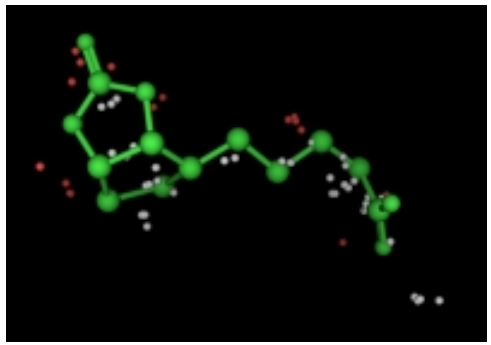


Figure 4.13 (a) The active site in open twisted α/β domains is in a crevice outside the carboxy ends of the β strands. This crevice is formed by two adjacent loop regions that connect the two strands with α helices on opposite sides of the β sheet. This is illustrated by the curled fingers of two hands (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.

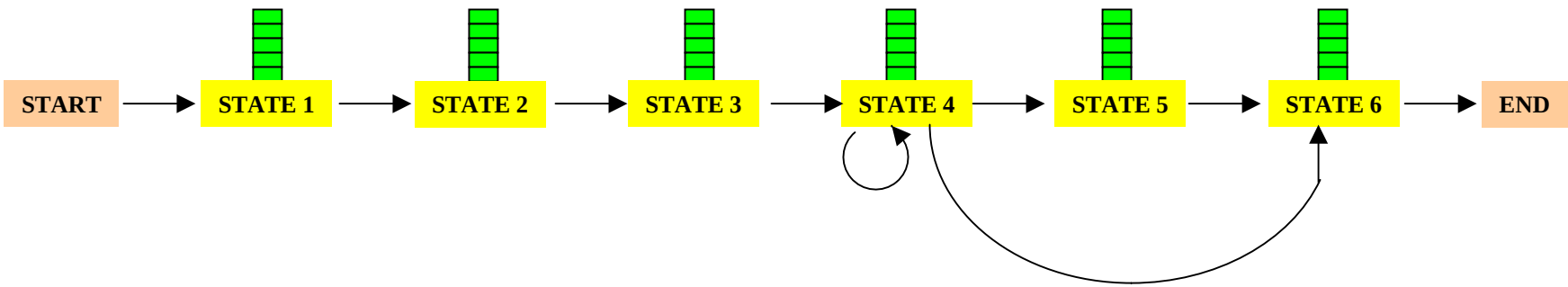
Active Sites

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.

Simple Models

- Helps to model simple sequence features.
  - single sequences e.g. **TTGACA** or **TATATT** **[??]**
  - sets of sequences e.g. **[AT] C [GC] TC [AGC]**
  - sets of sequences with inserts e.g. **GCA [AT] [AT]* AG**
  - & deletes too, e.g. **TATA [G –] T**



- long sequences with a sequence of domains **H-B-T-B-H**

Profile Method

PROFILE METHOD, [M. Gribskov et al., '90]

Location in Seq.	Sequence 1	2	3	4	5	6	7	Protein Name
14	G	V	S	A	S	A	V	Ka RbtR
32	G	V	S	E	M	T	I	Ec DeoR
33	G	V	S	P	G	T	I	Ec RpoD
76	G	A	G	I	A	T	I	Ec TrpR
178	G	C	S	R	E	T	V	Ec CAP
205	C	L	S	P	S	R	L	Ec AraC
210	C	L	S	P	S	R	L	St AraC
13	G	V	N	K	E	T	I	Br MerR

FREQUENCY TABLE

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	0	2	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	4	0	0
3	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	6	0	0	0	0
4	1	0	0	1	0	0	0	1	1	0	0	0	3	0	1	0	0	0	0	0
5	1	0	0	2	0	1	0	0	0	0	1	0	0	0	0	3	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	5	0	0	0
7	0	0	0	0	0	0	0	4	0	2	0	0	0	0	0	0	0	2	0	0

7

Profile Method

FREQUENCY TABLE

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	0	2	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	4	0	0
3	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	6	0	0	0	0
4	1	0	0	1	0	0	0	1	1	0	0	0	3	0	1	0	0	0	0	0
5	1	0	0	2	0	1	0	0	0	0	1	0	0	0	0	3	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	5	0	0	0
7	0	0	0	0	0	0	0	4	0	2	0	0	0	0	0	0	0	2	0	0

WEIGHT MATRIX

	A	C	E	G	I	K	L	M	N	P	R	S
1	0	108	0	101	0	0	0	0	0	0	0	0
2	21	78	0	0	0	0	44	0	0	0	0	0
3	0	0	0	23	0	0	0	0	46	0	0	102
4	21	0	32	0	38	32	0	0	0	86	39	0
5	21	0	62	23	0	0	0	74	0	0	0	72
6	21	0	0	0	0	0	0	0	0	0	69	0
7	0	0	0	0	98	0	44	0	0	0	0	0

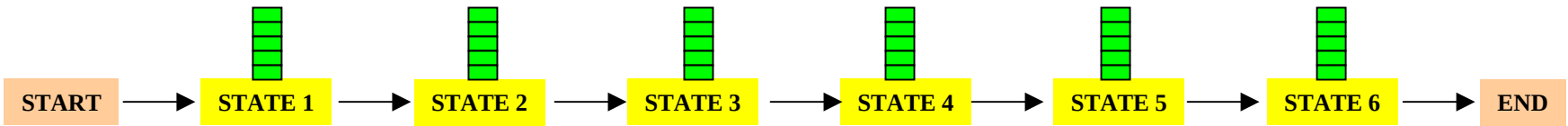
$$Weight[i, AA] = \log\left(\frac{Freq[i,AA]}{p[AA] \cdot N}\right) \cdot 100$$

8

Profile HMMs

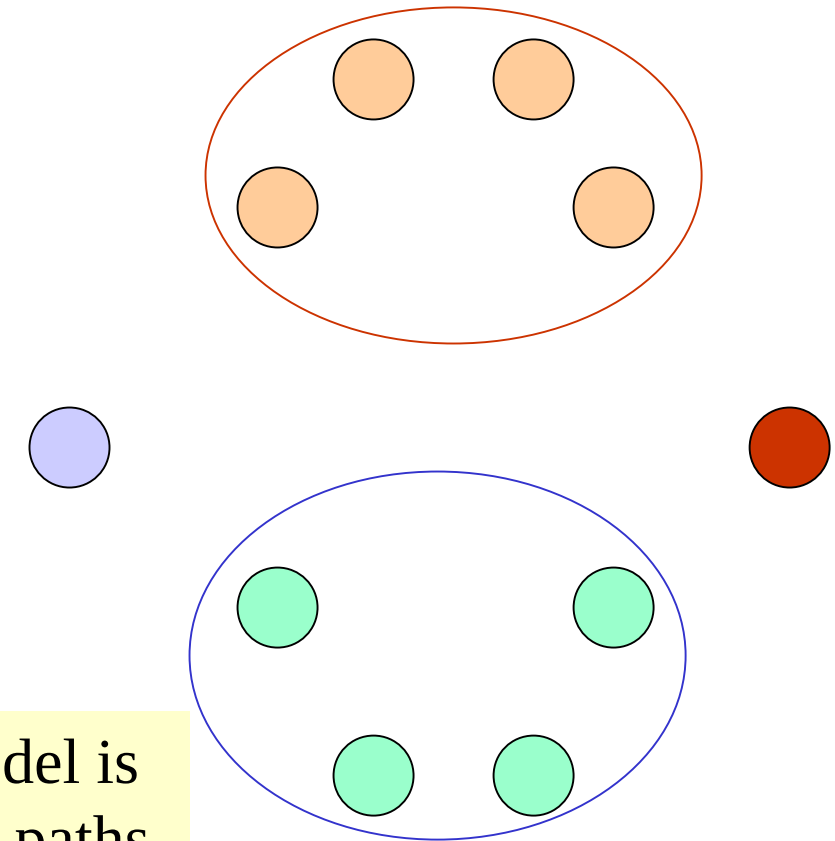
PROFILE METHOD, [M. Gribskov et al., '90]

Location in Seq.	Sequence 1	2	3	4	5	6	Protein Name
14	G	V	S	A	S	A	Ka RbtR
32	G	V	S	E	M	T	Ec DeoR
33	G	V	S	P	G	T	Ec RpoD
76	G	A	G	I	A	T	Ec TrpR
178	G	C	S	R	E	T	Ec CAP
205	C	L	S	P	S	R	Ec AraC
210	C	L	S	P	S	R	St AraC
13	G	V	N	K	E	T	Br MerR



Hidden Markov Model (HMM)

- States
- Transitions
- Transition Probabilities
- Emissions
- Emission Probabilities

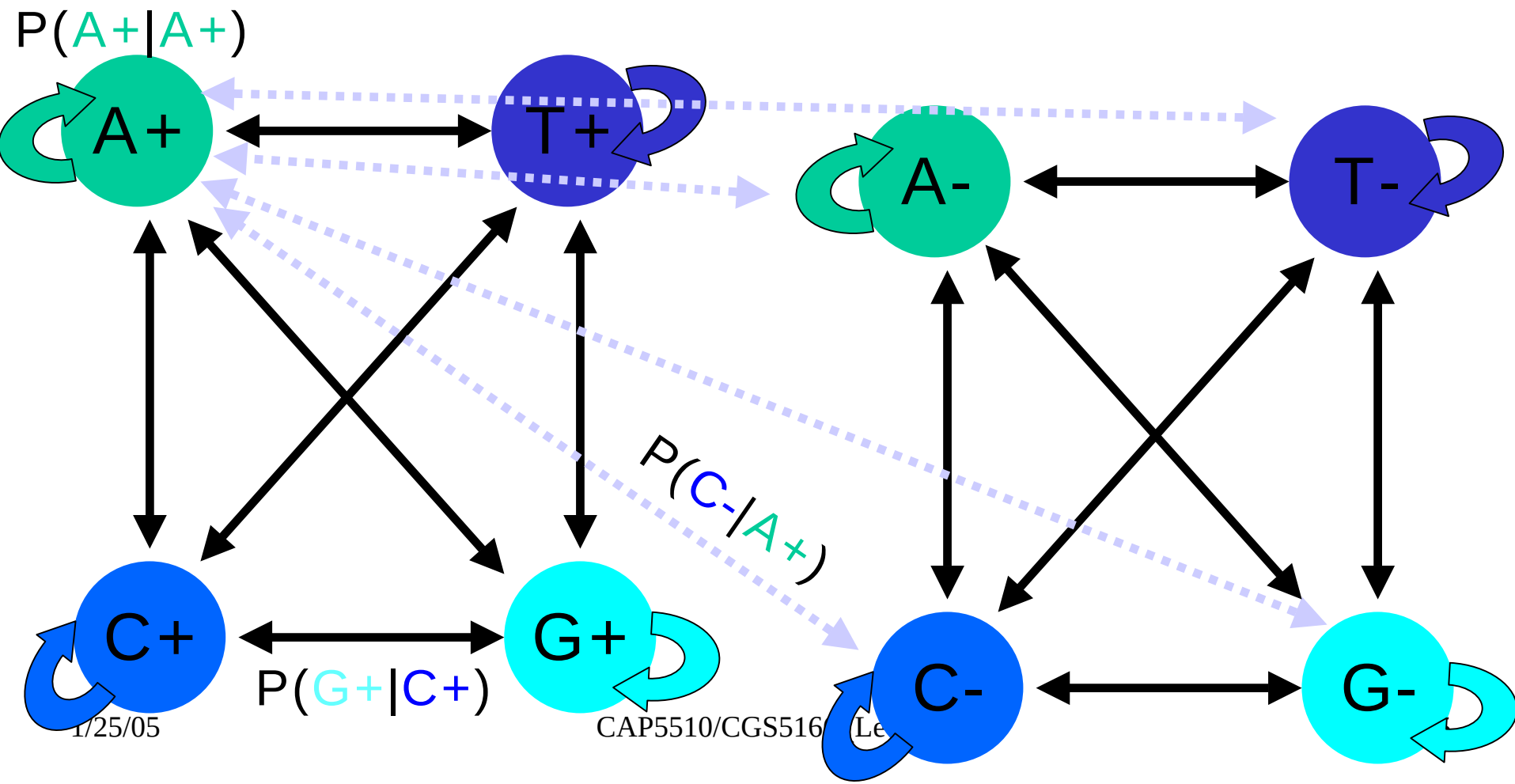


- What is hidden about HMMs?

Answer: The path through the model is hidden since there are many valid paths.

CpG Island + in an ocean of –
First order Markov Model

MM=16, HMM= 64 transition probabilities (adjacent bp)



How to Solve Problem 2?

- Solve the following problem:

Input: Hidden Markov Model M,
parameters Θ, emitted sequence S

Output: Most Probable Path Π

How: Viterbi's Algorithm (Dynamic Programming)

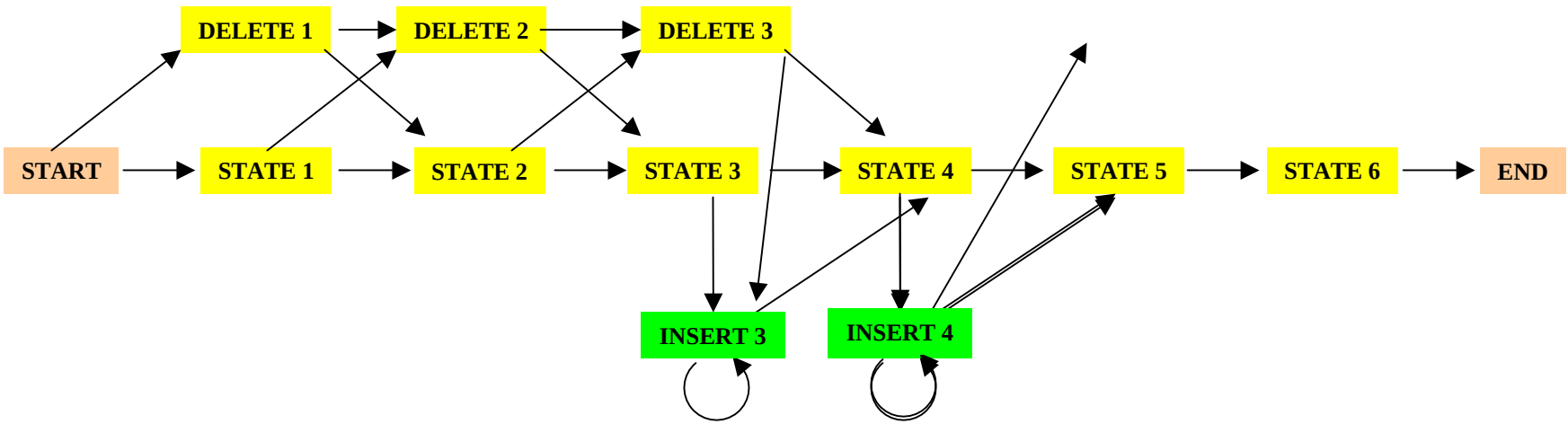
Define $\Pi[i,j]$ = MPP for first j characters of S ending in state i

Define $P[i,j]$ = Probability of $\Pi[i,j]$

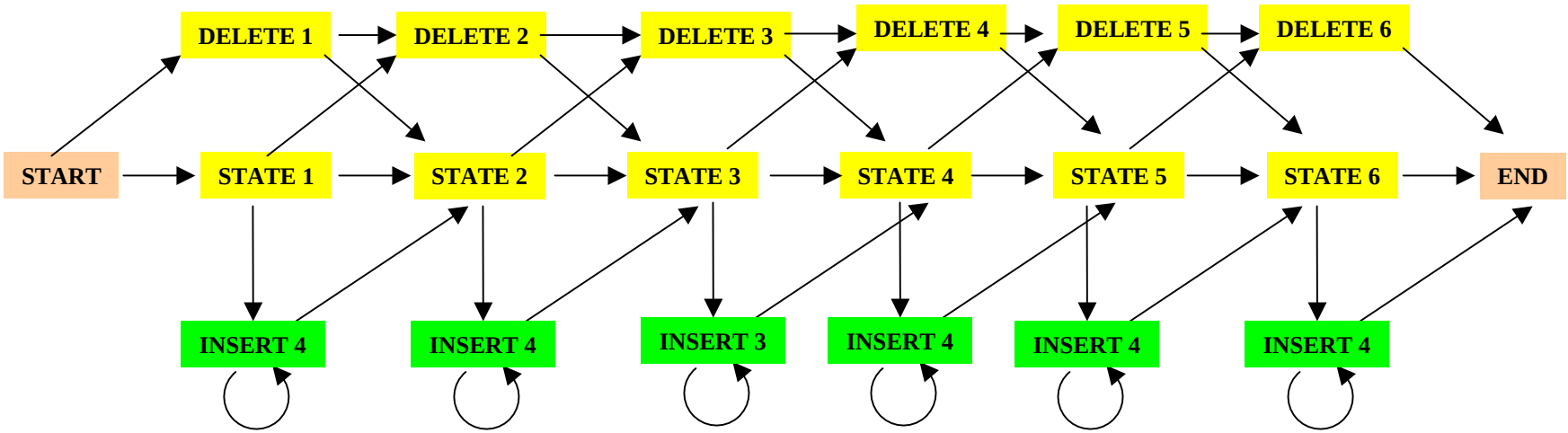
- Compute state i with largest $P[i,j]$.

Profile HMMs with InDels

- Insertions
- Deletions
- Insertions & Deletions



Profile HMMs with InDels

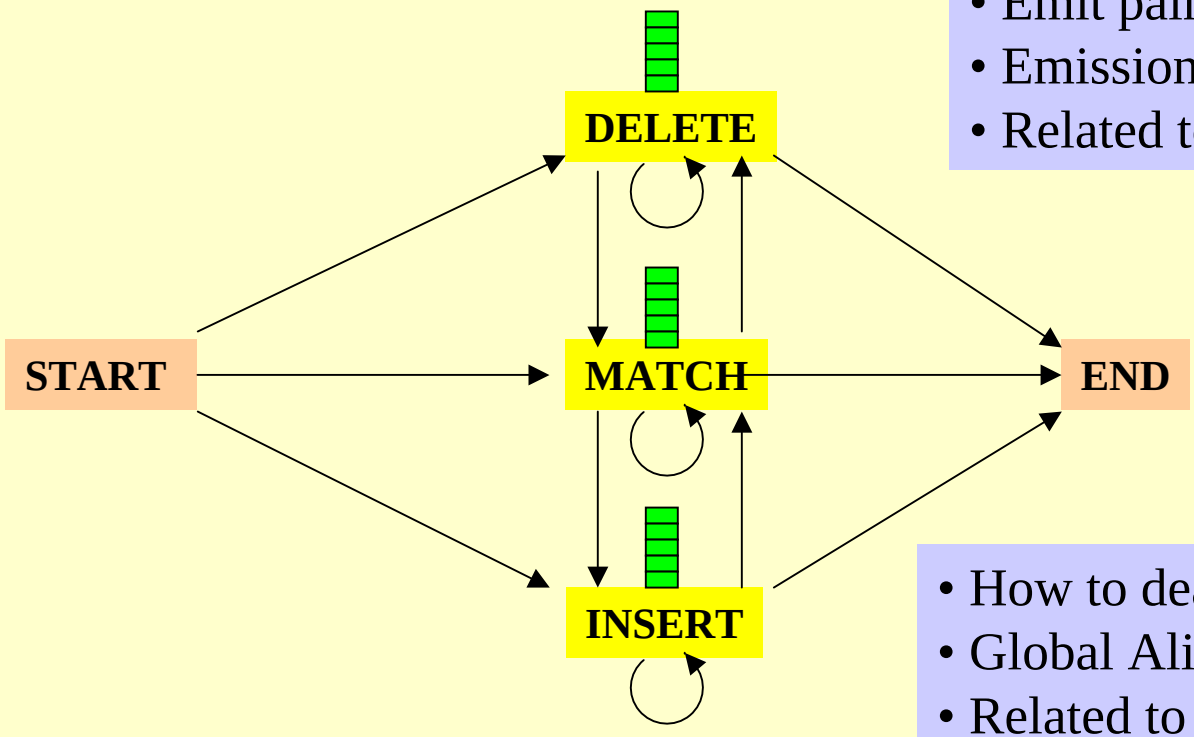


Missing transitions from DELETE j to INSERT j and from INSERT j to DELETE j+1.

How to model Pairwise Sequence Alignment

LEAPVE

LAPVIE



Pair HMMs
- Emit pairs of synbols
- Emission probs?
- Related to Sub. Matrices

- How to deal with InDels?
- Global Alignment? Local?
- Related to Sub. Matrices

How to model Pairwise Local Alignments?

How to model Pairwise Local Alignments with gaps?