CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

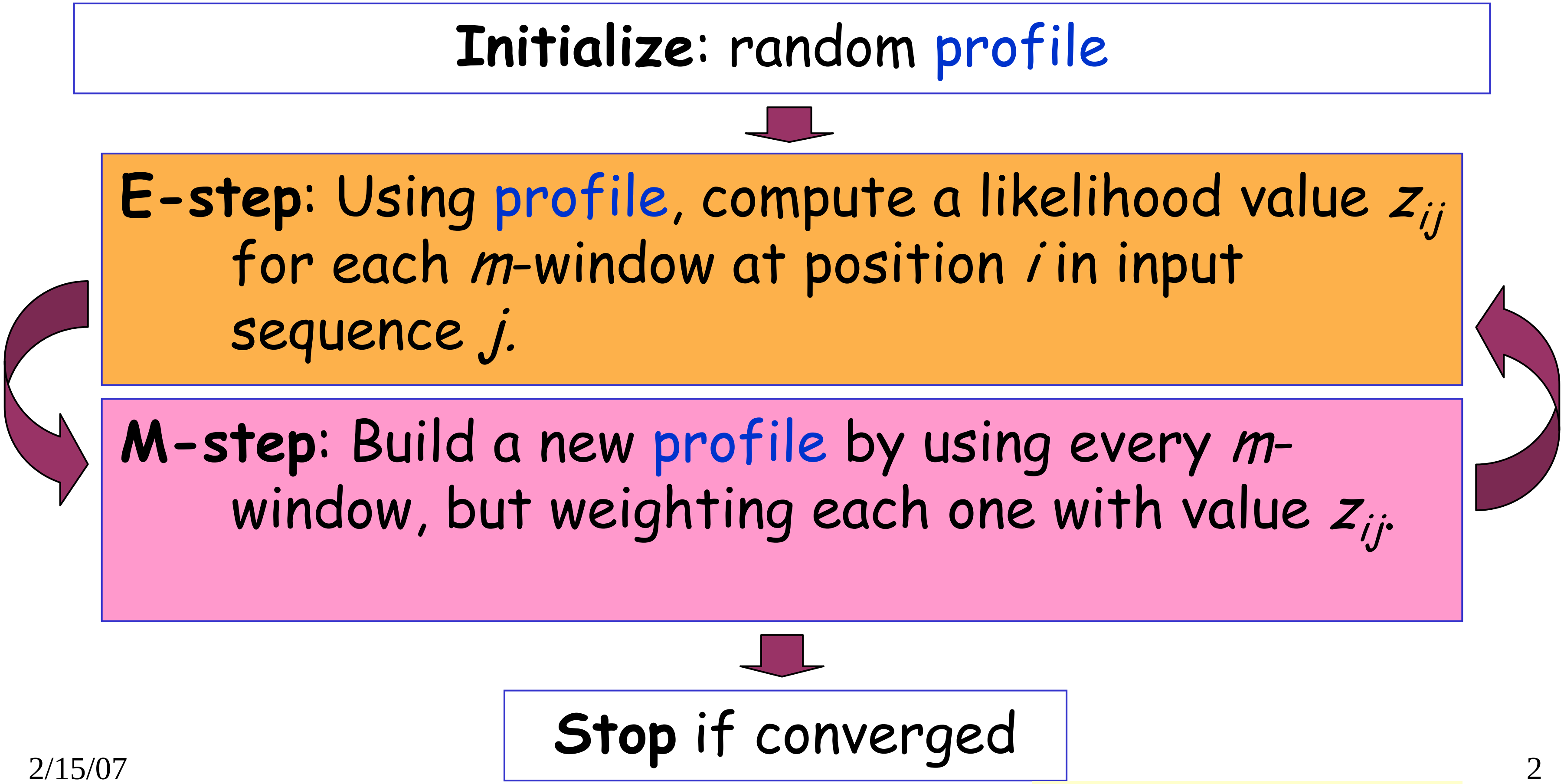
ECS 254; Phone: x3748

giri@cis.fiu.edu

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

EM Algorithm

Goal: Find θ, Z that maximize Pr (X, Z | θ)



Initialize: random profile

E-step: Using profile, compute a likelihood value z_{ij} for each *m*-window at position *i* in input sequence *j*.

M-step: Build a new profile by using every *m*-window, but weighting each one with value z_{ij}.

Stop if converged

MEME [Bailey, Elkan 1994]

EM Method: Model Parameters

- Input: upstream sequences
  - $X = \{X_1, X_2, \ldots, X_n\}$,
- Motif profile: 4×k matrix $\theta = (\theta_{rp})$,
  - $r \in \{A,C,G,T\}$
  - $1 \leq p \leq k$
  - θ_{rp} = Pr(residue r in position p of motif)
- Background distribution:
  - θ_{r0} = Pr(residue r in background)

EM Method: Hidden Information

- $Z = \{Z_{ij}\}$, where

$$Z_{ij} = \begin{cases} 1, & \text{if motif instance starts at position } i \text{ of } X_j \\ 0, & \text{otherwise} \end{cases}$$

- Iterate over probabilistic models that could generate X and Z, trying to converge on this solution, i.e., maximize $\Pr(X, Z \mid \theta)$.

Statistical Evaluation

- ❑ Z-score of a motif with a certain frequency: →

$$z(w) = \frac{Obs(w) - Exp(w)}{\sqrt{Var(w)}}$$

- ❑ Information Content or Relative Entropy of an alignment or profile: →

$$IC(M) = \sum_{i=1}^{4}\sum_{j=1}^{m} m_{i,j} \log \frac{m_{i,j}}{b_i}$$

- ❑ Maximum a Posteriori (MAP) Score: →

$$MAP(M) = -\sum_{i=1}^{4}\sum_{j=1}^{m} n_{i,j} \log \frac{m_{i,j}}{b_i}$$

- ❑ Model Vs Background Score: →

$$L(w) = \frac{\Pr(w \mid M)}{\Pr(w \mid Bg)} = \prod_{j=1}^{m} \frac{m_{i,j}}{b_i}$$

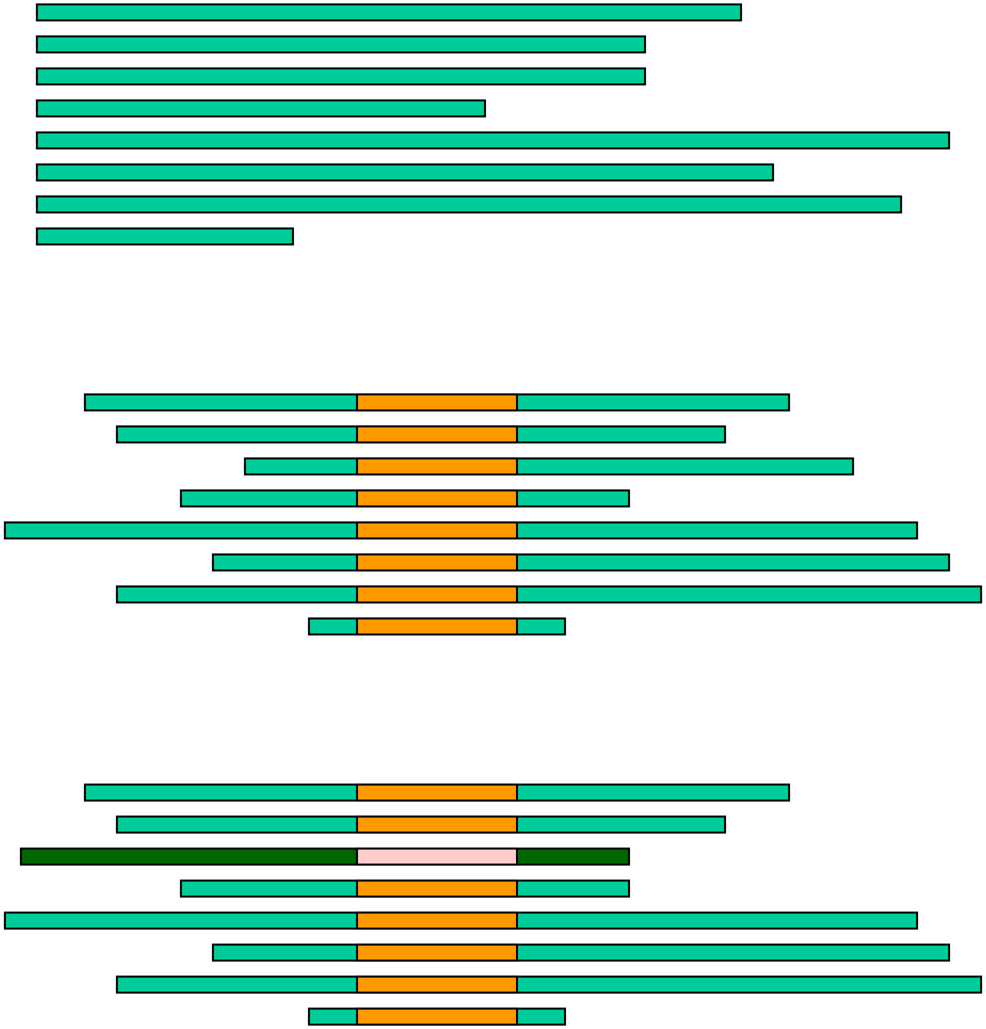
Counts

Frequencies

Predicting Motifs in Whole Genome

- **MEME: EM algorithm** [Bailey *et al.*, 1994]
- **AlignACE: Gibbs Sampling Approach** [Hughes *et al.*, 2000]
- **Consensus: Greedy Algorithm Based** [Hertz *et al.*, 1990]
- **ANN-Spec: Artificial Neural Network and a Gibbs sampling method** [Workman *et al.*, 2000]
- **YMF: Enumerative search** [Sinha *et al.*, 2003]
- **…**

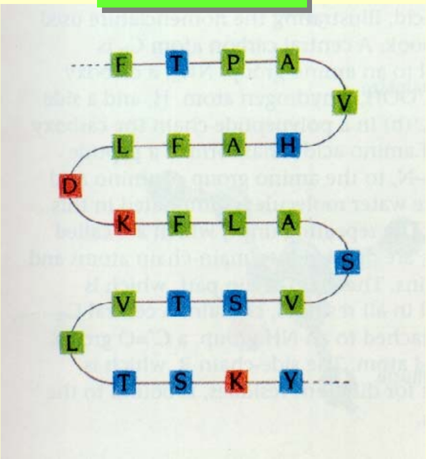
Gibbs Sampling for Motif Detection



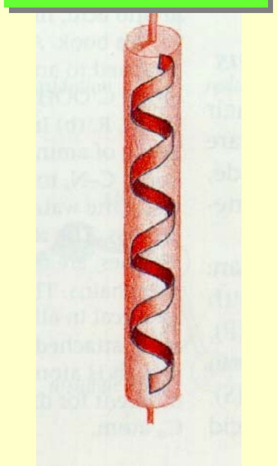
Protein Structures

- ❑ Sequences of amino acid residues
- ❑ 20 different amino acids

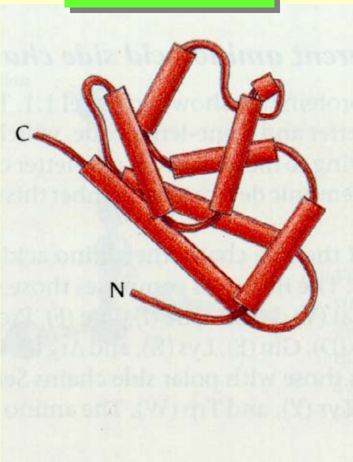
Primary



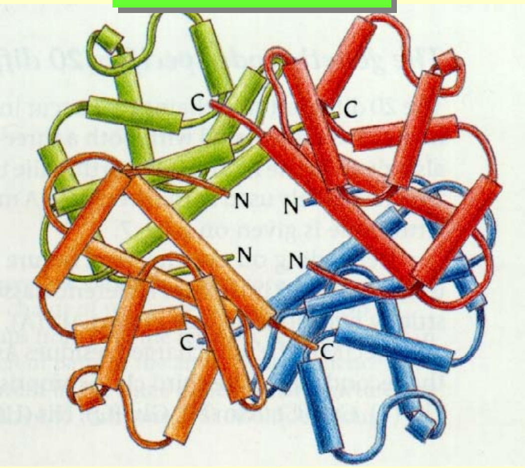
Secondary



Tertiary



Quaternary



Proteins

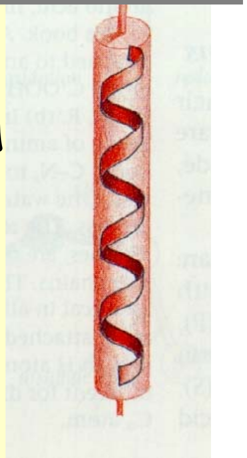
- **Primary structure** is the sequence of amino acid residues of the protein, e.g., Flavodoxin:

  AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADA…

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

  AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADA…

Secondary



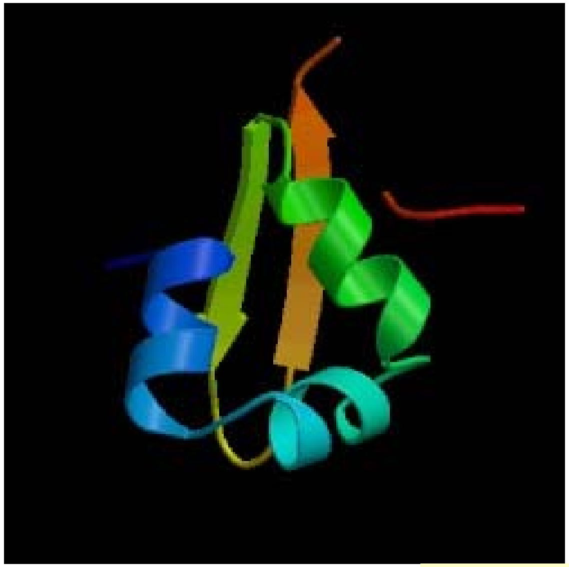
More on Secondary Structures

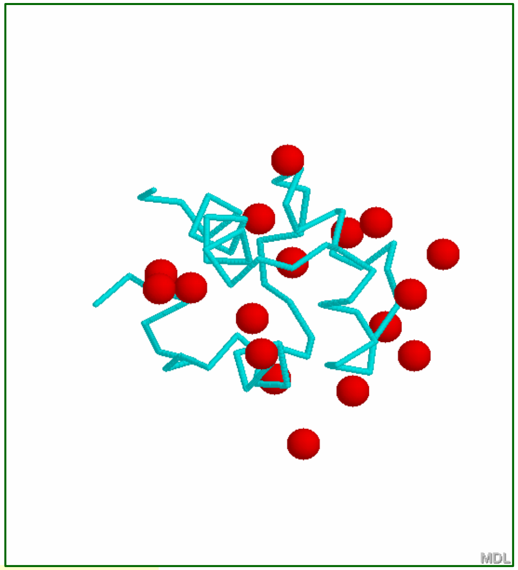
- **α-helix**
  - Main chain with peptide bonds
  - Side chains project outward from helix
  - Stability provided by H-bonds between CO and NH groups of residues 4 locations away.
- **β-strand**
  - Stability provided by H-bonds with one or more β-strands, forming β-sheets. Needs a β-turn.

Proteins

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

Myoglobin



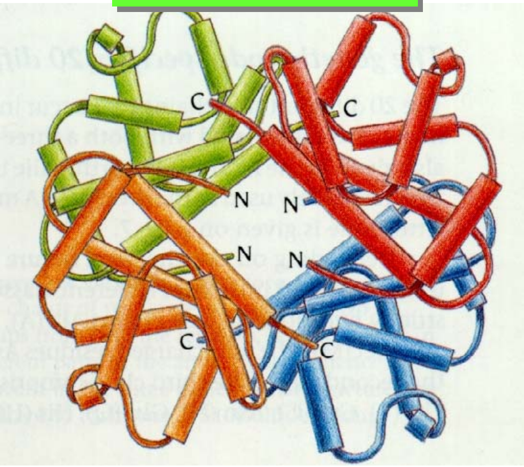


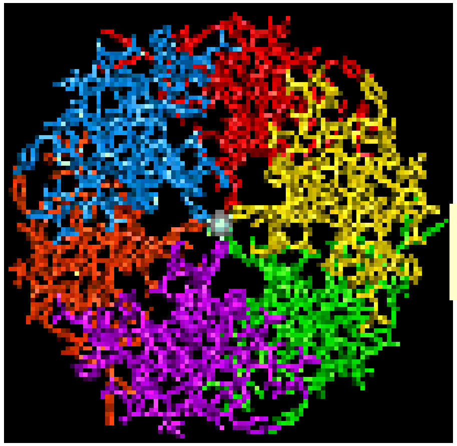
Lambda Cro

Quaternary Structures in Proteins

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

Quaternary





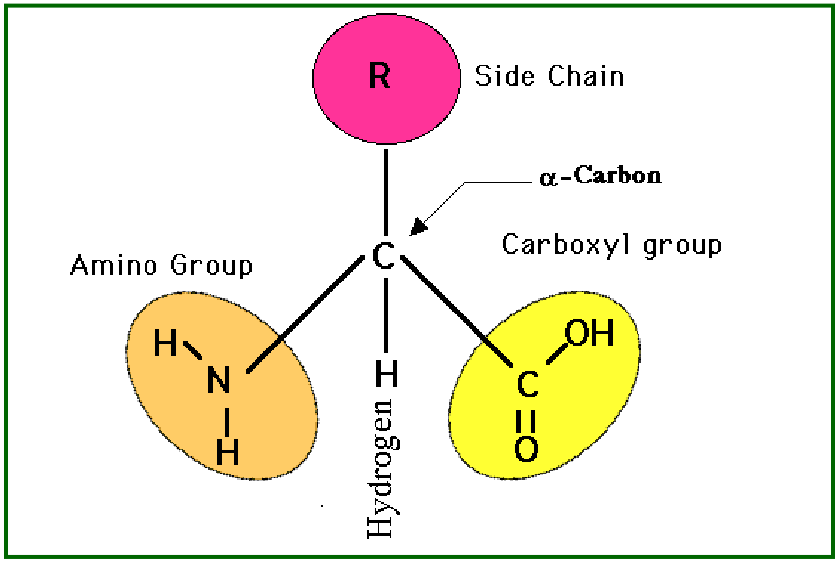
Insulin Hexamer

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.



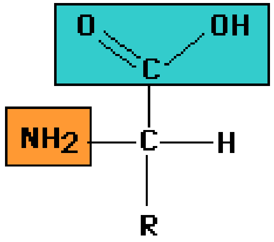
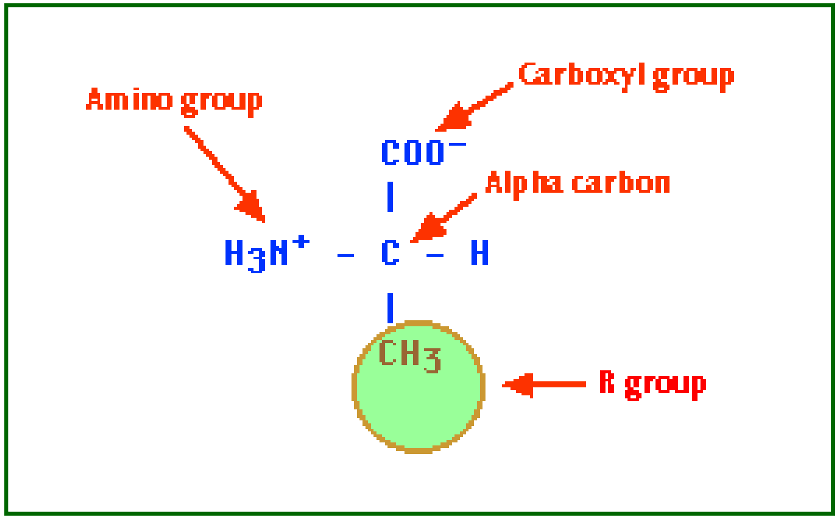
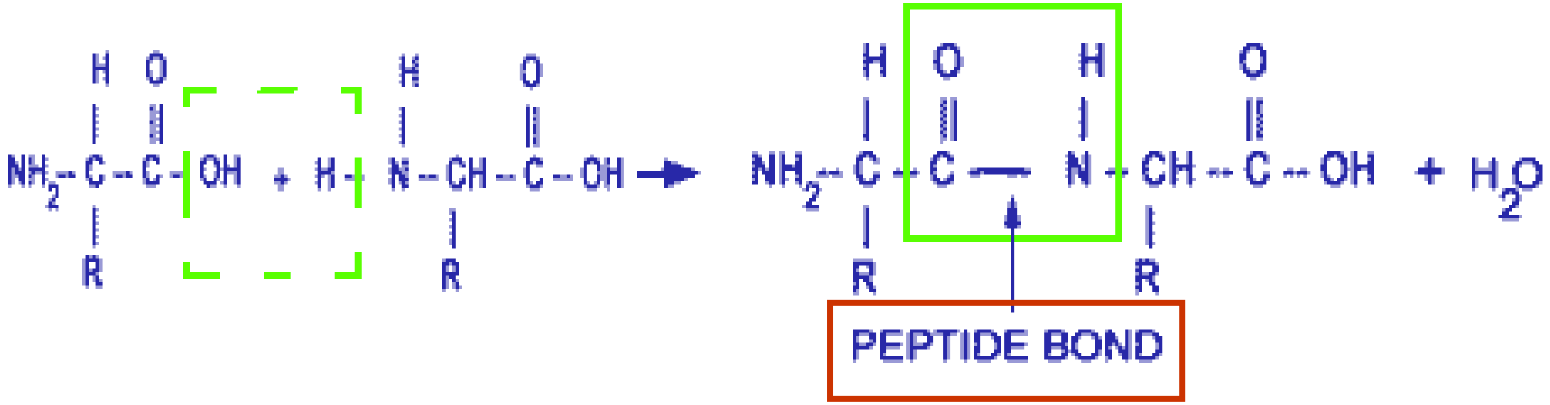


Fig. General formula for an amino acid molecule. "R" represents the variable groups that are attached to this basic molecule to make up the 20 common amino acids



Chains of amino acids



Amino acids vs Amino acid residues

Angles φ and ψ in the polypeptide chain

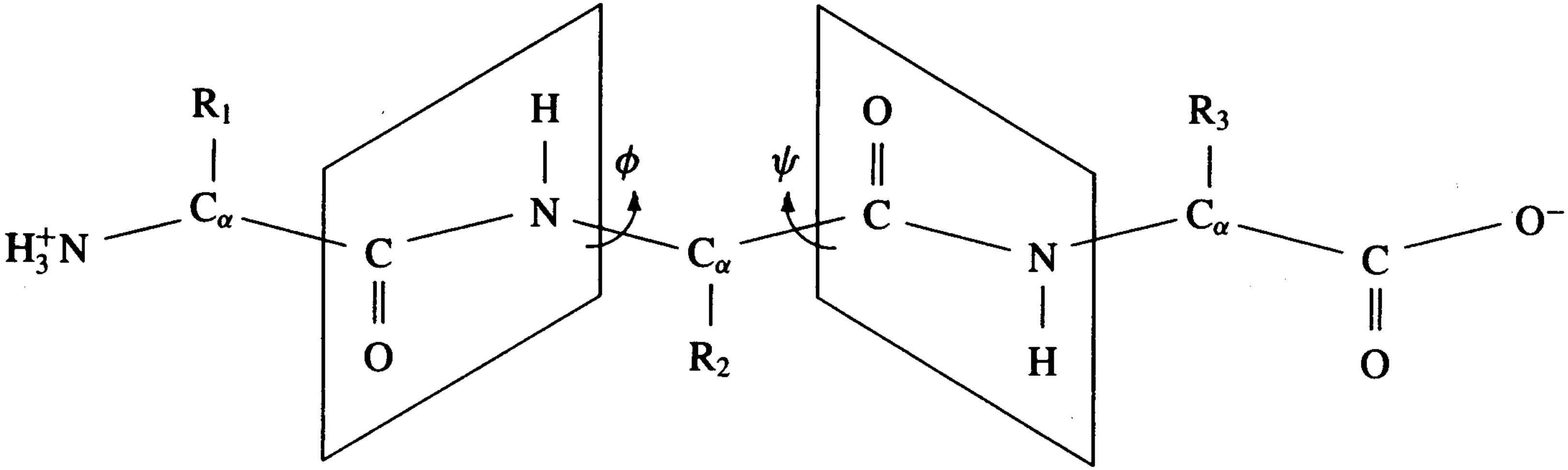
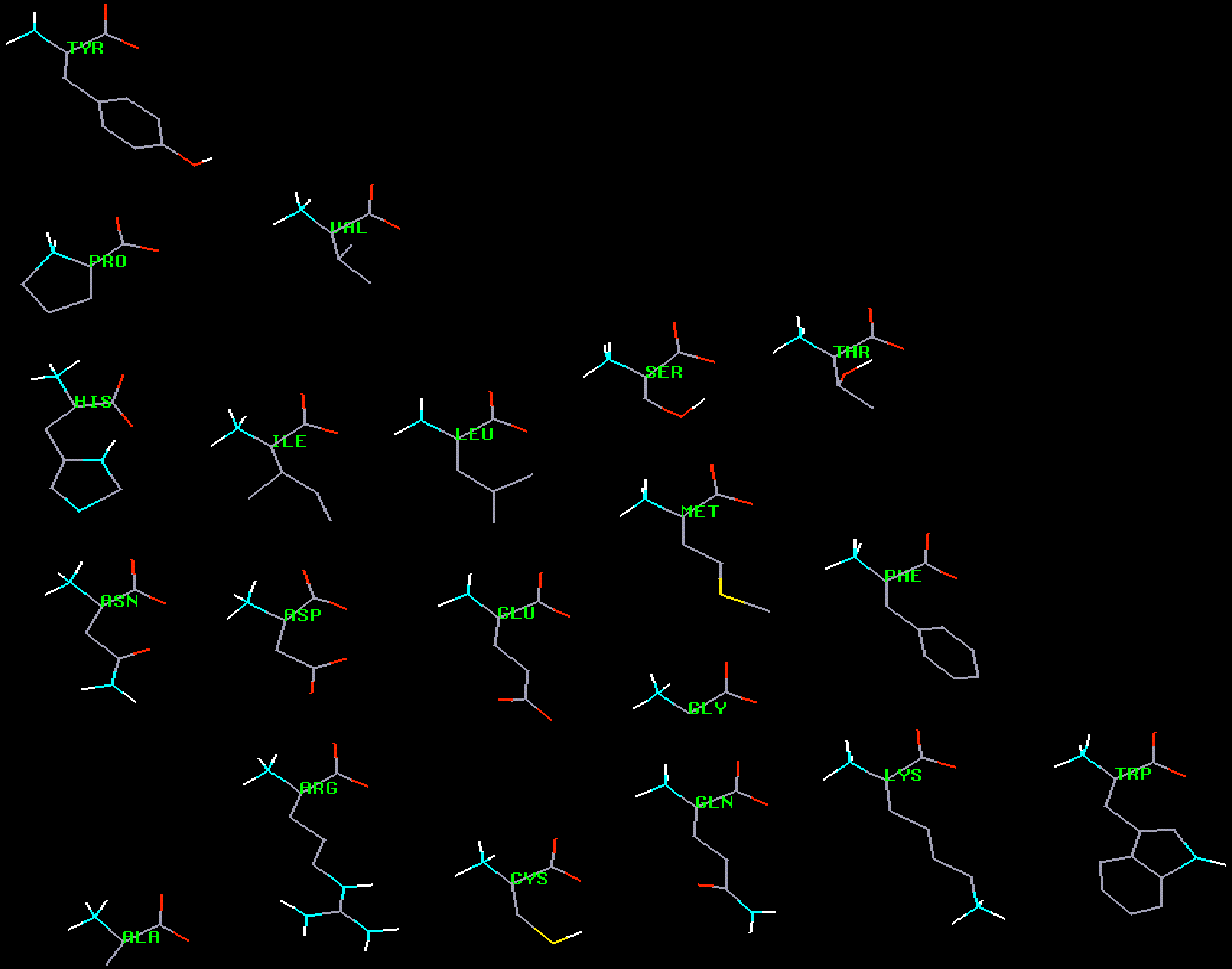
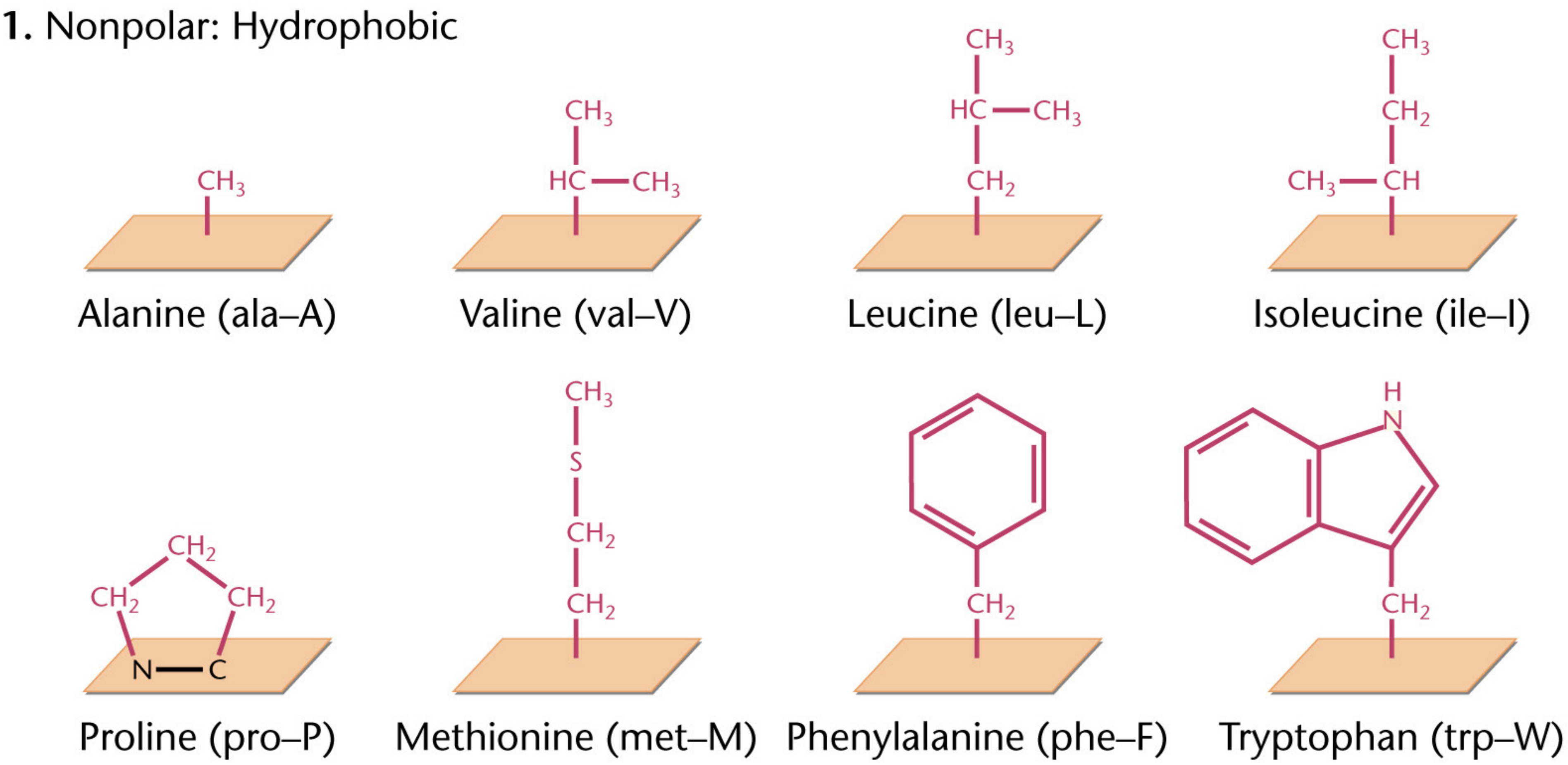


FIGURE 1.2

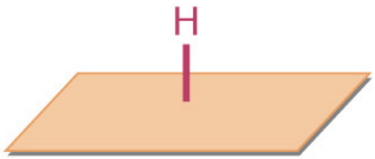
A polypeptide chain. The R_i *side chains identify the component amino acids. Atoms inside each quadrilateral are on the same plane, which can rotate according to angles* ϕ *and* ψ.


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

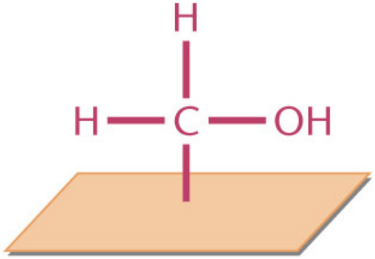


Amino Acid Structures from Klug & Cummings

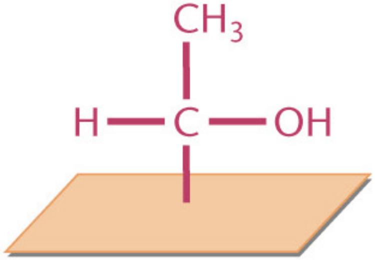
2. Polar: Hydrophilic



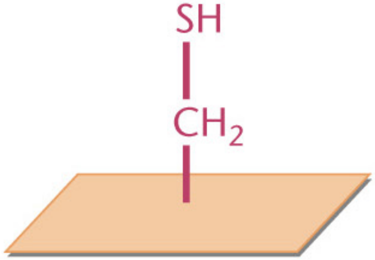
Glycine (gly–G)



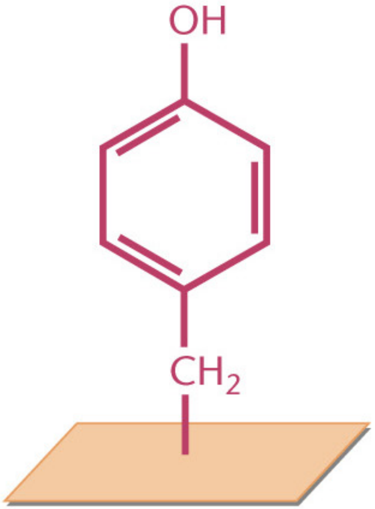
Serine (ser–S)



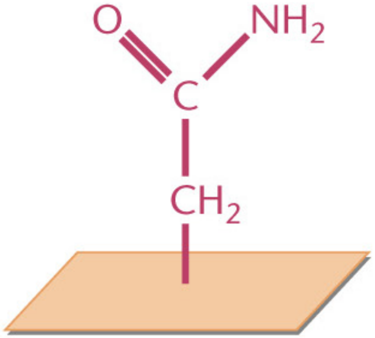
Threonine (thr–T)



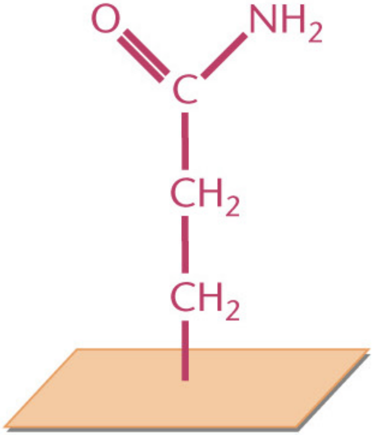
Cysteine (cys–C)



Tyrosine (tyr–Y)



Asparagine (asn–N)

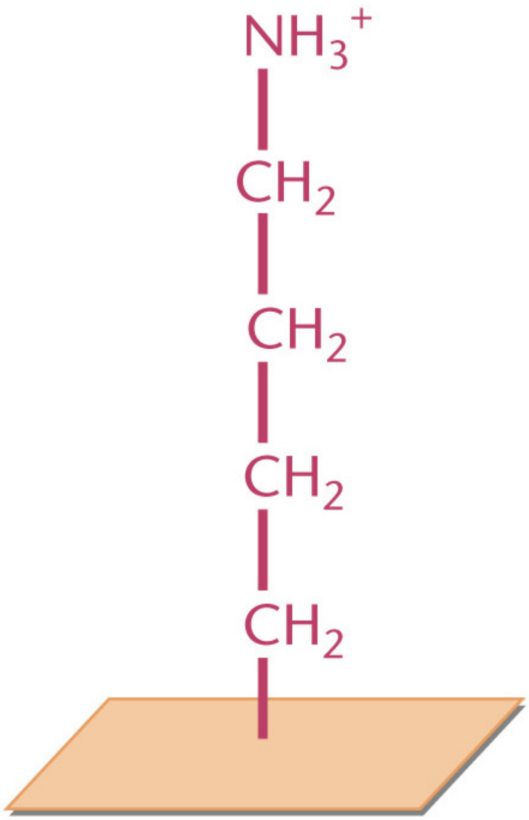


Glutamine (gln–Q)

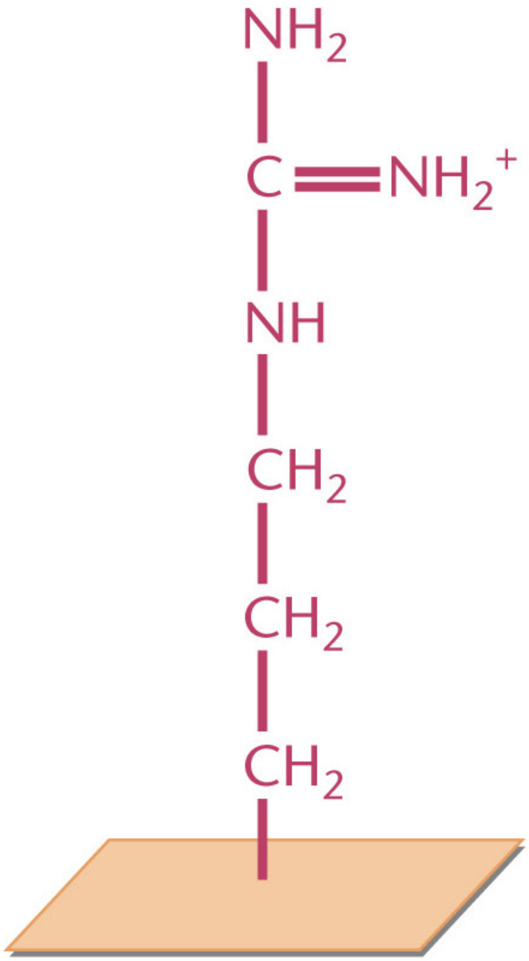
Amino Acid Structures from Klug & Cummings

3. Polar: positively charged (basic)

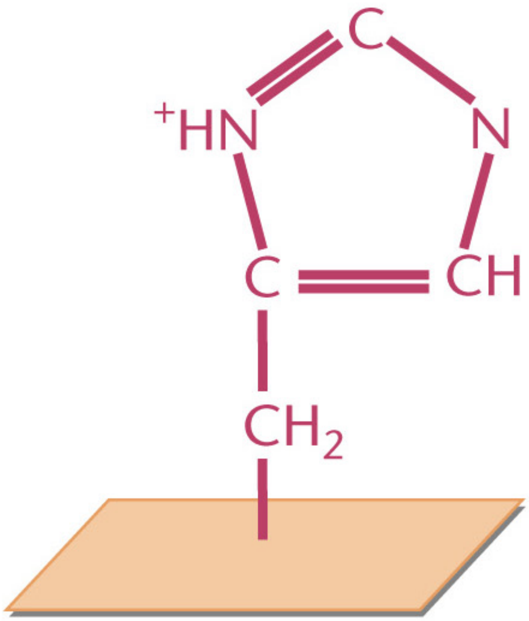
Amino Acid Structures from Klug & Cummings



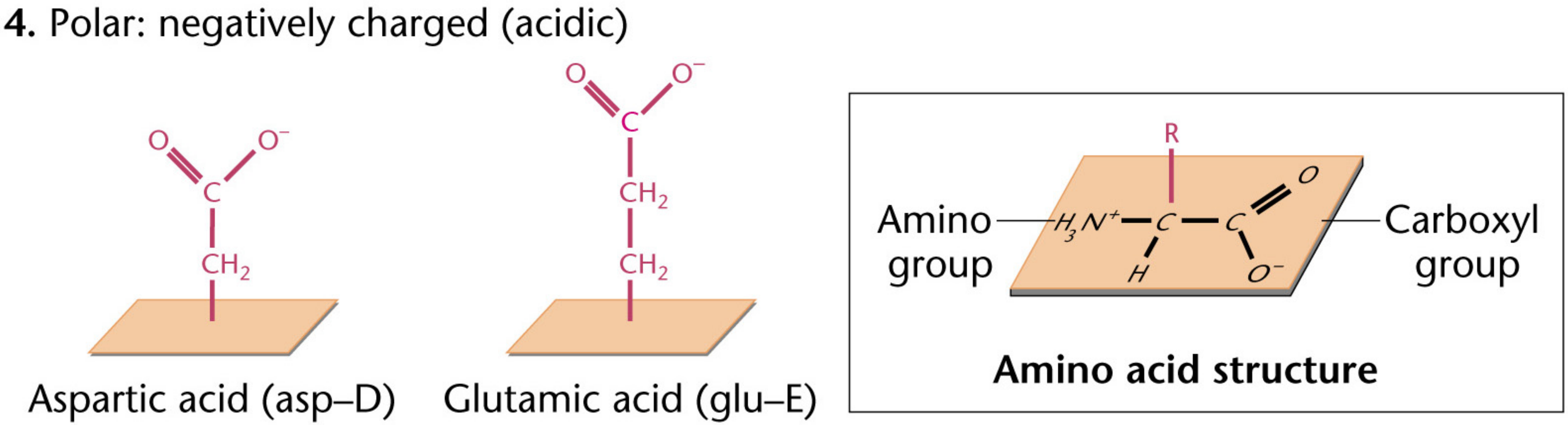
Lysine (lys–K)



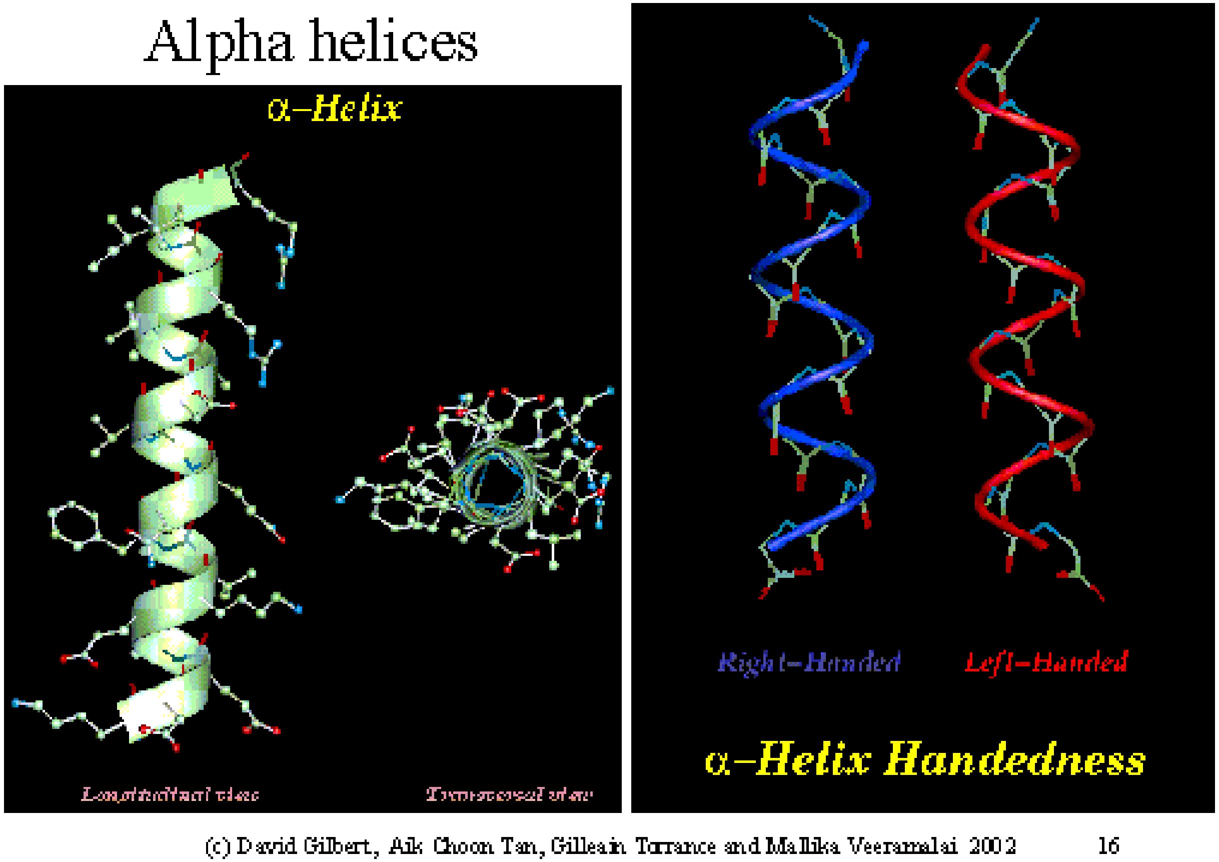
Arginine (arg–R)



Histidine (his–H)

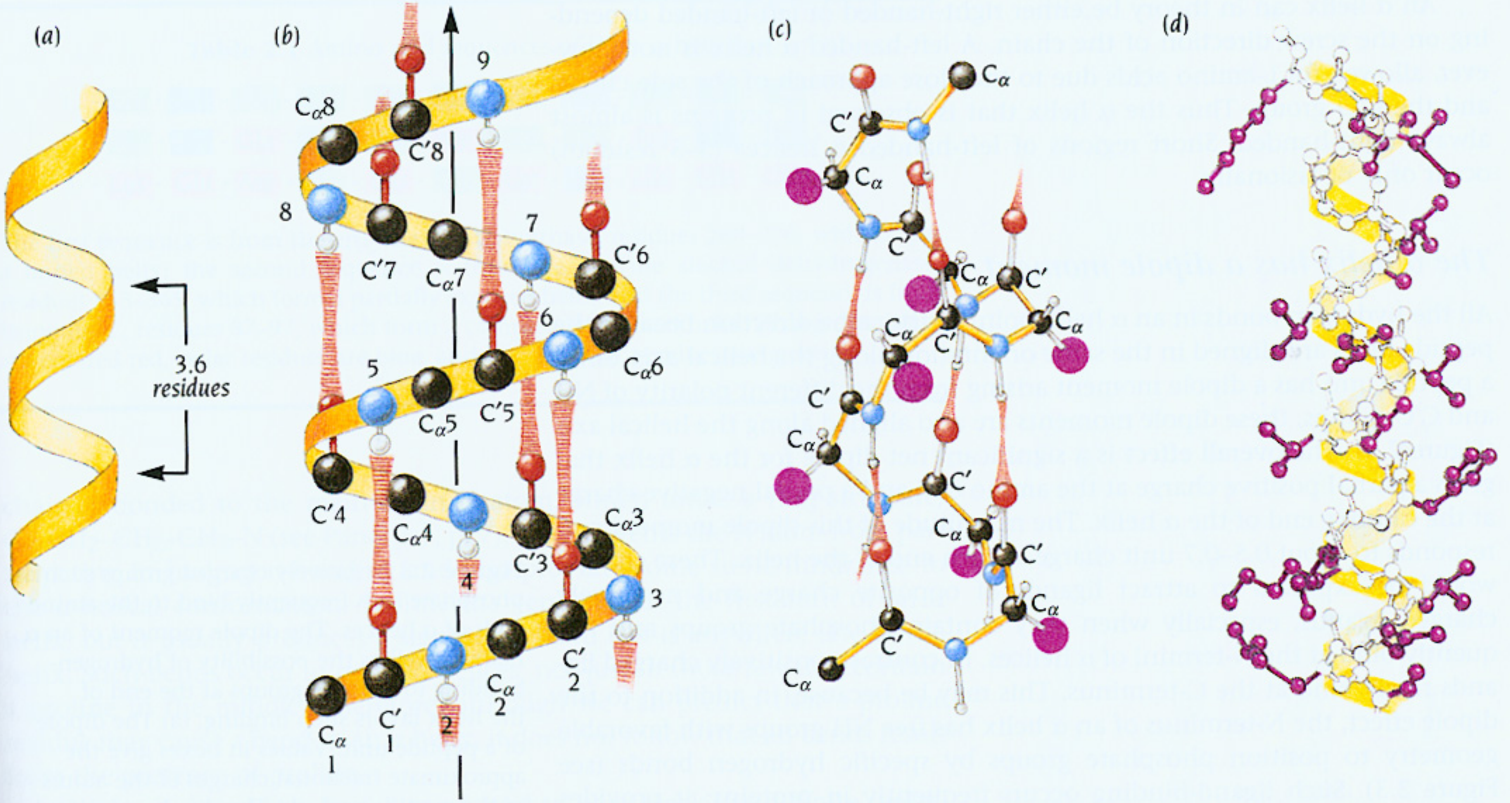


Amino Acid Structures from Klug & Cummings



Alpha helices

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



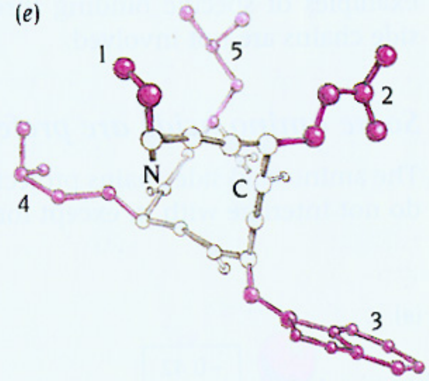
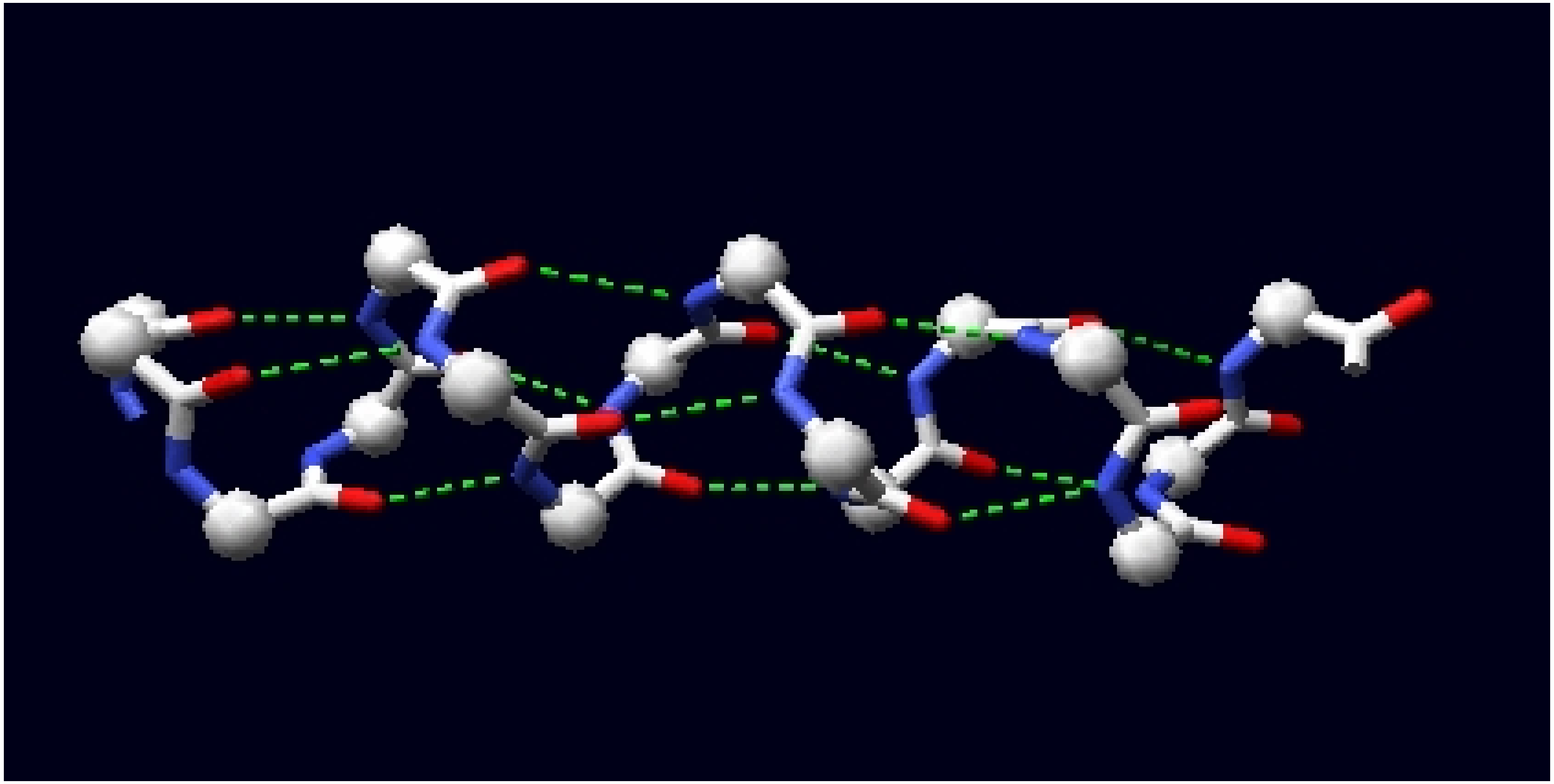
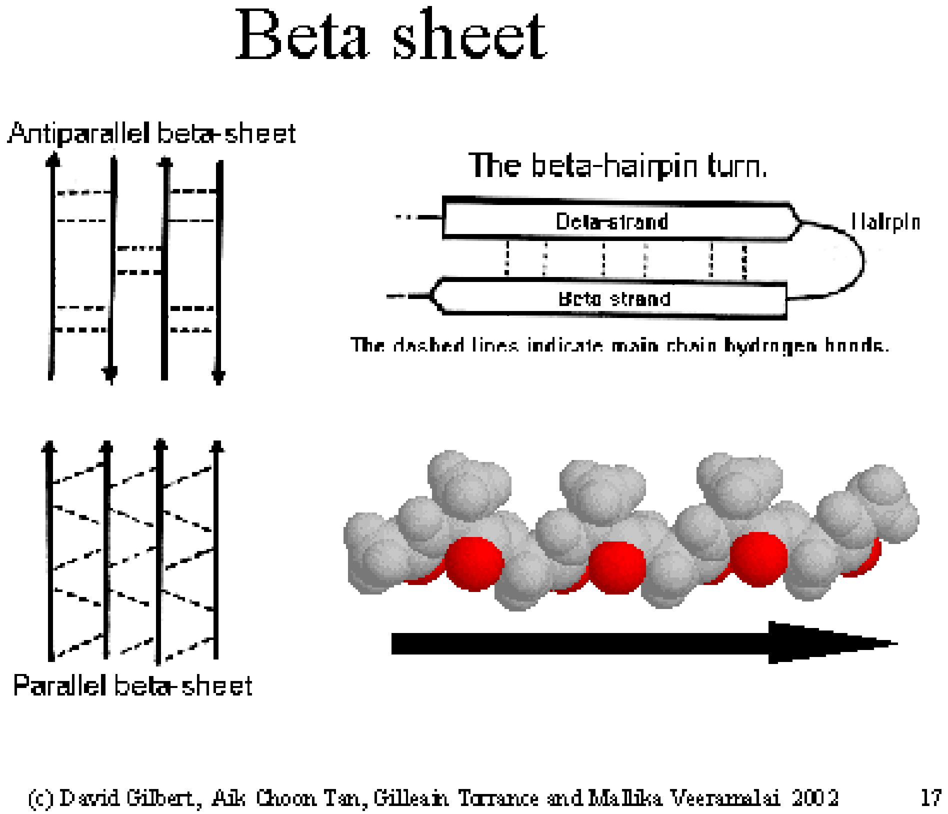


Figure 2.2 The α helix is one of the major elements of secondary structure in proteins. Main-chain N and O atoms are hydrogen-bonded to each other within α helices. (a) Idealized diagram of the path of the main chain in an α helix. Alpha helices are frequently illustrated in this way. There are 3.6 residues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

Antiparallel beta-sheet

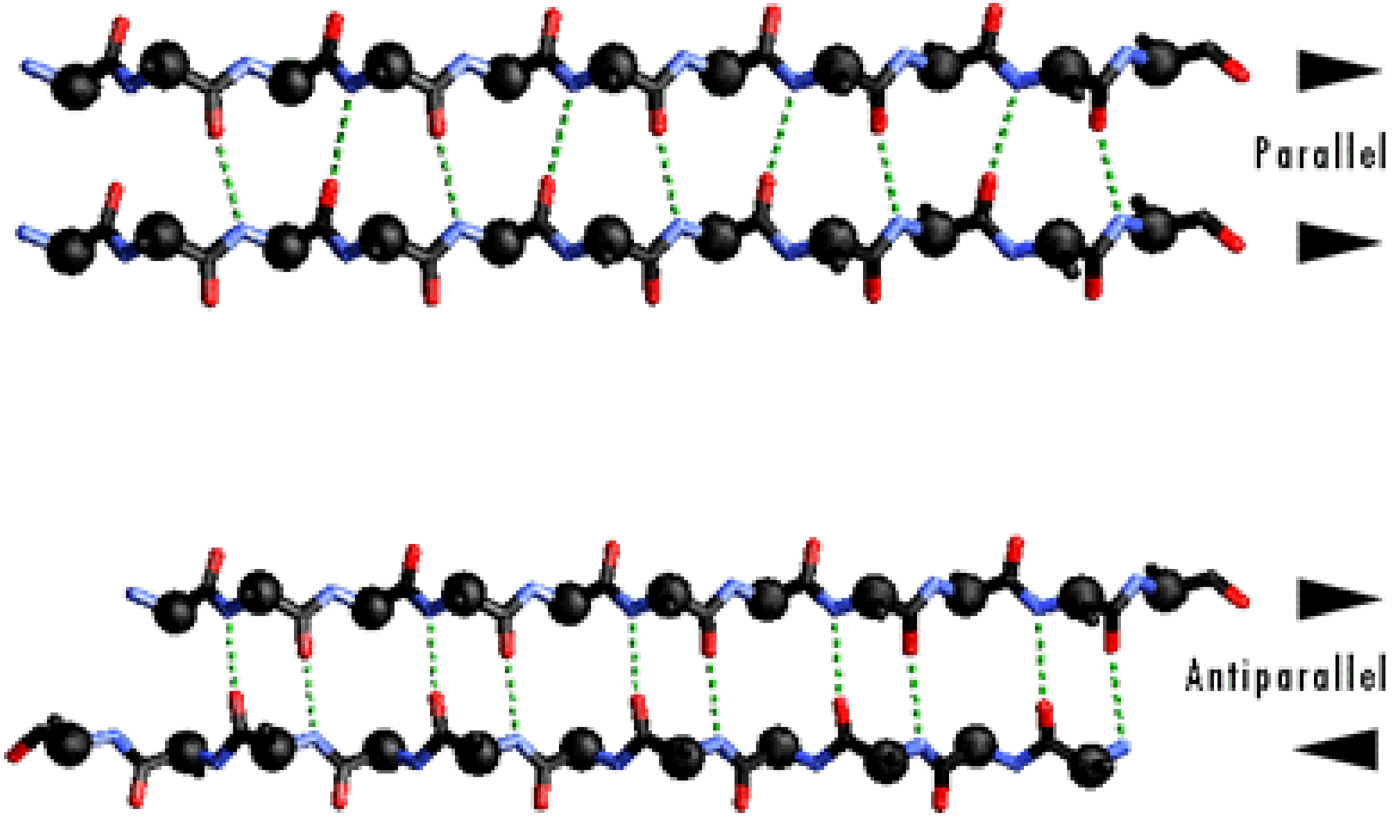
The beta-hairpin turn.

The dashed lines indicate main chain hydrogen bonds.

Parallel beta-sheet

(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.

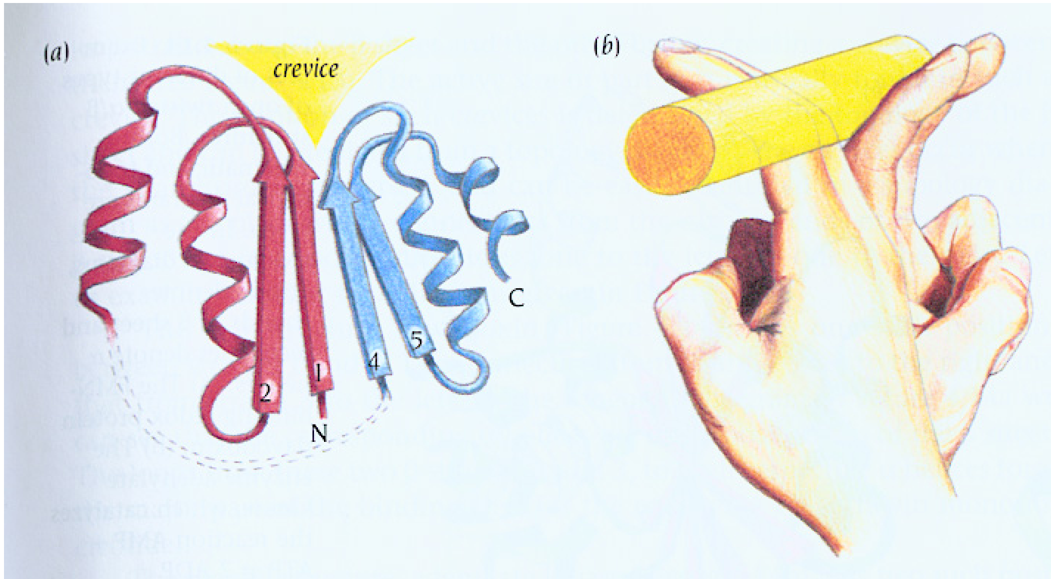
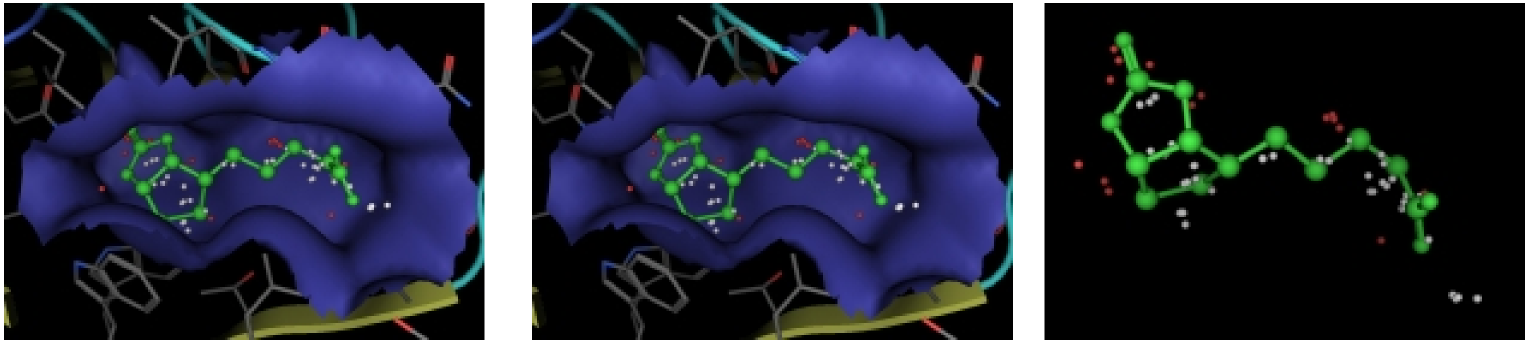


Figure 4.13 (a) The active site in open twisted α/β domains is in a crevice outside the carboxy ends of the β strands. This crevice is formed by two adjacent loop regions that connect the two strands with α helices on opposite sides of the β sheet. This is illustrated by the curled fingers of two hands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

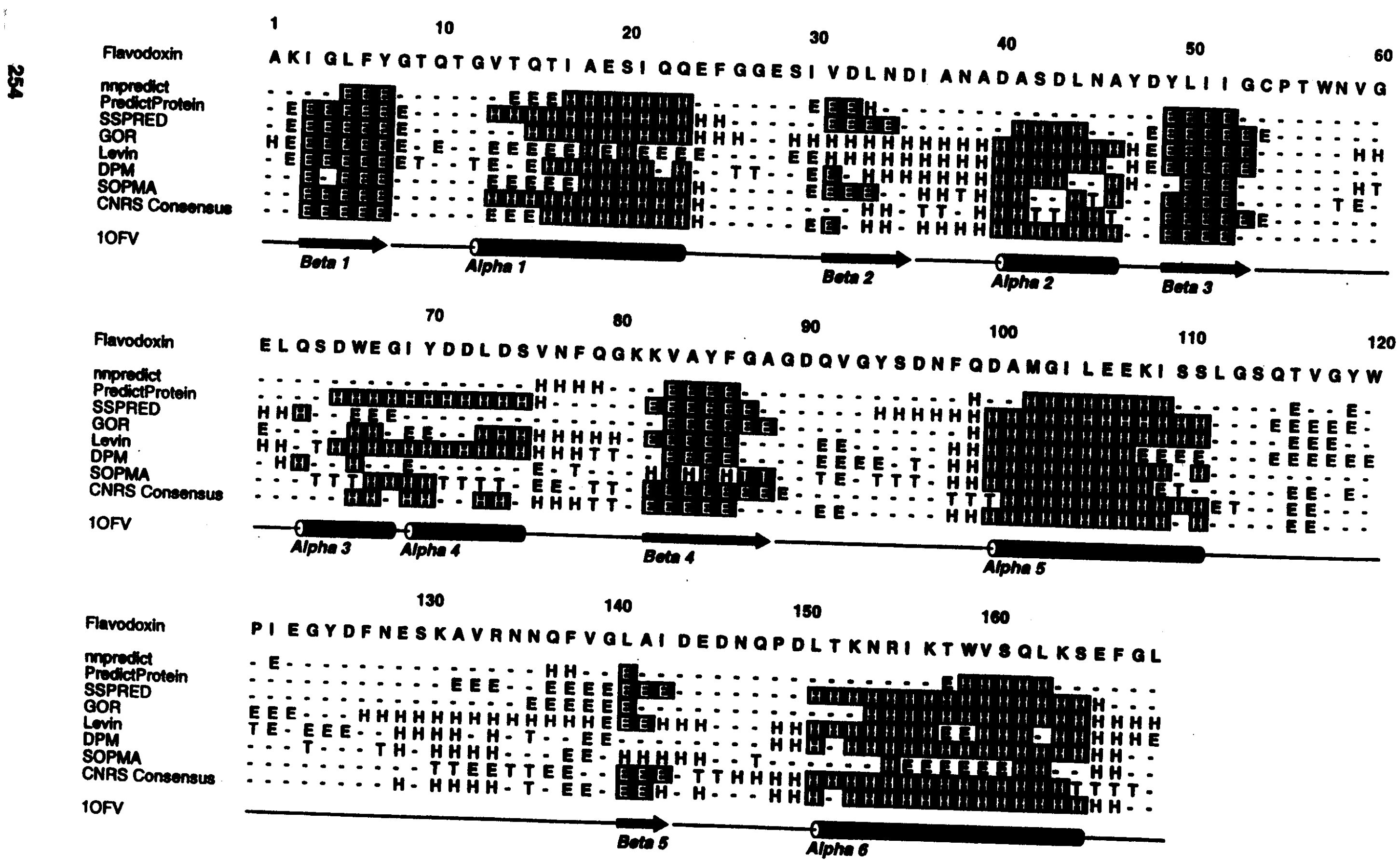
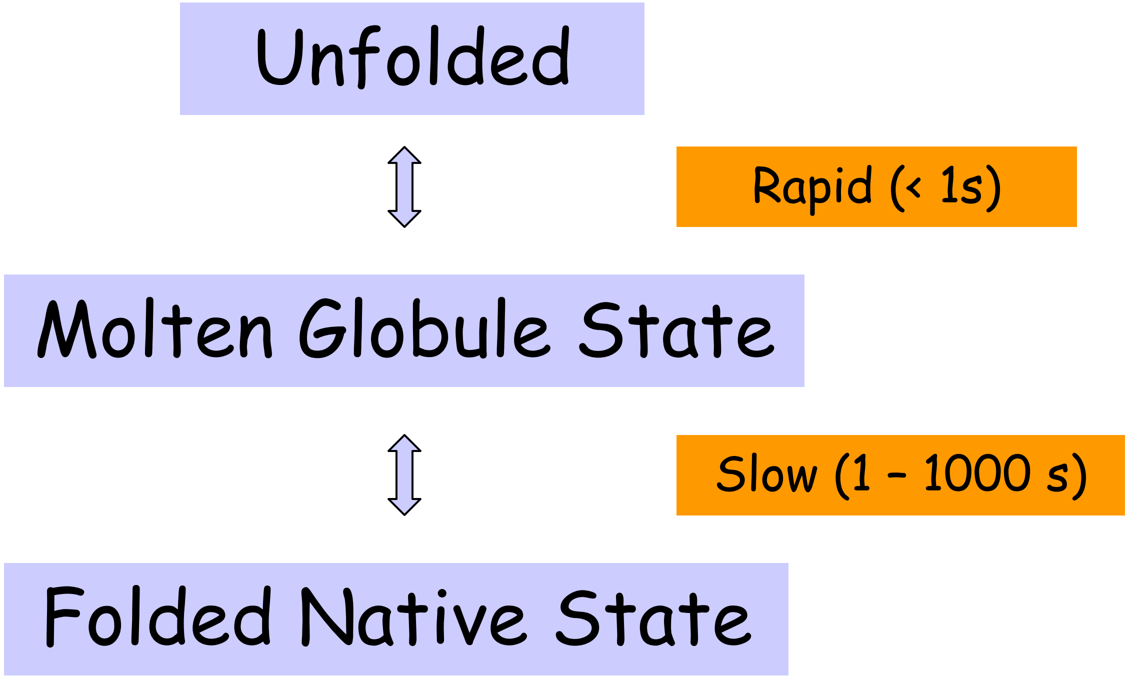

254
Flavodoxin
nnpredict
PredictProtein
SSPRED
GOR
Levin
DPM
SOPMA
CNRS Consensus
1OFV
1 10 20 30 40 50 60
A K I G L F Y G T Q T G V T Q T I A E S I Q Q E F G G E S I V D L N D I A N A D A S D L N A Y D Y L I I G C P T W N V G
Beta 1
Alpha 1
Beta 2
Alpha 2
Beta 3
70 80 90 100 110 120
E L Q S D W E G I Y D D L D S V N F Q G K K V A Y F G A G D Q V G Y S D N F Q D A M G I L E E K I S S L G S Q T V G Y W
Alpha 3
Alpha 4
Beta 4
Alpha 5
130 140 150 160
P I E G Y D F N E S K A V R N N Q F V G L A I D E D N Q P D L T K N R I K T W V S Q L K S E F G L
Beta 5
Alpha 6

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

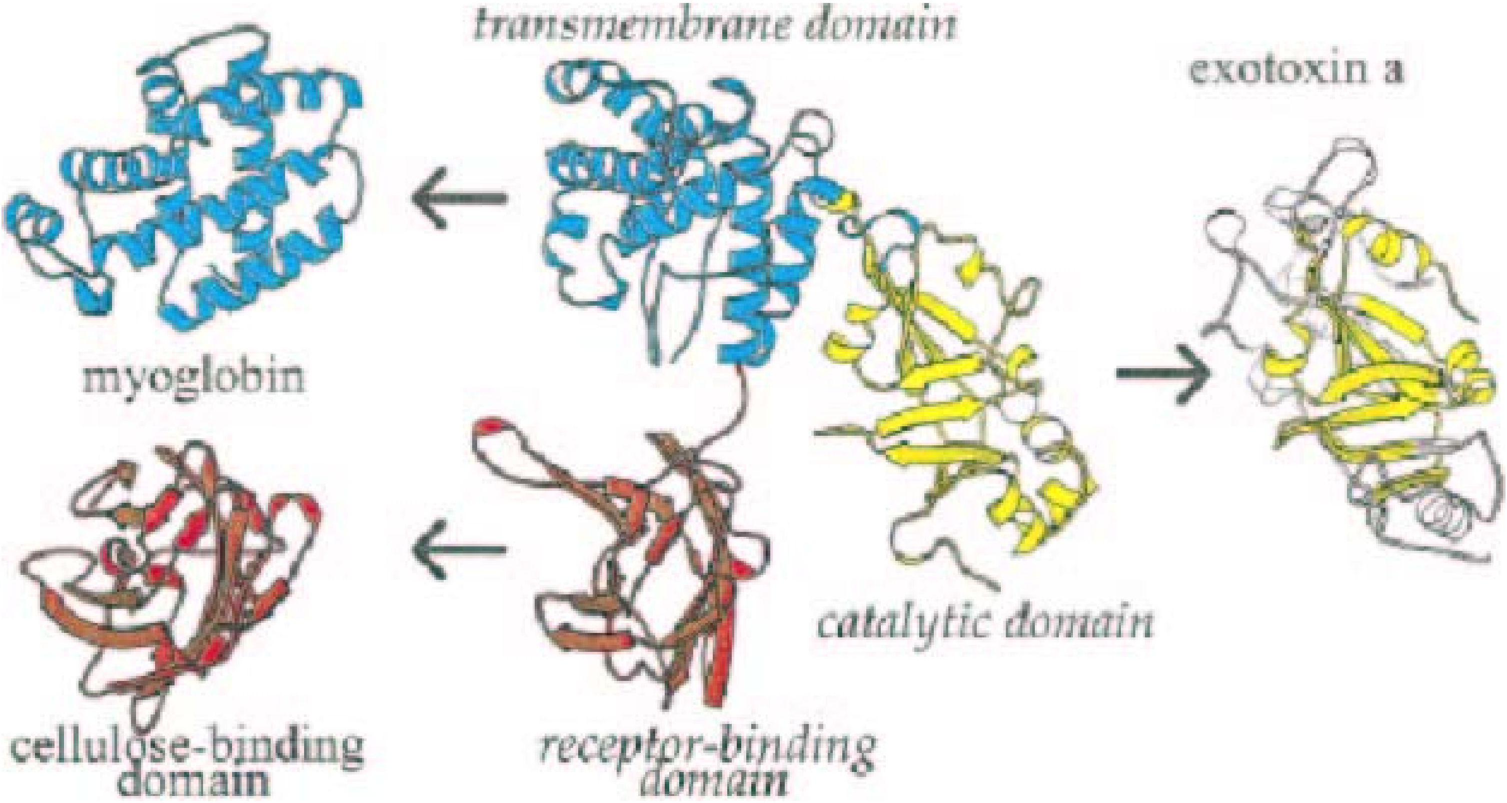
Slow (1 – 1000 s)

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

Structural Classification of Proteins

- Over 1000 protein families known
  - Sequence alignment, motif finding, block finding, similarity search
- SCOP (Structural Classification of Proteins)
  - Based on structural & evolutionary relationships.
  - Contains ~ 40,000 domains
  - Classes (groups of folds), Folds (proteins sharing folds), Families (proteins related by function/evolution), Superfamilies (distantly related proteins)

SCOP Family View

JMB—MS 422

538 *Communication*

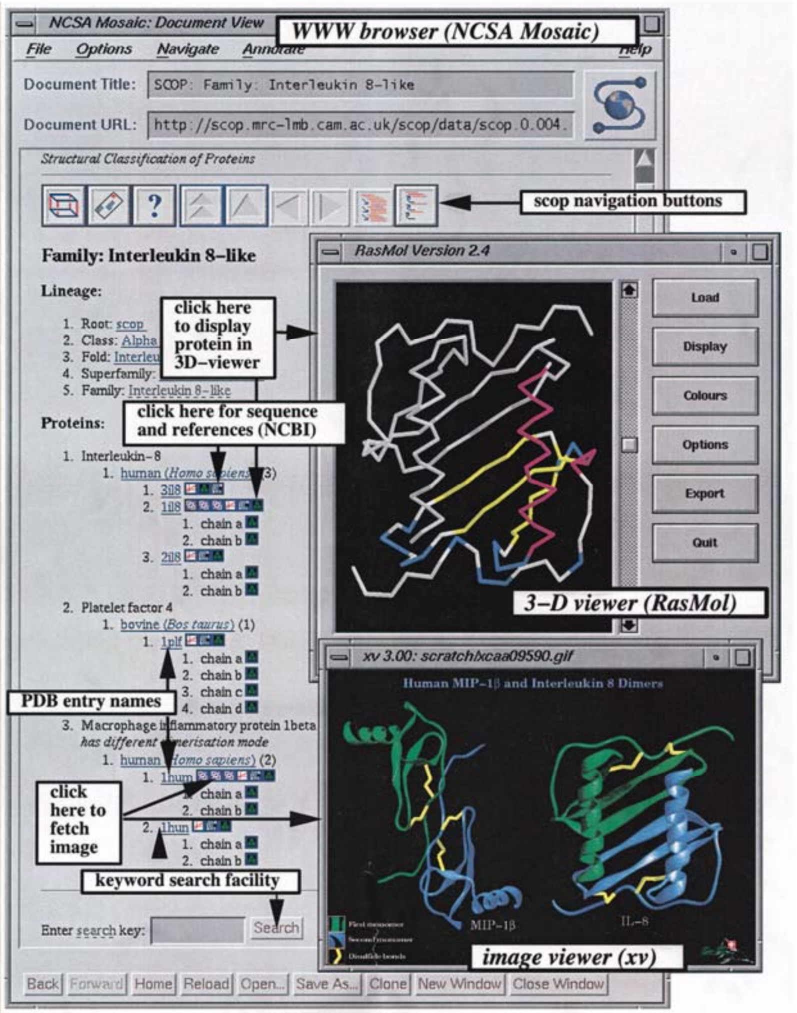


Figure 2. A typical scop session is shown on a unix workstation. A scop page, of the Interleukin 8-like family, is displayed by the *WWW browser program (NCSA Mosaic)* (Schatz & Hardin, 1994). Navigating through the tree structure is accomplished by selecting any underlined entry, by clicking on buttons (at the top of each page) and by keyword searching (at the bottom of each page). The static image comparing two proteins in this family was downloaded by clicking on the icon indicated and is displayed by image-viewer program *xv*. By clicking on one of the green icons, commands were sent to a molecular viewer program (*RasMol*) written by Roger Sayle (Sayle, 1994), instructing it to automatically display the relevant PDB file and colour the domain in question by secondary structure. Since sending large PDB files over the network can be slow, this feature of scop can be configured to use local copies of PDB files if they are available. Equivalent WWW browsers, image-display programs and molecular viewers are also available free for Windows-PC and Macintosh platforms.

CATH: Protein Structure Classification

- Semi-automatic classification; ~36K domains
- 4 levels of classification:
  - Class (C), depends on sec. Str. Content
    - α class, β class, α/β class, α+β class
  - Architecture (A), orientation of sec. Str.
  - Topolgy (T), topological connections &
  - Homologous Superfamily (H), similar str and functions.

DALI/FSSP Database

- Completely automated; 3724 domains
- Criteria of compactness & recurrence
- Each domain is assigned a Domain Classification number DC_l_m_n_p representing fold space attractor region (l), globular folding topology (m), functional family (n) and sequence family (p).

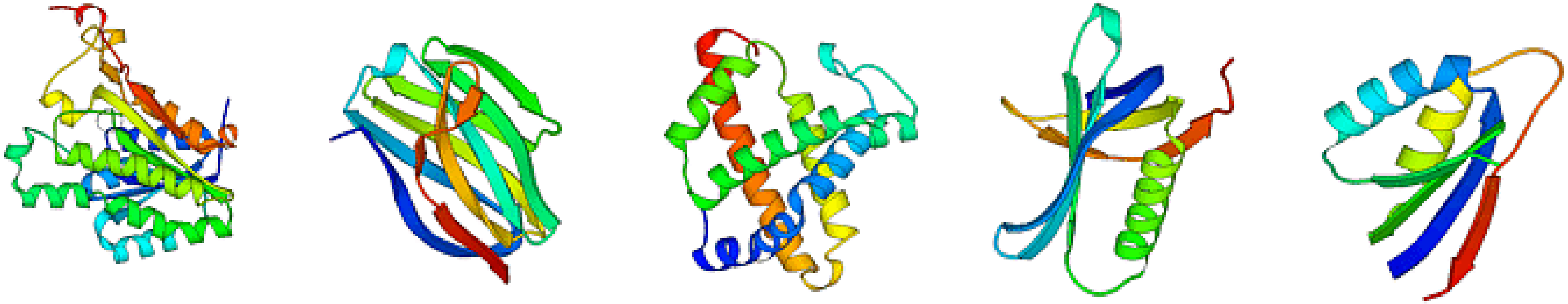
Structural Alignment

- What is structural alignment of proteins?
  - 3-d superimposition of the atoms as “best as possible”, i.e., to minimize RMSD (root mean square deviation).
  - Can be done using VAST and SARF
- Structural similarity is common, even among proteins that do not share sequence similarity or evolutionary relationship.

Other databases & tools

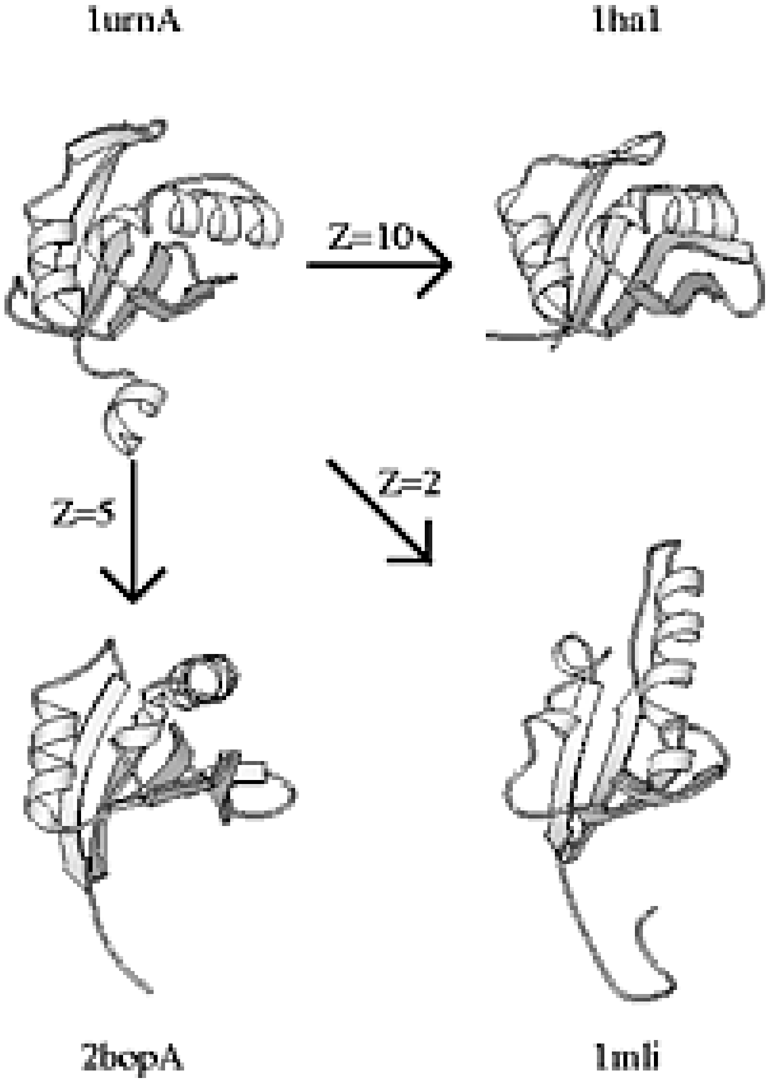
- ❑MMDB contains groups of structurally related proteins
- ❑SARF structurally similar proteins using secondary structure elements
- ❑VAST Structure Neighbors
- ❑SSAP uses double dynamic programming to structurally align proteins

5 Fold Space classes



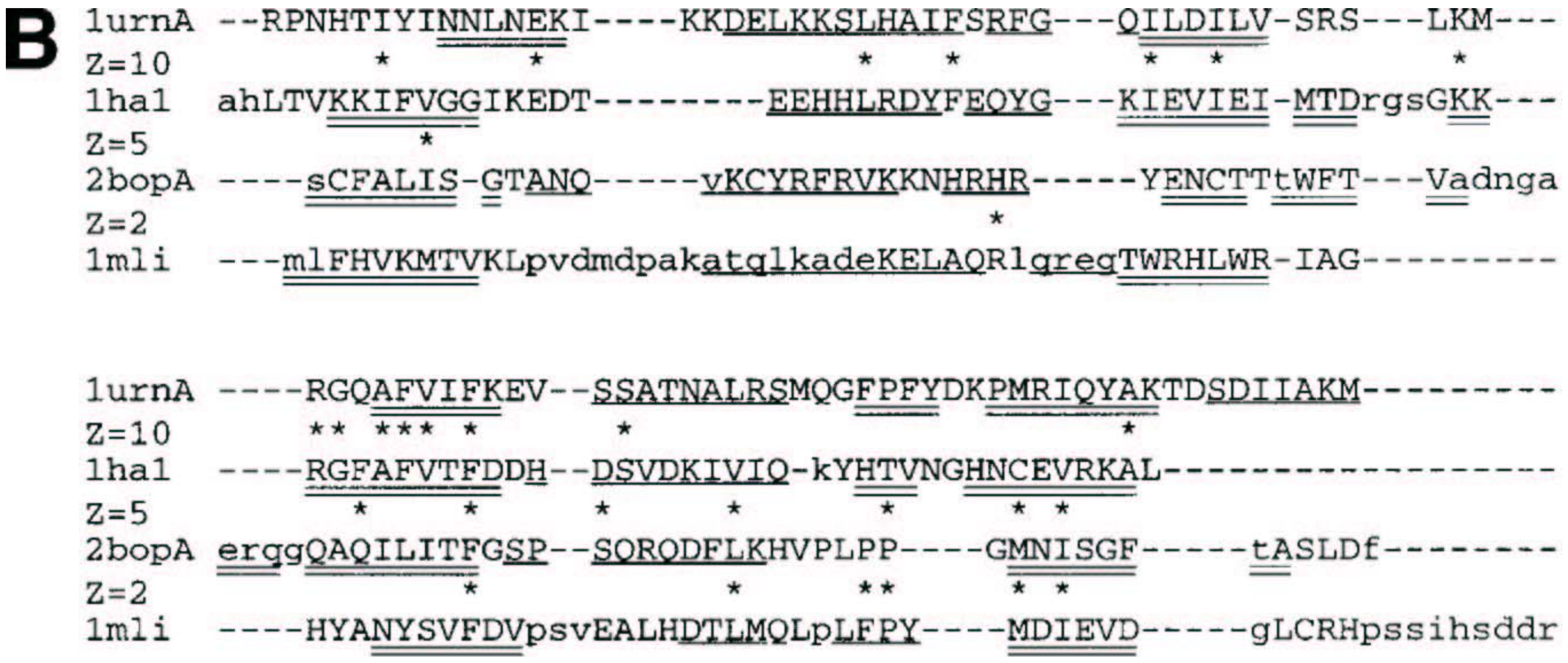
Attractor 1 can be characterized as alpha/beta,
attractor 2 as all-beta, attractor 3 as all-alpha,
attractor 5 as alpha-beta meander (1mli), and
attractor 4 contains antiparallel beta-barrels e.g. OB-fold (1prtF).

Fold Types & Neighbors



Structural neighbours of 1urnA (top left). 1mli (bottom right) has the same topology even though there are shifts in the relative orientation of secondary structure elements.

Sequence Alignment of Fold Neighbors

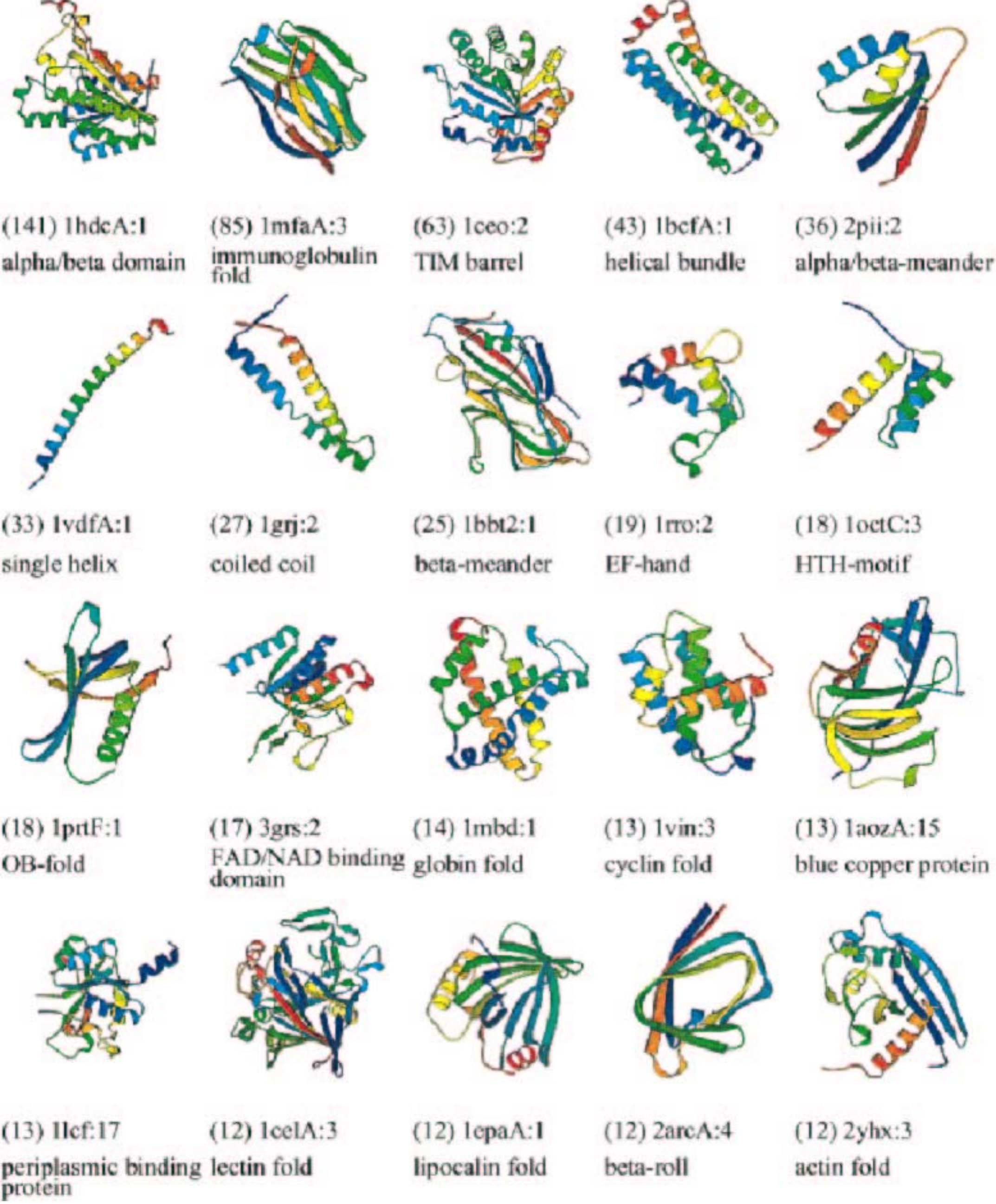


B

```
1urnA --RPNHTIYINNLNEKI----KKDELKKSLHAIFSRFG---QILDILV-SRS---LKM---
Z=10         *      *              *   *        *  *          *
1ha1  ahLTVKKIFVGGIKEDT--------EEHHLRDYFEQYG---KIEVIEI-MTDrgsGKK---
Z=5            *
2bopA ----sCFALIS-GTANQ-----vKCYRFRVKKNHRHR-----YENCTTtWFT---Vadnga
Z=2                                      *
1mli  ---mlFHVKMTVKLpvdmdpakatqlkadeKELAQRlqreqTWRHLWR-IAG---------

1urnA ----RGQAFVIFKEV--SSATNALRSMQGFPFYDKPMRIQYAKTDSDIIAKM---------
Z=10      ** *** *      *                      *
1ha1  ----RGFAFVTFDDH--DSVDKIVIQ-kYHTVNGHNCEVRKAL------------------
Z=5         *    *     *     *      *     * *
2bopA erqgQAQILITFGSP--SQRQDFLKHVPLPP----GMNISGF-----tASLDf--------
Z=2              *           *     **     * *
1mli  ----HYANYSVFDVpsvEALHDTLMQLpLFPY----MDIEVD-----gLCRHpssihsddr
```

Frequent Fold Types



Protein Structure Prediction

- Holy Grail of bioinformatics
- Protein Structure Initiative to determine a set of protein structures that span protein structure space sufficiently well. WHY?
  - Number of folds in natural proteins is limited. Thus a newly discovered proteins should be within modeling distance of some protein in set.
- CASP: Critical Assessment of techniques for structure prediction
  - To stimulate work in this difficult field

PSP Methods

- homology-based modeling
- methods based on fold recognition
  - Threading methods
- *ab initio* methods
  - From first principles
  - With the help of databases

ROSETTA

- Best method for PSP
- As proteins fold, a large number of partially folded, low-energy conformations are formed, and that local structures combine to form more global structures with minimum energy.
- Build a database of known structures (I-sites) of short sequences (3-15 residues).
- Monte Carlo simulation assembling possible substructures and computing energy

Threading Methods

- See p471, Mount
  - http://www.bioinformaticsonline.org/links/ch_10_t_7.html

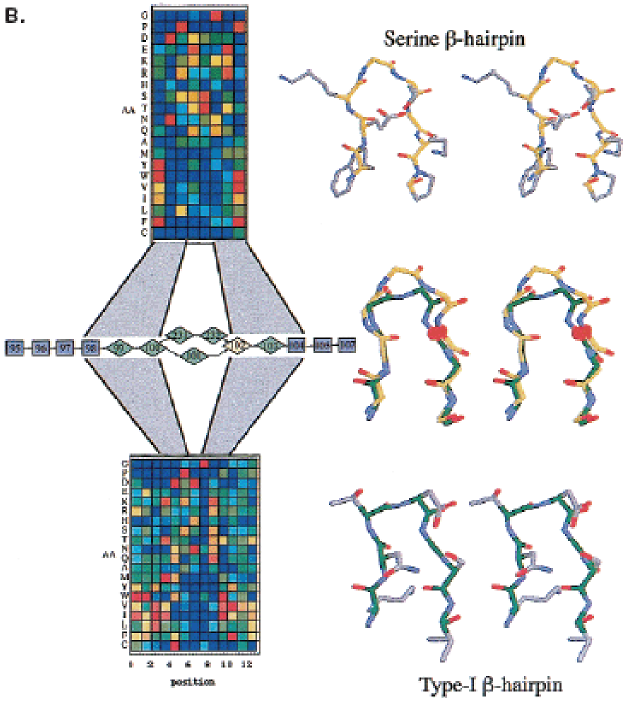


FIGURE 10.30. A hidden Markov model (discrete state-space model) of protein three-dimensional structure. (*B*) HMM called HMMSTR based on I-sites, 3- to 15-amino-acid patterns that are associated with three-dimensional structural features. The two matrices with colored squares represent alignment of sets of patterns that are found to be associated with a structure, in this case the hairpin turns shown on the right. Each column in the table corresponds to the amino acid variation found for one structural position in one of the turns. (*Blue* side chains) Conserved nonpolar residues; (*green*) conserved polar residues; (*red*) conserved proline; and (*orange*) conserved glycine. The two hairpins are aligned structurally in the middle structure on the right and the observed variation in the corresponding amino acid positions is represented by the HMM between the matrices on the left. The HMM represents an alignment of the two hairpin structural motifs in three-dimensional space and an alignment of the sequences. A short mismatch in the turn is represented by splitting the model into two branches. The shaped icons represent states, each of which represents a structure and a sequence position. Each state contains probability distributions about the sequence and structural attributes of a single position in the motif, including the probability of observing a particular amino acid, secondary structure, Φ-Ψ backbone angles, and structural context, e.g., location of β strand in a β sheet. Rectangles are predominantly β-strand states, and diamonds are predominantly turns. The color of the icon indicates a sequence preference as follows: (*blue*) hydrophobic; (*green*) polar; and (*yellow*) glycine. Numbers in icons are arbitrary identification numbers for the HMM states. There is a transition probability of moving from each state in the model to the next, as in HMMs that represent msa's. This model is a small component of the main HMMSTR model that represents a merging of the entire I-sites library. Three different models, designated λ^D, λ^C, and λ^R, are included in HMMSTR, which differ in details as to how the alignment of the I-sites was obtained to design the branching patterns (topology) of the model and which structural data were used to train the model. HMMSTR may be used for a variety of different predictions, including secondary structure prediction, structural context prediction, and Φ-Ψ dihedral angle prediction. Predictions are made by aligning the model with a sequence, finding if there is a high-scoring alignment, and deciphering the highest-scoring path through the model. The HMMSTR program may be downloaded or used on a server that can be readily located by a Web search. (*B*, reprinted, with permission, from Bystroff et al. 2000 [©2000 Elsevier].)