CAP 5510: Introduction to Bioinformatics

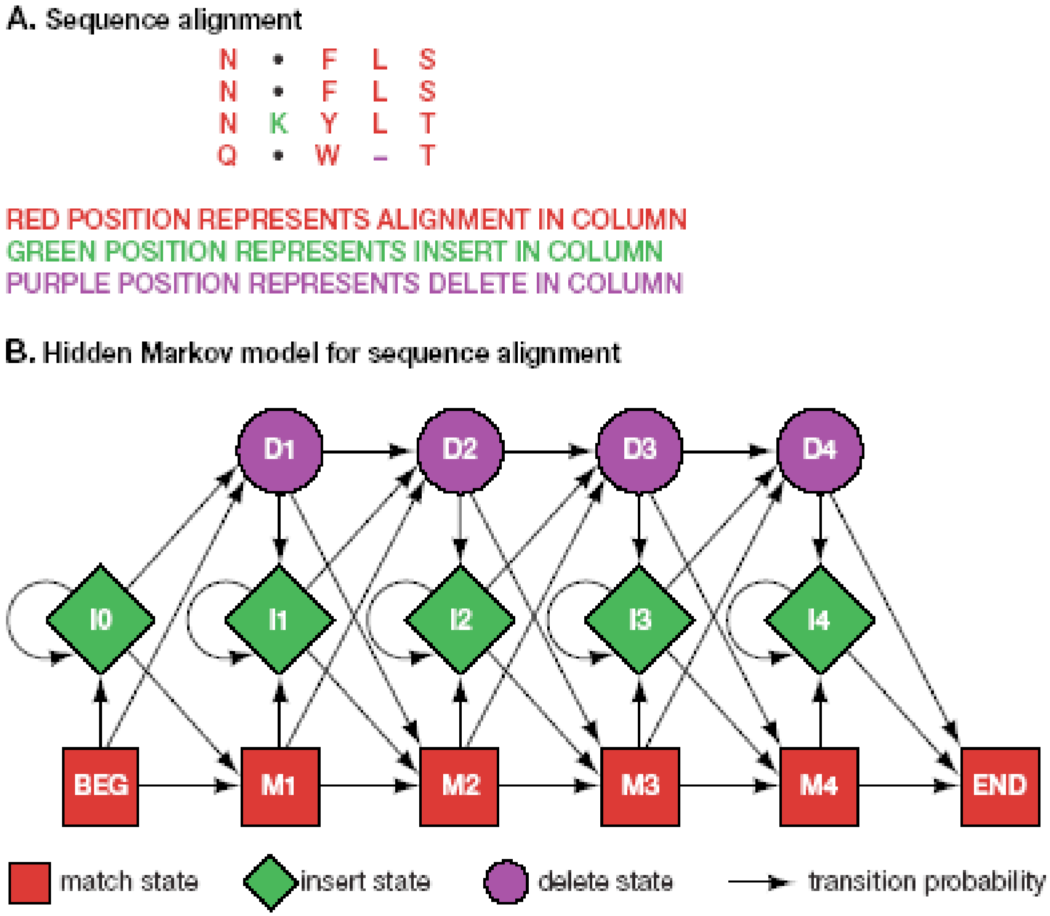
Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS07.html

HMM for Sequence Alignment



A. Sequence alignment

```
N  •  F  L  S
N  •  F  L  S
N  K  Y  L  T
Q  •  W  –  T
```

RED POSITION REPRESENTS ALIGNMENT IN COLUMN
GREEN POSITION REPRESENTS INSERT IN COLUMN
PURPLE POSITION REPRESENTS DELETE IN COLUMN

B. Hidden Markov model for sequence alignment

FIGURE 5.16. Relationship between the sequence alignment and the hidden Markov model of the alignment (Krogh et al. 1994). This particular form for the HMM was chosen to represent the sequence, structural, and functional variation expected in proteins. The model accommodates the identities, mismatches, insertions, and deletions expected in a group of related proteins. (*A*) A section of an msa. The illustration shows the columns generated in an msa. Each column may include matches and mismatches (*red* positions), insertions (*green* positions), and deletions (*purple* positions). (*B*) The HMM. Each column in the model represents the possibility of a match, insert, or delete in each column of the alignment in *A*. The HMM is a probabilistic representation of a section of the msa. Sequences can be generated from the HMM by starting at the beginning state labeled BEG and then by following

Problem 3: LIKELIHOOD QUESTION

- Input: Sequence S, model M, state i
- Output: Compute the probability of reaching state i with sequence S using model M
  - Backward Algorithm (DP)

Problem 4: LIKELIHOOD QUESTION

- Input: Sequence S, model M
- Output: Compute the probability that S was emitted by model M
  - Forward Algorithm (DP)

Problem 5: LEARNING QUESTION

- Input: model structure M, Training Sequence *S*
- Output: Compute the parameters Θ
- Criteria: ML criterion
  - maximize P(*S* | M, Θ)   HOW???

Problem 6: DESIGN QUESTION

- Input: Training Sequence *S*
- Output: Choose model structure M, and compute the parameters Θ
  - No reasonable solution
  - Standard models to pick from

Iterative Solution to the LEARNING QUESTION (Problem 5)

- Pick initial values for parameters Θ_0
- Repeat

```
        Run training set $S$ on model $M$
        Count # of times transition $i \Rightarrow j$ is made
        Count # of times letter x is emitted from state $i$
        Update parameters $\Theta$
```

- Until (some stopping condition)

Entropy

- Entropy measures the variability observed in given data.

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

- Entropy is useful in multiple alignments & profiles.

- Entropy is max when uncertainty is max.

G-Protein Couple Receptors

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

GPCR - Analysis

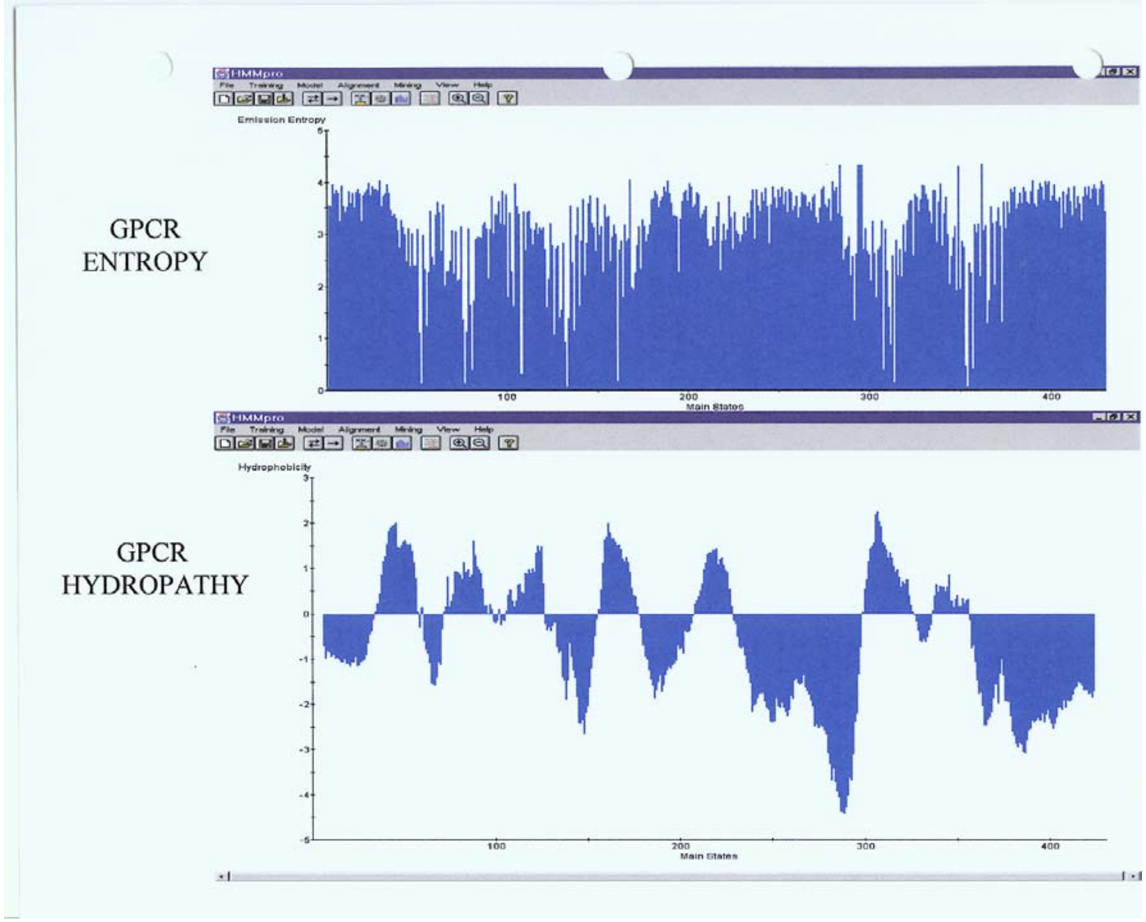
- Compute main state entropy values

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

- 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

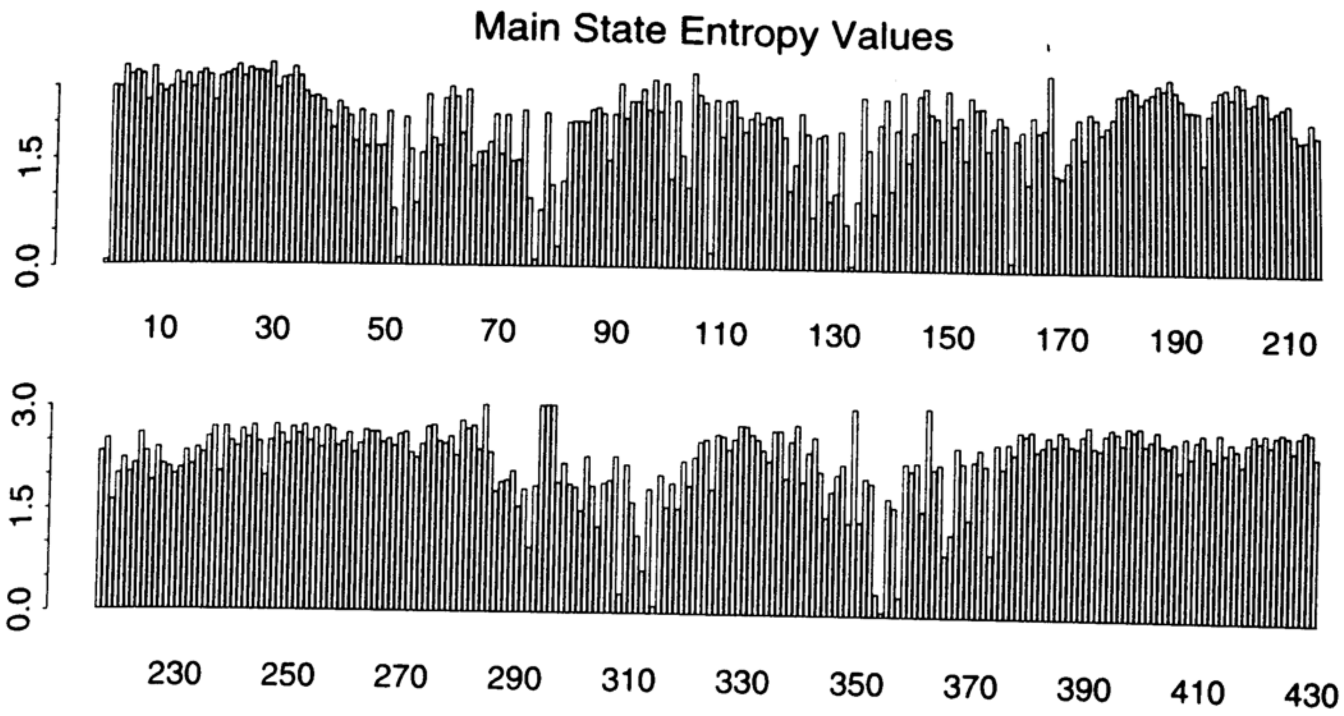


Figure 8.1: Entropy Profile of the Emission Probability Distributions Associated with the Main States of the HMM After 12 Cycles of Training.

GPCR Analysis (Cont’d)

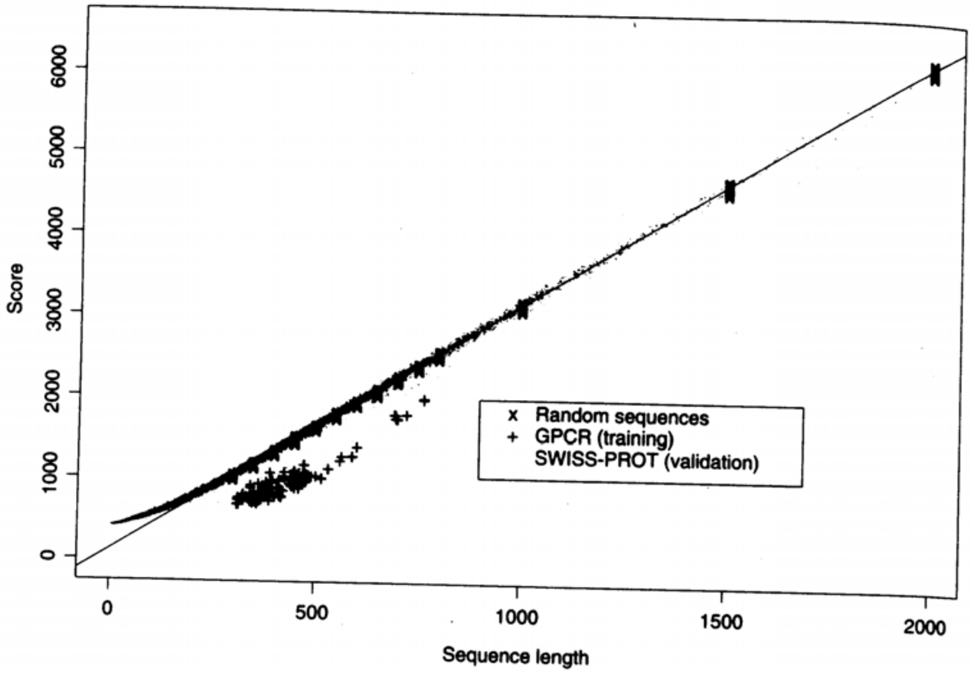


Figure 8.2: Scores (Negative Log-likelihoods of Optimal Viterbi Paths). Represented sequences consist of 142 GPCR training sequences, all sequences from the SWISS-PROT database of length less than or equal to 2000, and 220 randomly generated sequences with same average composition as the GPCRs of length 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
  - 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.

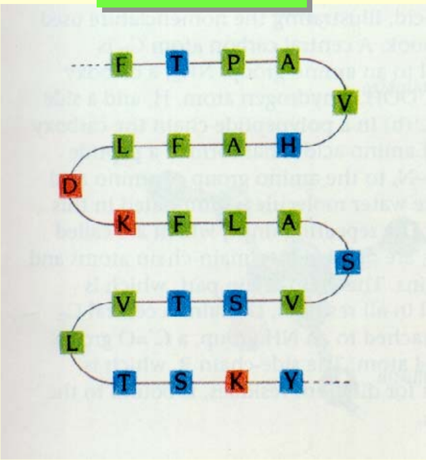
HMMs – Disadvantages

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

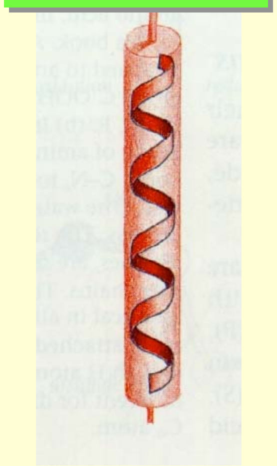
Protein Structures

- Sequences of amino acid residues
- 20 different amino acids

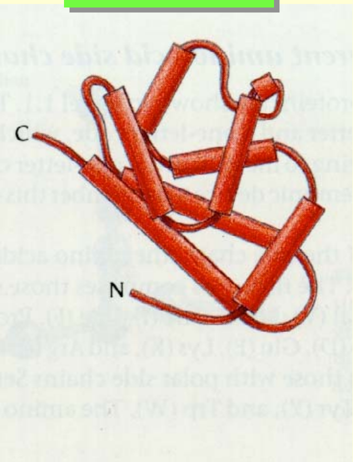
Primary



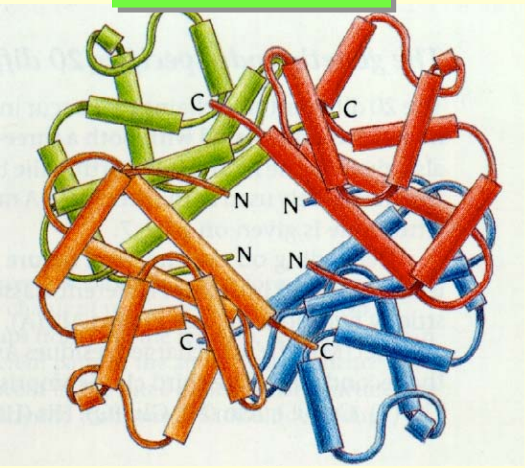
Secondary



Tertiary



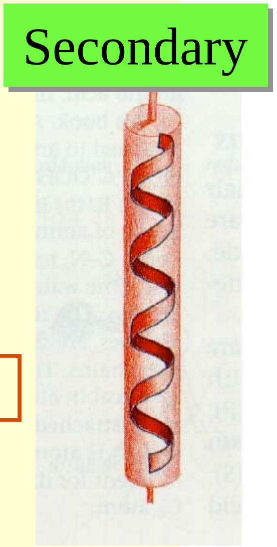
Quaternary



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…



More on Secondary Structures

θ **α-helix**

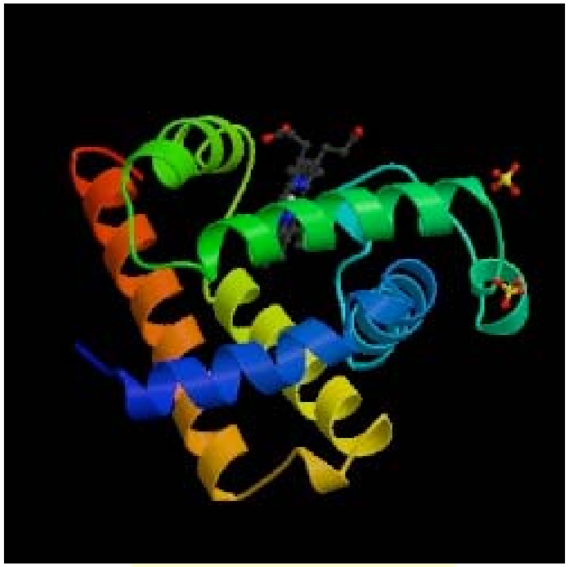
- Main chain with peptide bonds
- Side chains project outward from helix
- Stability provided by H-bonds between CO and NH groups of residues 4 locations away.

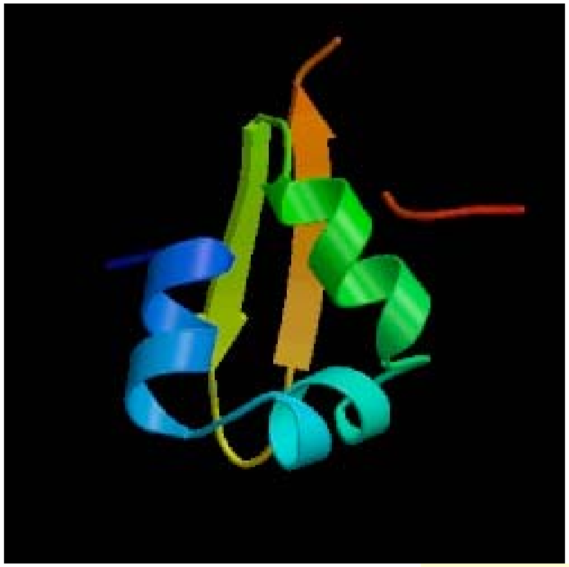
θ **β-strand**

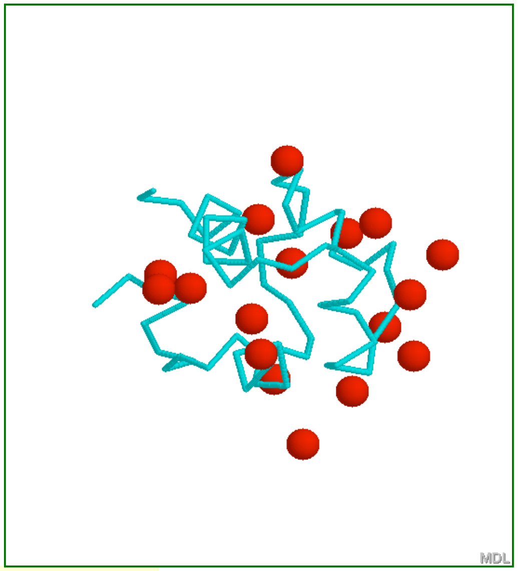
- Stability provided by H-bonds with one or more β-strands, forming β-sheets. Needs a β-turn.

Proteins

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







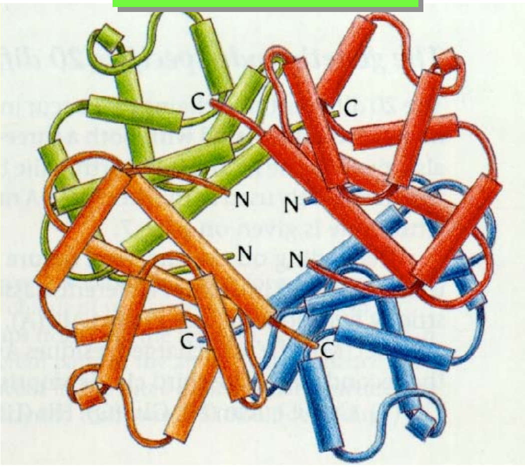
Myoglobin

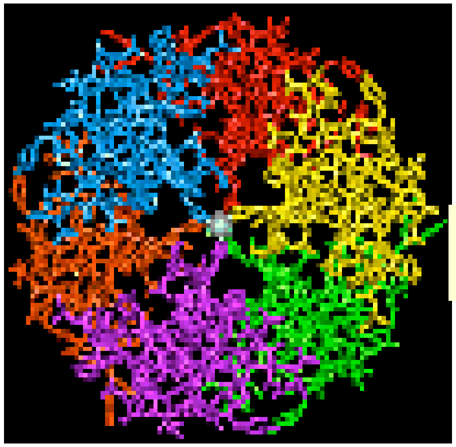
Lambda Cro

Quaternary Structures in Proteins

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

Quaternary





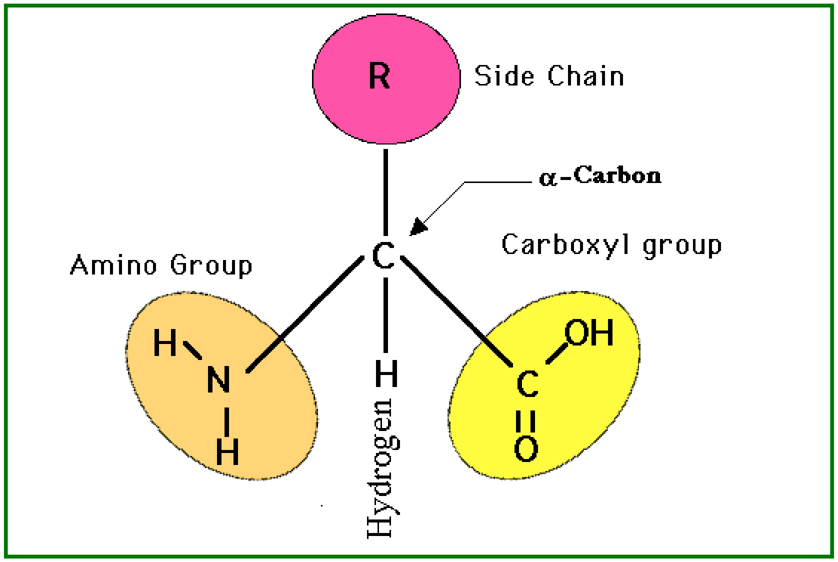
Insulin Hexamer

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

Structure of a single amino acid

All 3 figures are cartoons of an amino acid residue.



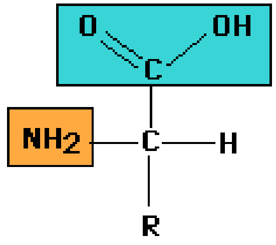
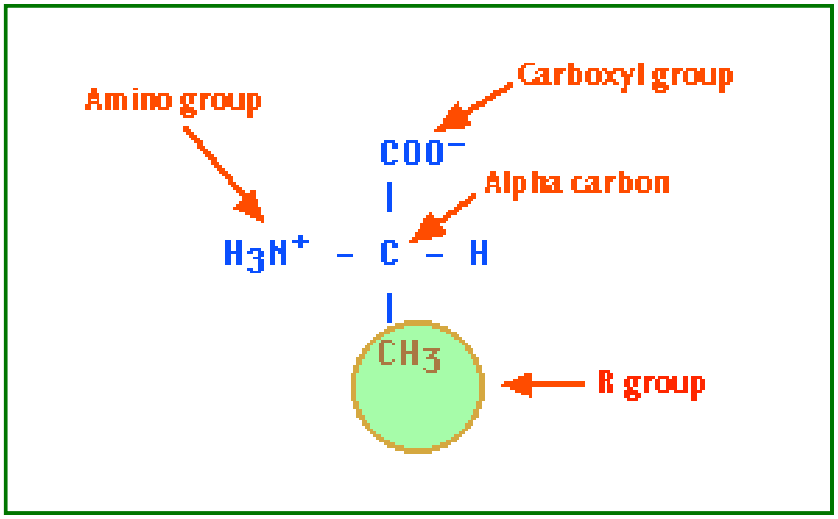
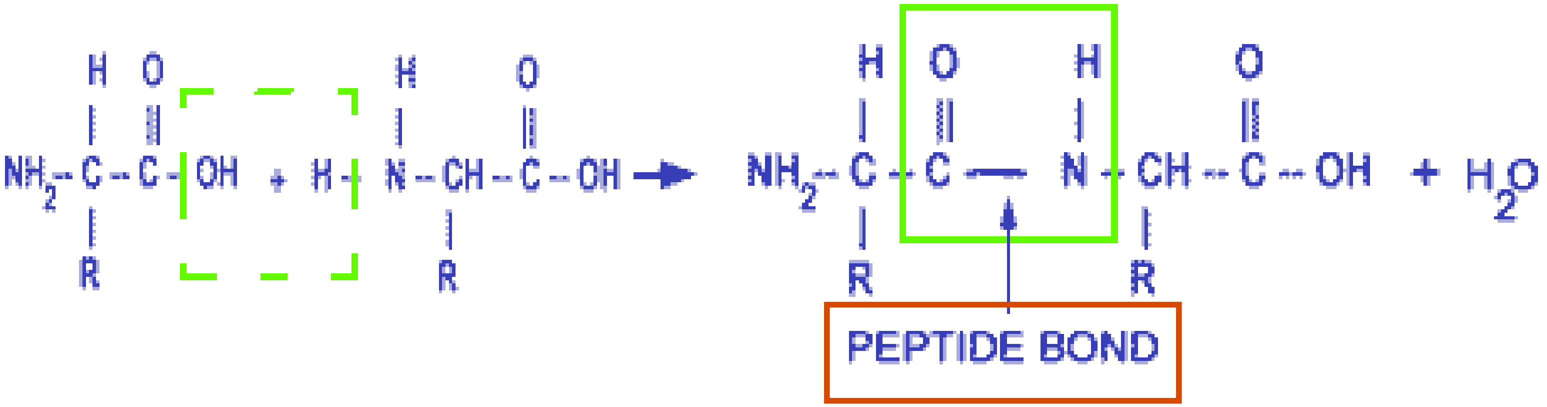


Fig. General formula for an amino acid molecule. "R" represents the variable groups that are attached to this basic molecule to make up the 20 common amino acids



Chains of amino acids



Amino acids vs Amino acid residues

Angles φ and ψ in the polypeptide chain

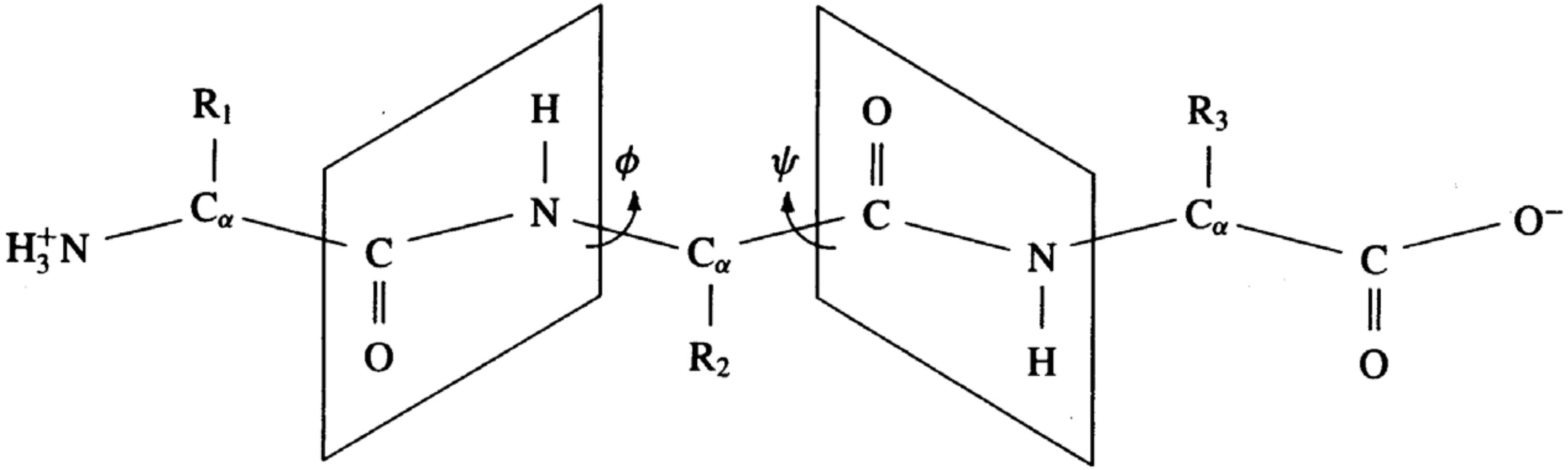
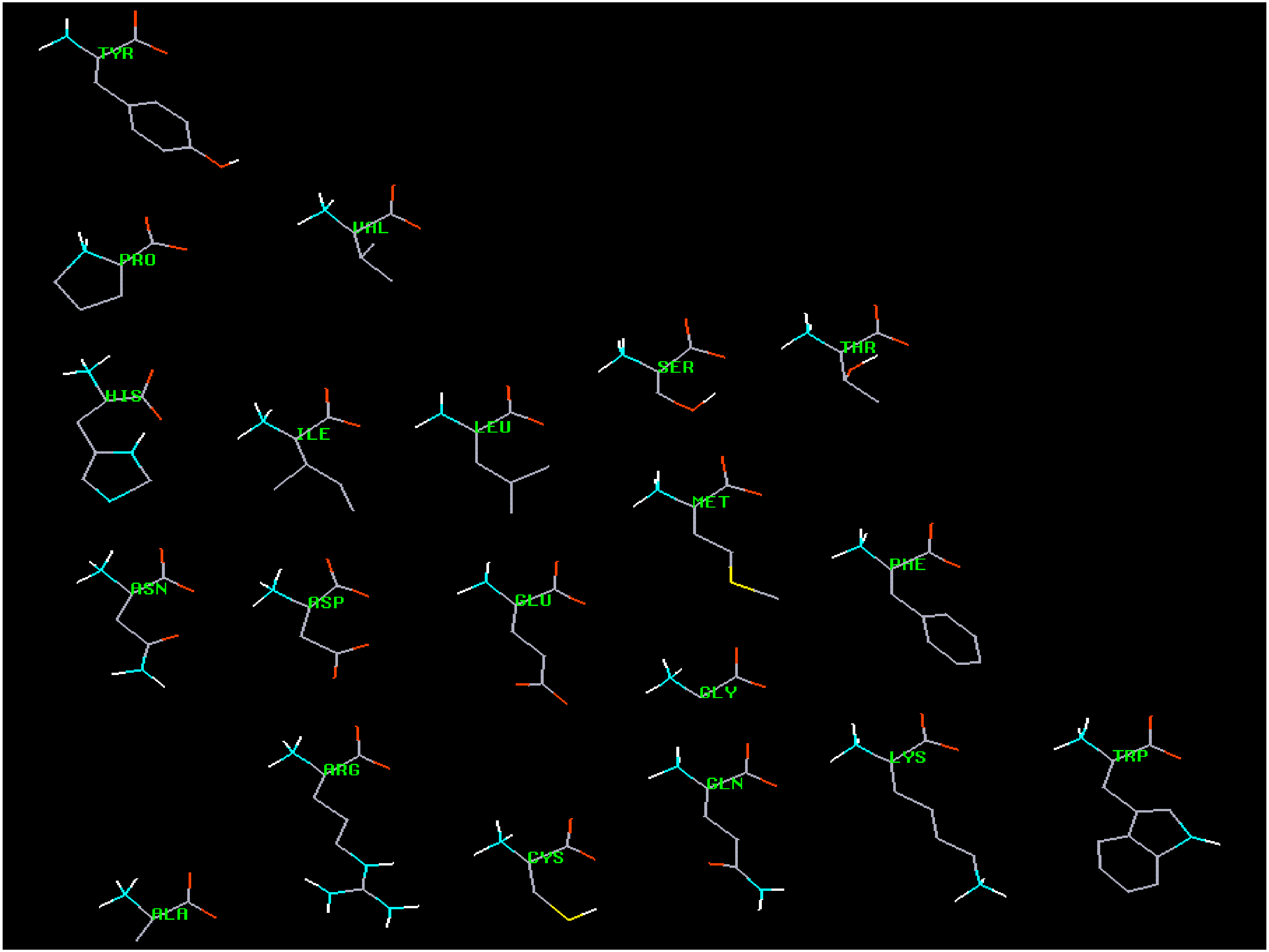
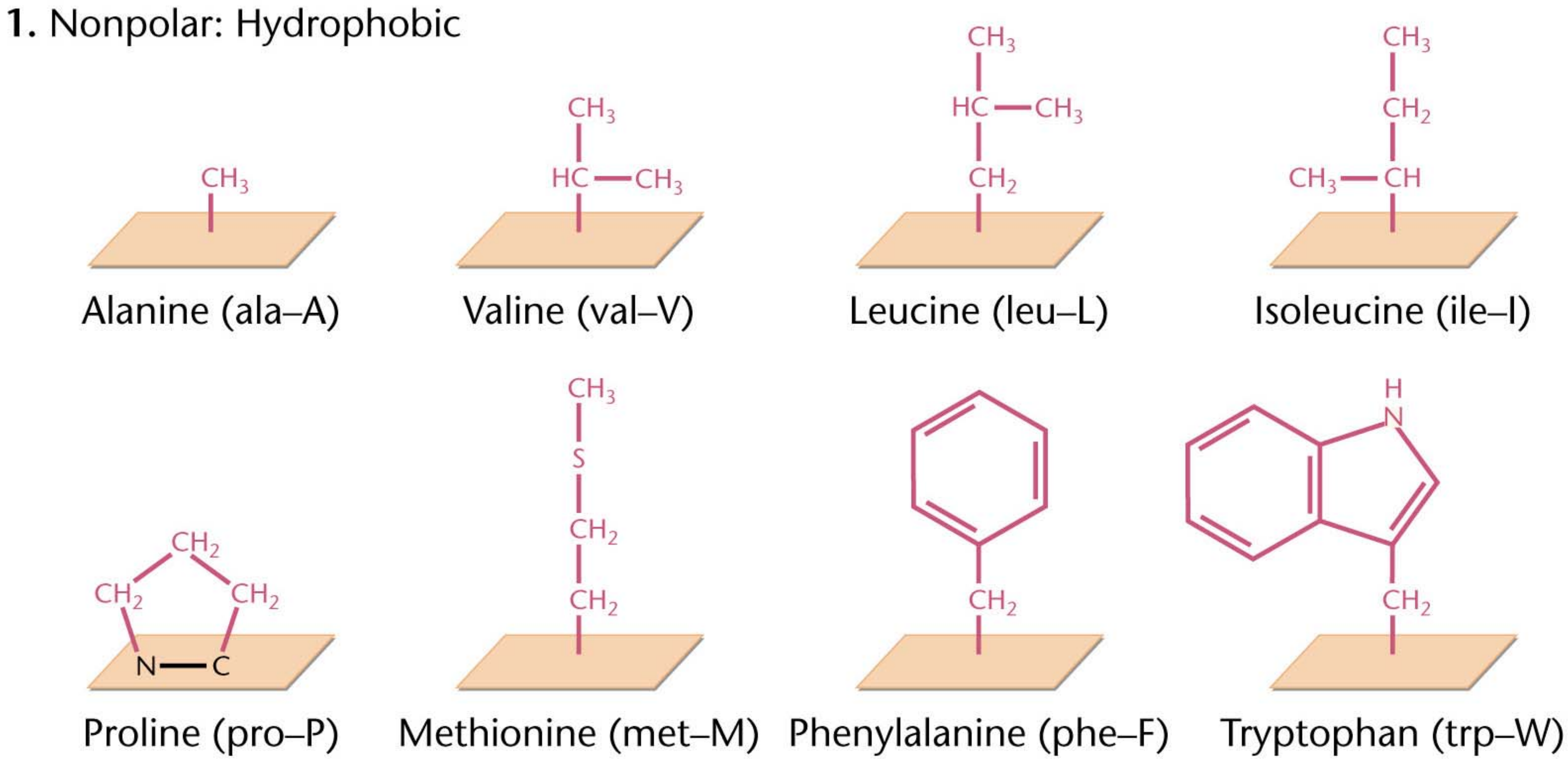


FIGURE 1.2

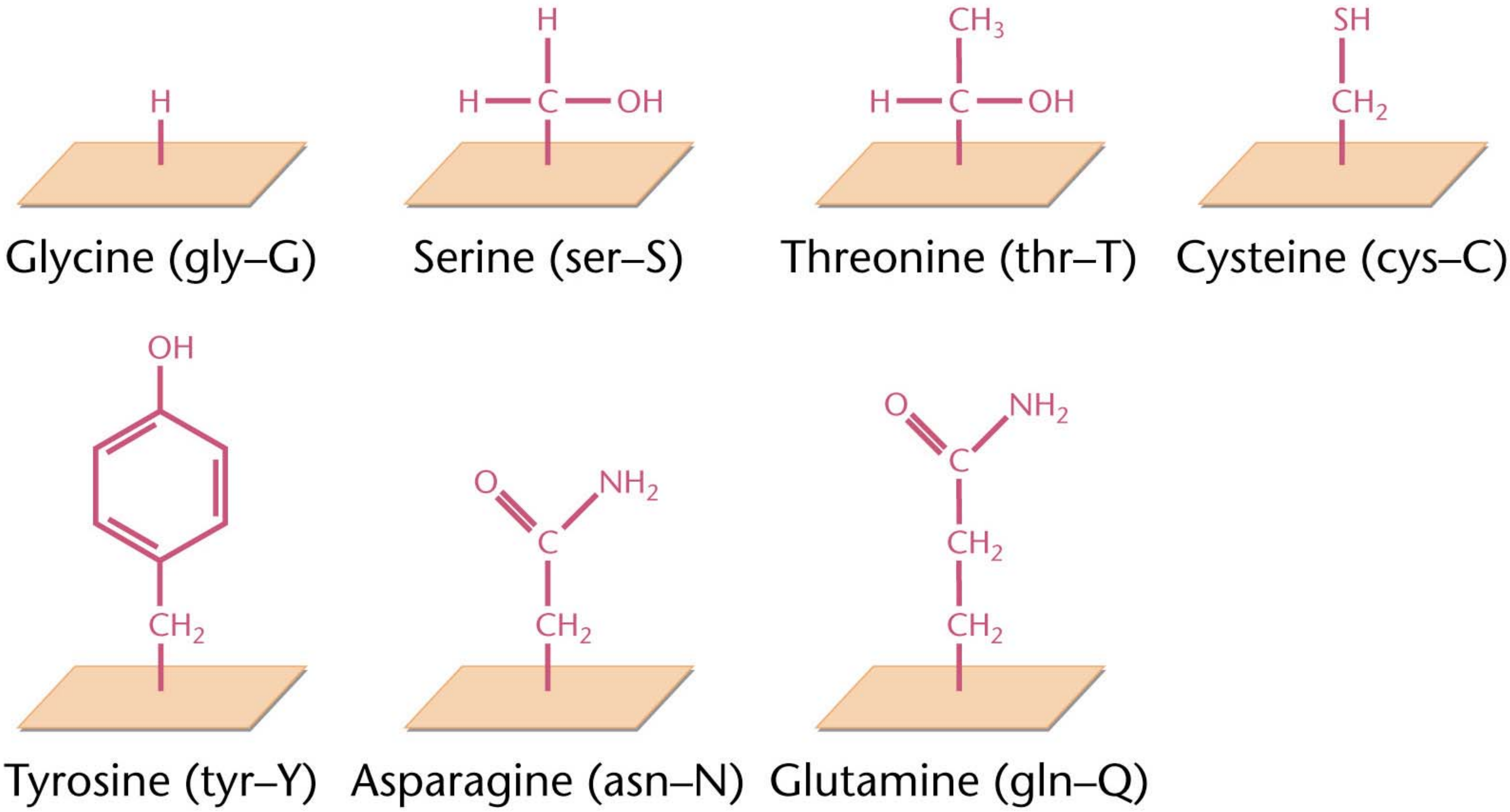
A polypeptide chain. The R_i *side chains identify the component amino acids. Atoms inside each quadrilateral are on the same plane, which can rotate according to angles* ϕ *and* ψ.


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



Amino Acid Structures from Klug & Cummings

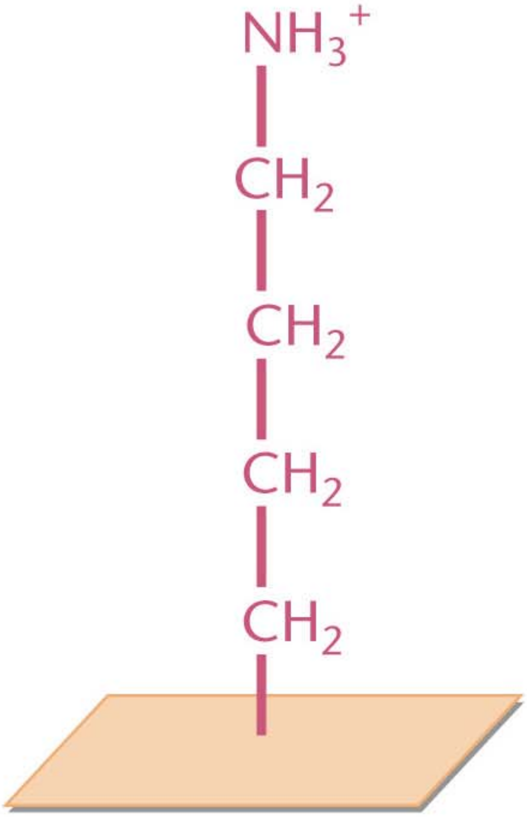
2. Polar: Hydrophilic



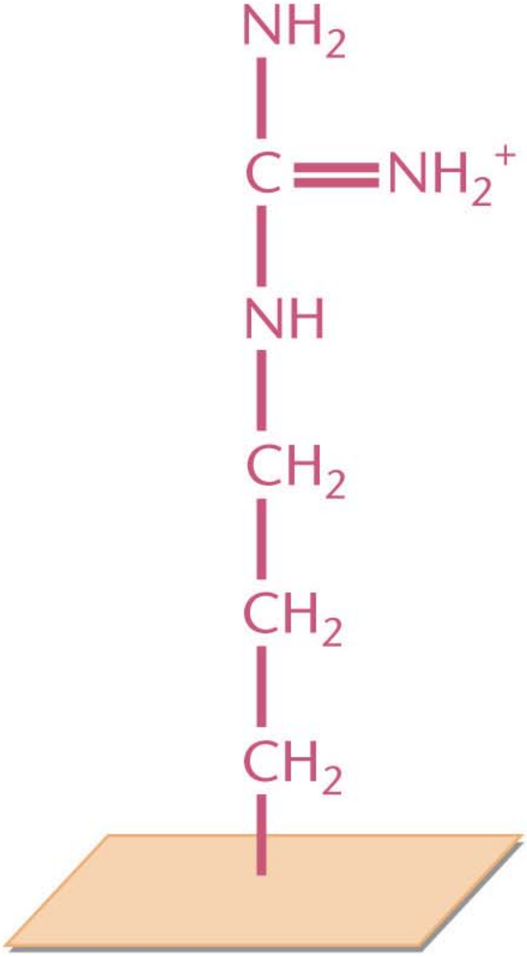
Amino Acid Structures from Klug & Cummings

3. Polar: positively charged (basic)

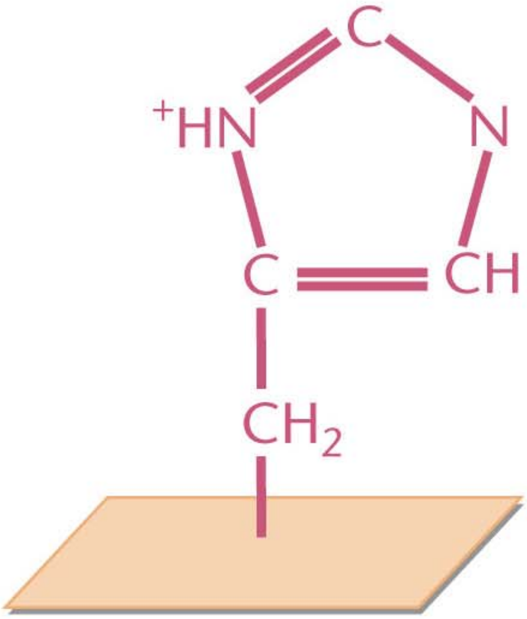
Amino Acid Structures
from Klug & Cummings



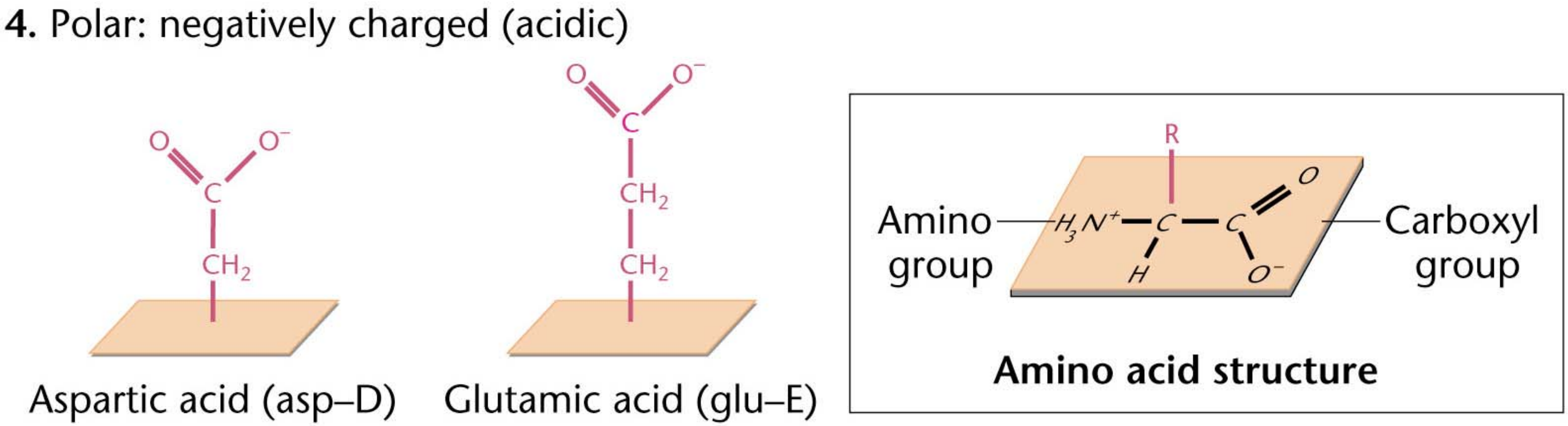
Lysine (lys–K)



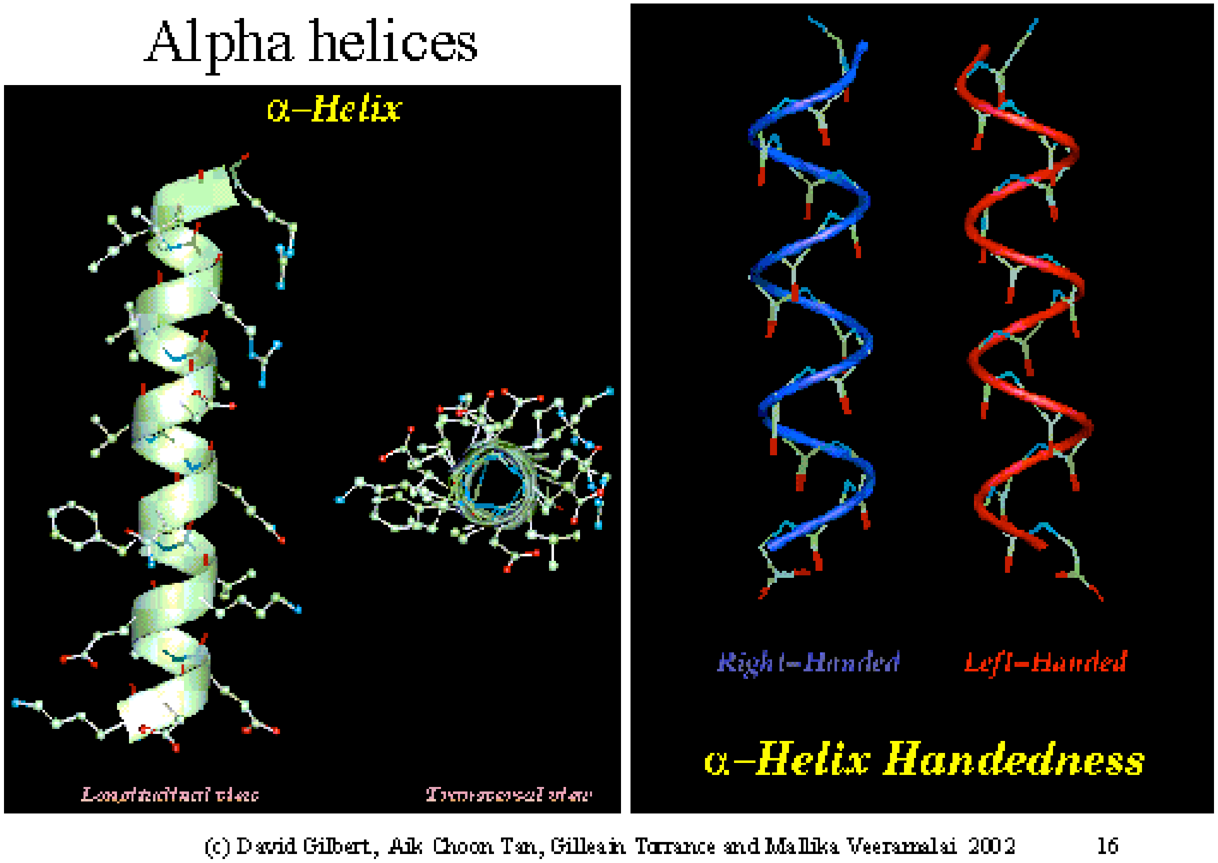
Arginine (arg–R)



Histidine (his–H)



Amino Acid Structures from Klug & Cummings



Alpha helices

(c) David Gilbert, Aik Choon Tan, Gilleain Torrance and Mallika Veeramalai 2002 16

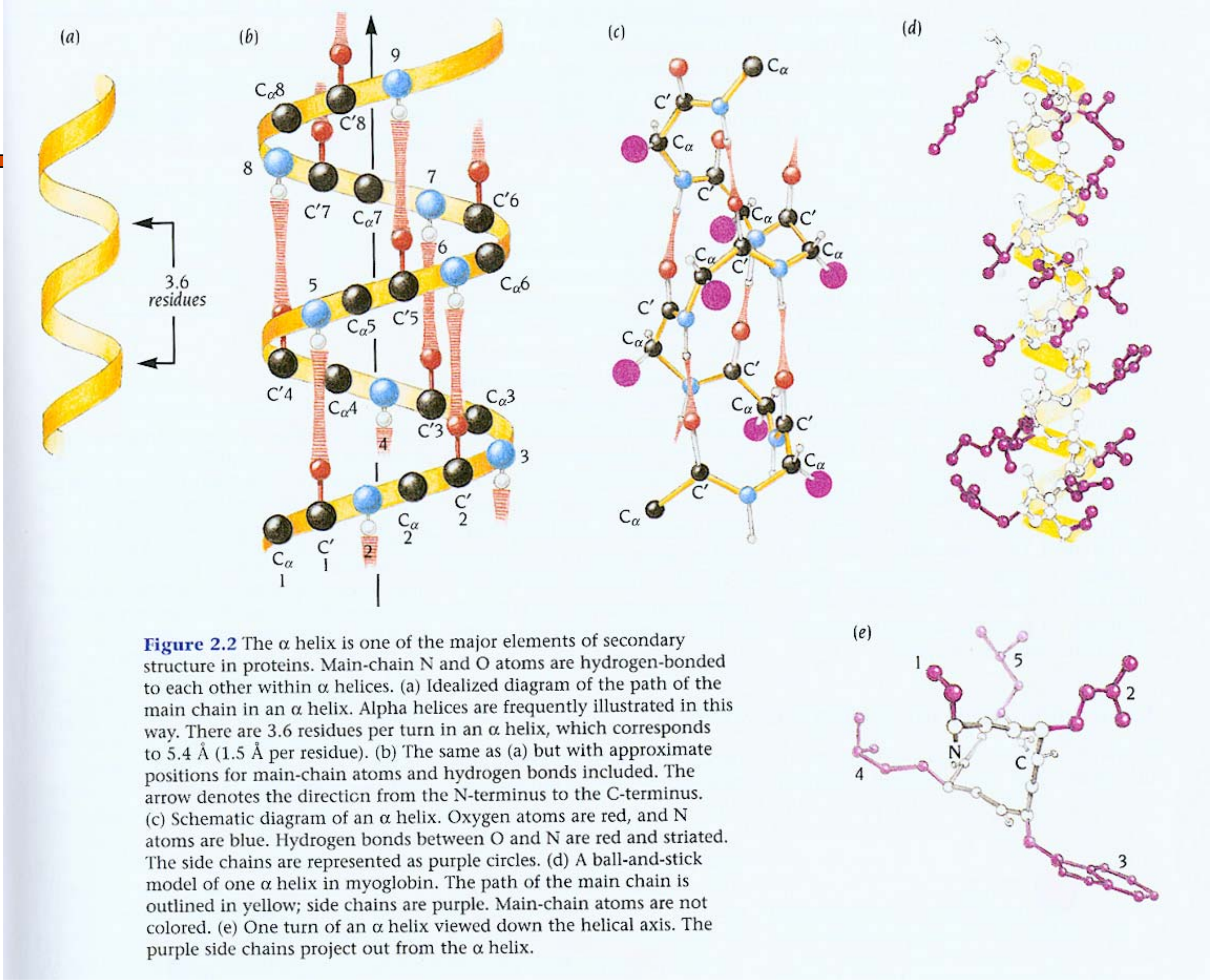
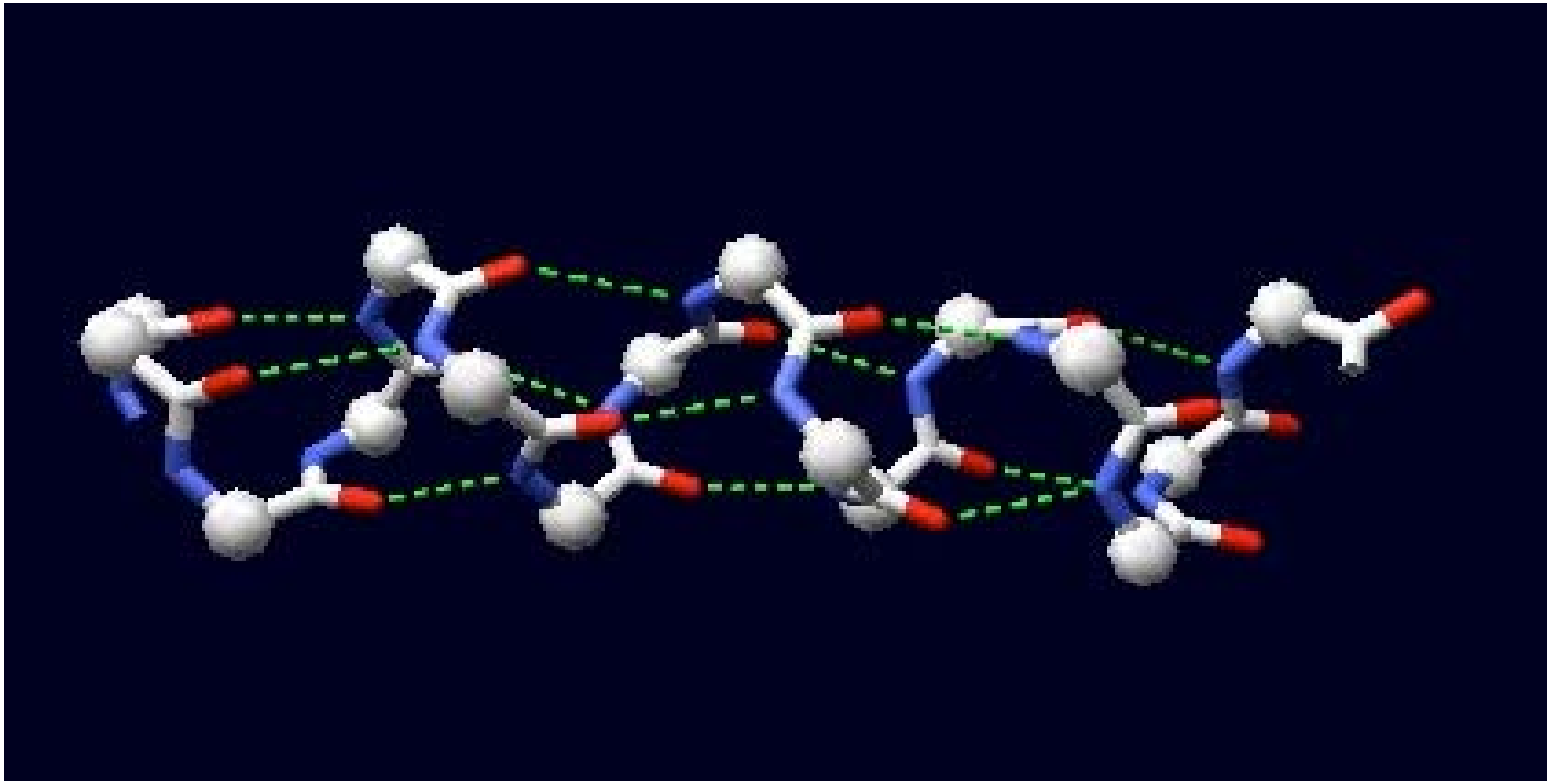
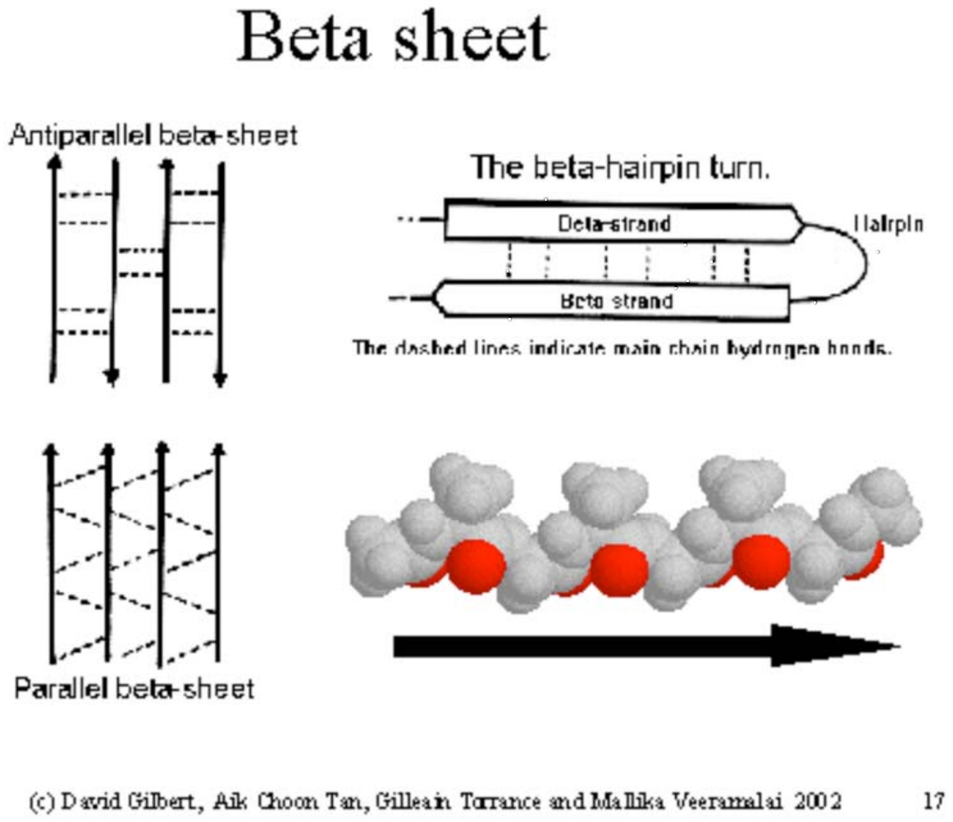


Figure 2.2 The α helix is one of the major elements of secondary structure in proteins. Main-chain N and O atoms are hydrogen-bonded to each other within α helices. (a) Idealized diagram of the path of the main chain in an α helix. Alpha helices are frequently illustrated in this way. There are 3.6 residues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



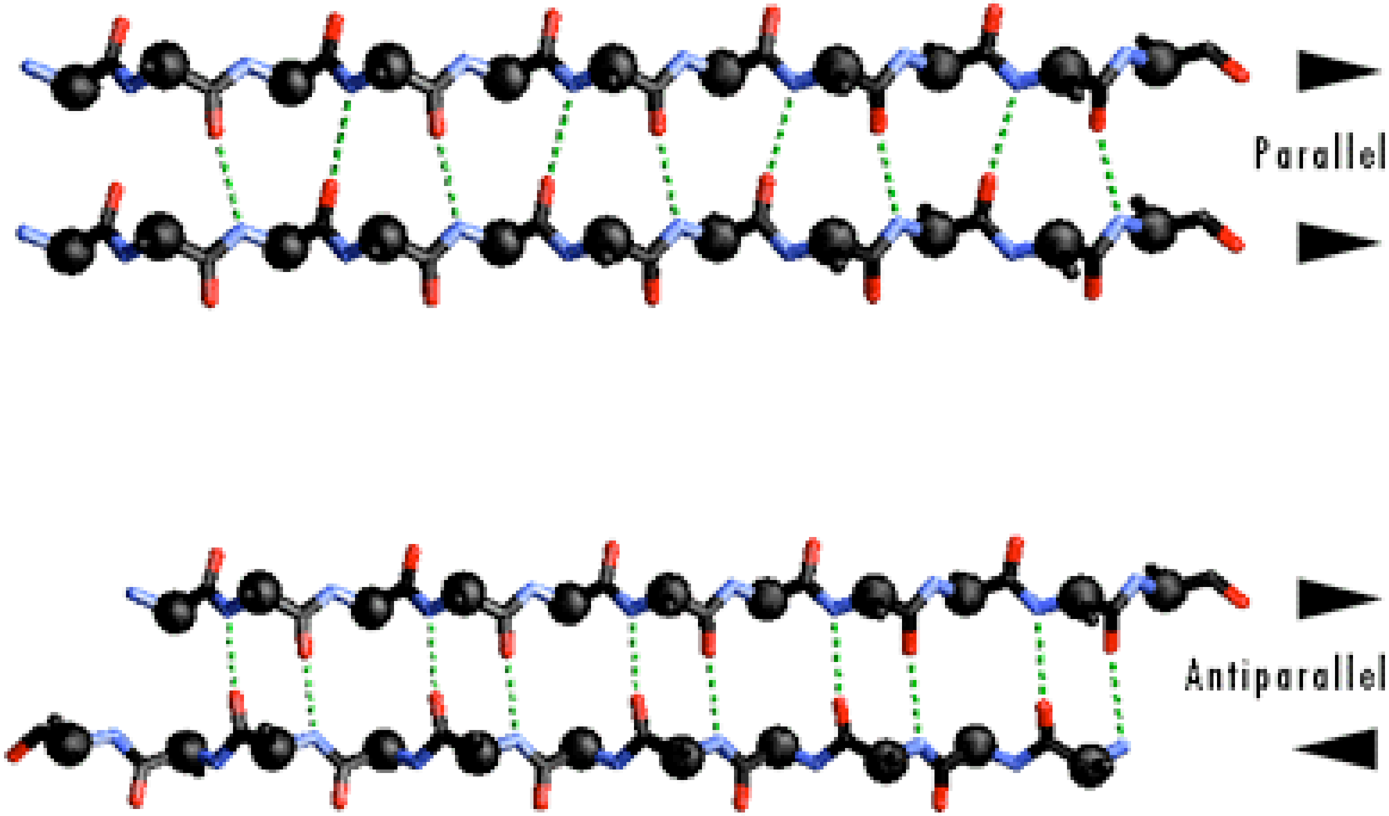


Beta sheet

The dashed lines indicate main chain hydrogen bonds.

(c) David Gilbert, Aik Choon Tan, Gilleain Torrance and Mallika Veeramalai 2002 17

Beta Strand



Active Sites

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

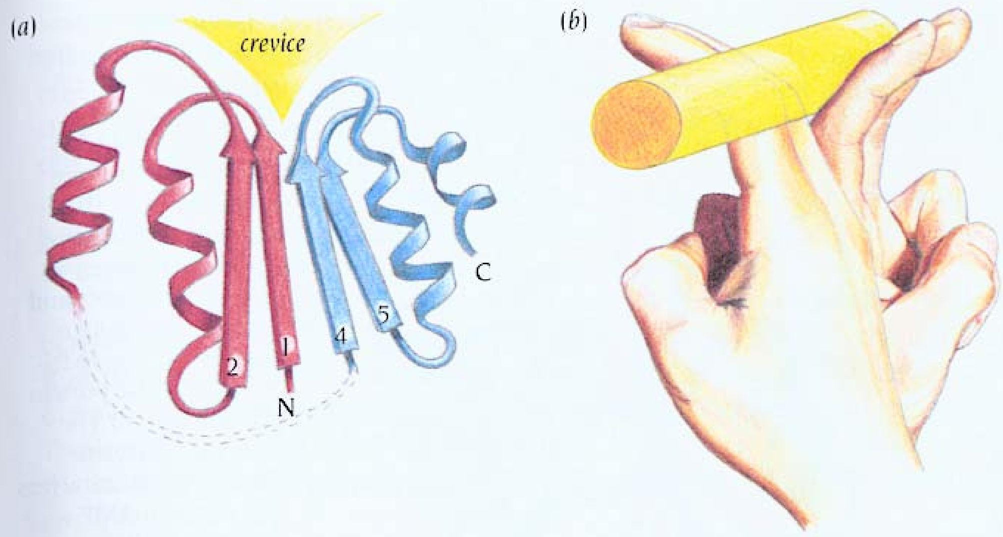
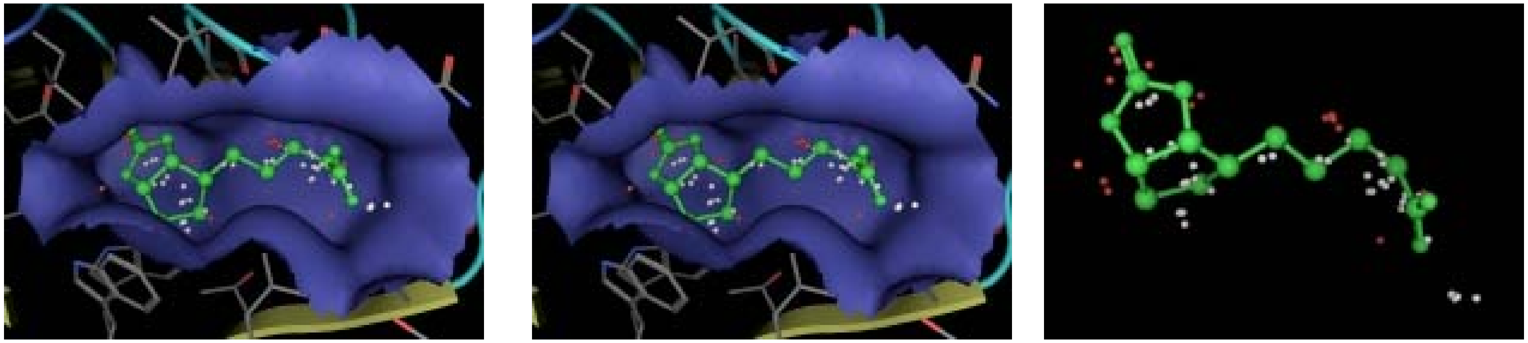


Figure 4.13 (a) The active site in open twisted α/β domains is in a crevice outside the carboxy ends of the β strands. This crevice is formed by two adjacent loop regions that connect the two strands with α helices on opposite sides of the β sheet. This is illustrated by the curled fingers of two hands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

Secondary Structure Prediction Software

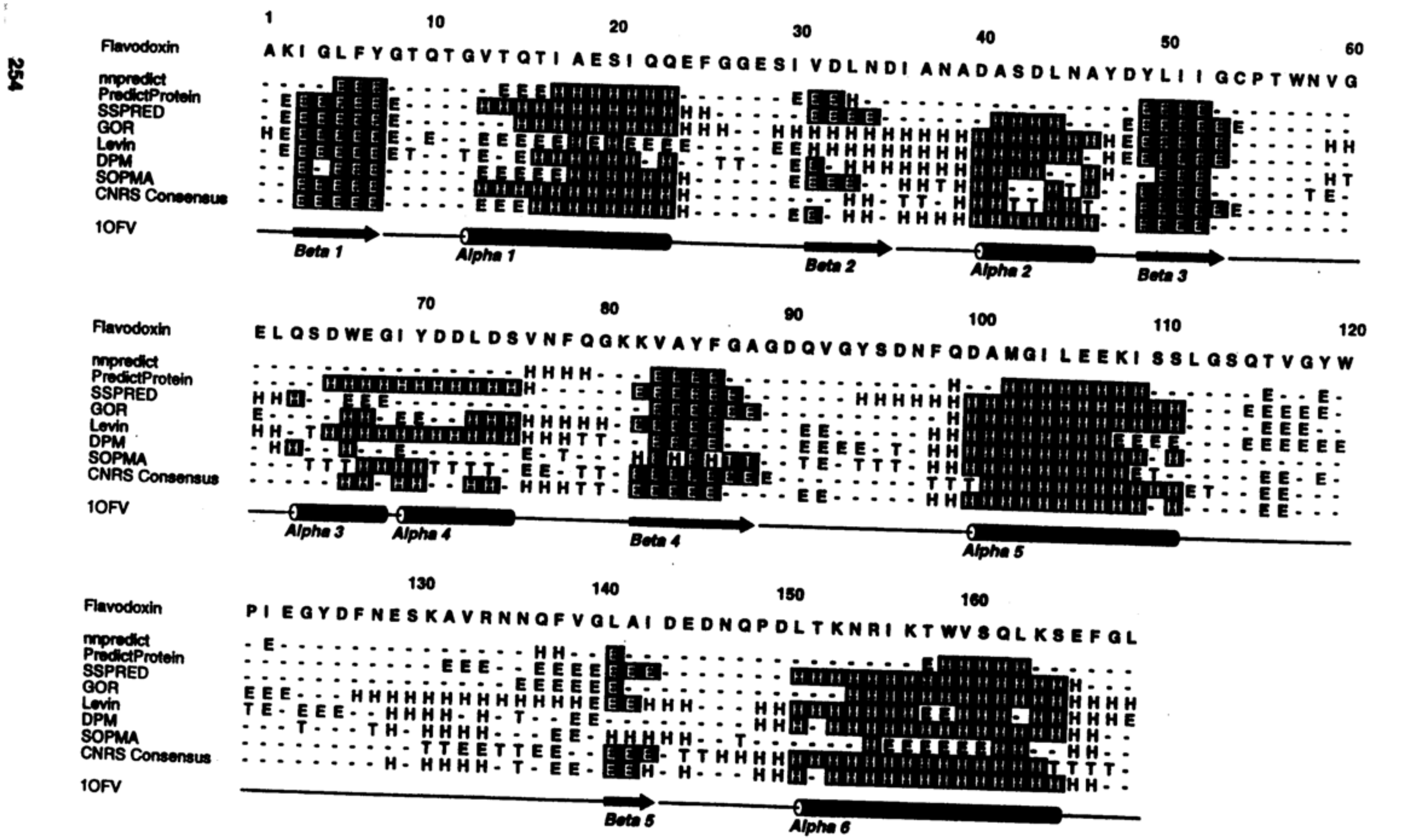
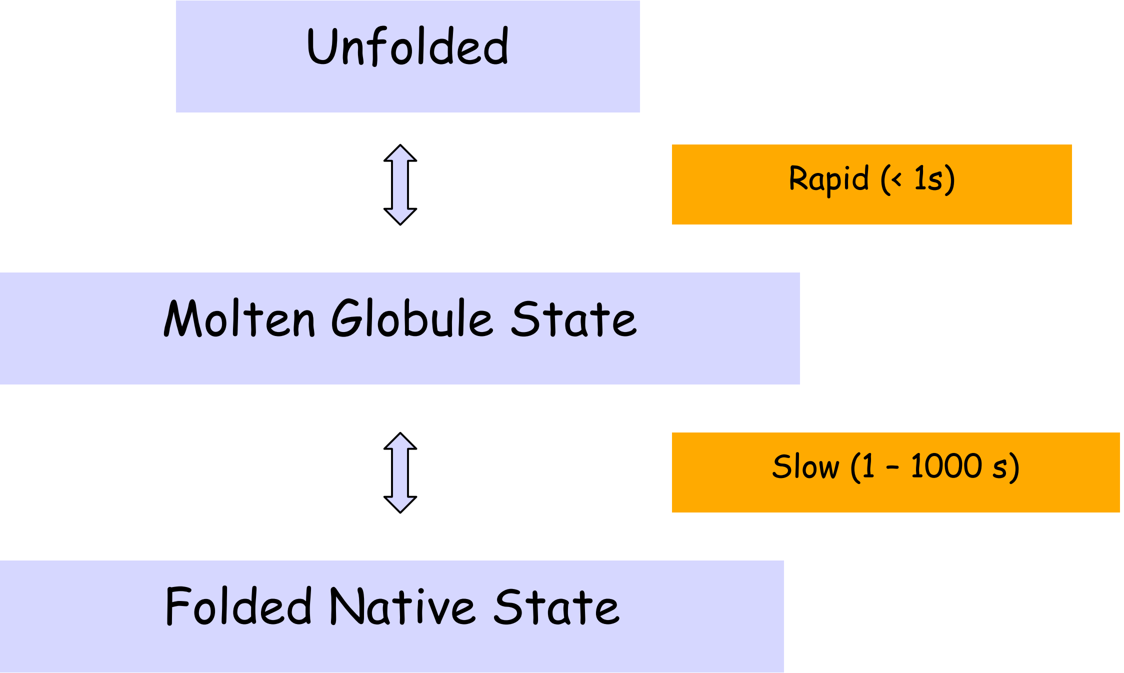


Figure 11.3 Comparison of secondary structure predictions by various methods. The sequence of flavodoxin, an α/β protein, was used as the query and is shown on the first line of the alignment. For each prediction, H denotes an α helix, E a β strand, T a β turn; all other positions are assumed to be random 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 [ ]

Protein Folding



Unfolded

Rapid (< 1s)

Molten Globule State

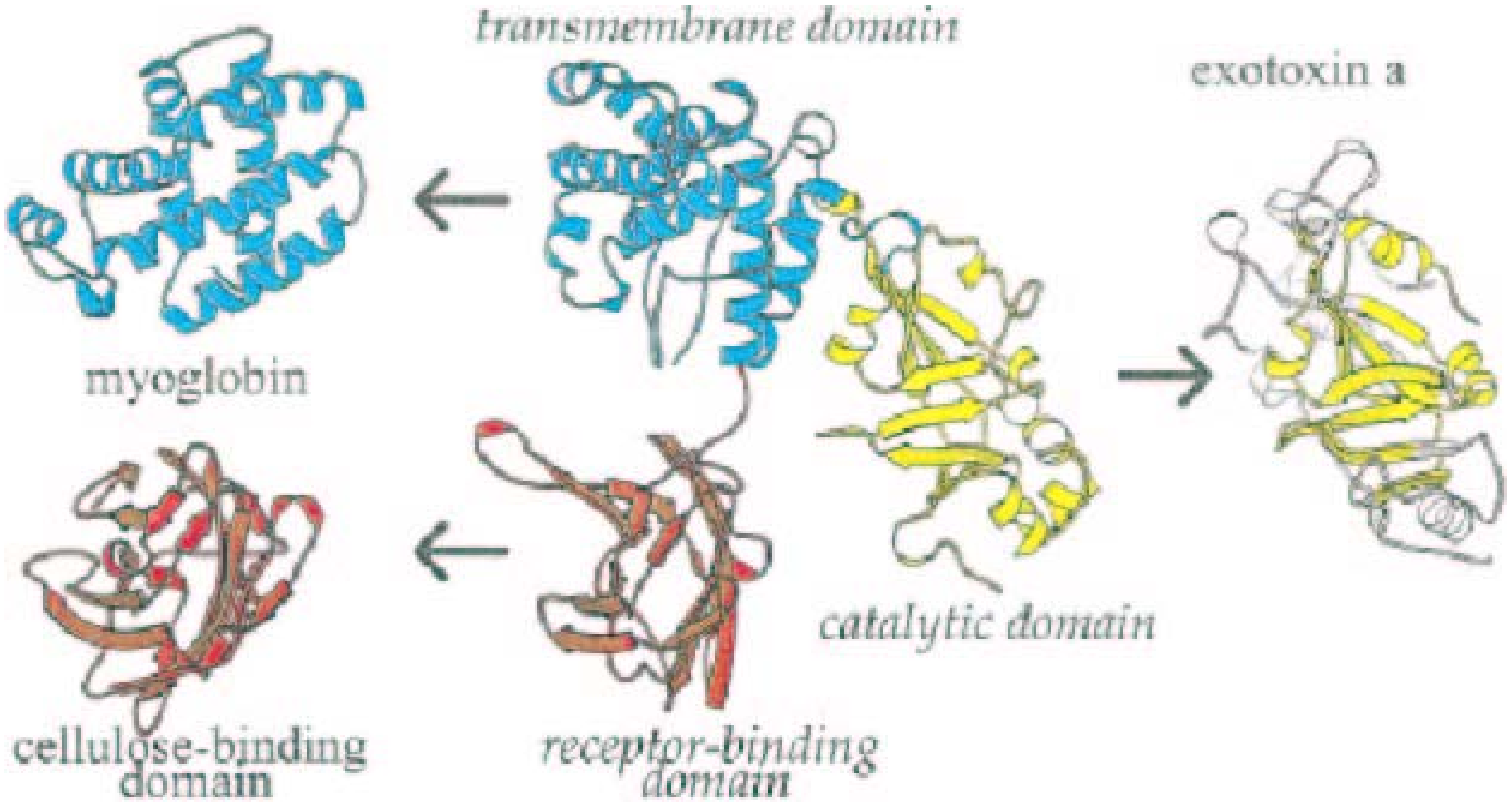
Slow (1 – 1000 s)

Folded Native State

- How to find minimum energy configuration?

Modular Nature of Protein Structures

Example: Diphtheria Toxin



Protein Structures

- Most proteins have a hydrophobic core.
- Within the core, specific interactions take place between amino acid side chains.
- Can an amino acid be replaced by some other amino acid?
  - Limited by space and available contacts with nearby amino acids
- Outside the core, proteins are composed of loops and structural elements in contact with water, solvent, other proteins and other structures.

Viewing Protein Structures

- SPDBV
- RASMOL
- CHIME