

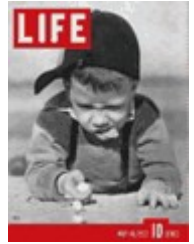
COMP598: Introduction to Protein Structure Prediction

Jérôme Waldispühl

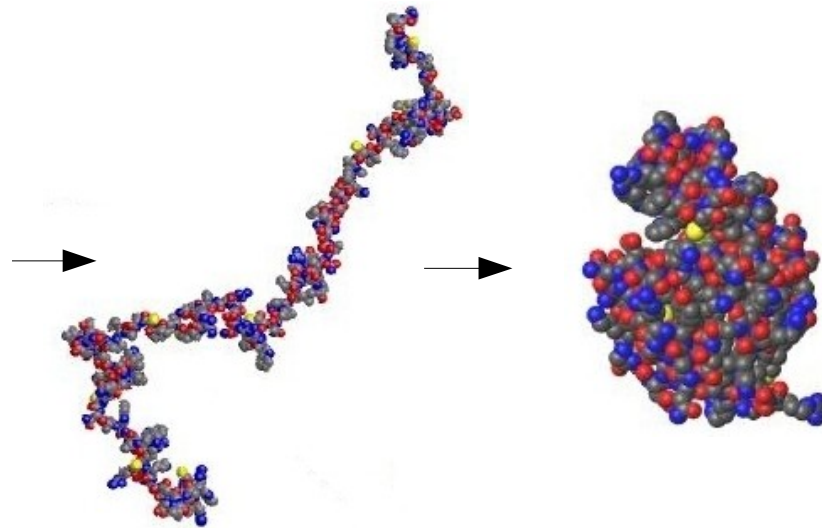
School of Computer Science &
McGill Centre of Bioinformatics
jeromew@cs.mcgill.ca

Features slides from Jinbo Xu – TTI-Chicago

Folding problem



K
L
H
G
G
P
M
L
D
S
D
Q
K
F
W
R
T
P
A
A
L
H
Q
N
E
G
F
T



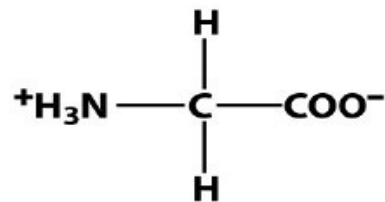
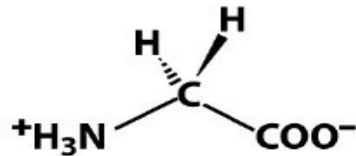
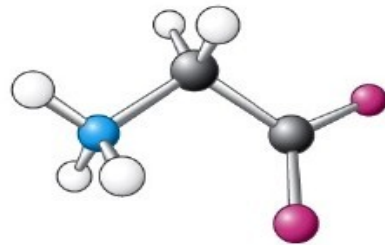
$$N_{\text{états}} \sim 10^n$$

$$n = 100-300$$

Levinthal paradox

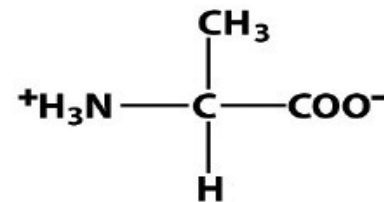
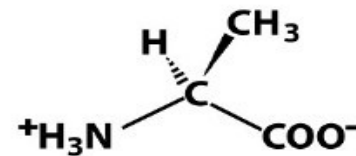
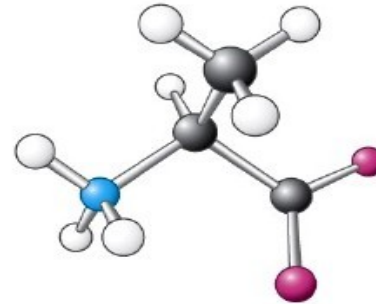
Amino acids: The simple ones

**Glycine
(Gly, G)**



**Glycine
(Gly, G)**

**Alanine
(Ala, A)**



**Alanine
(Ala, A)**

Figure 2-7
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

Amino acids: Aliphatics

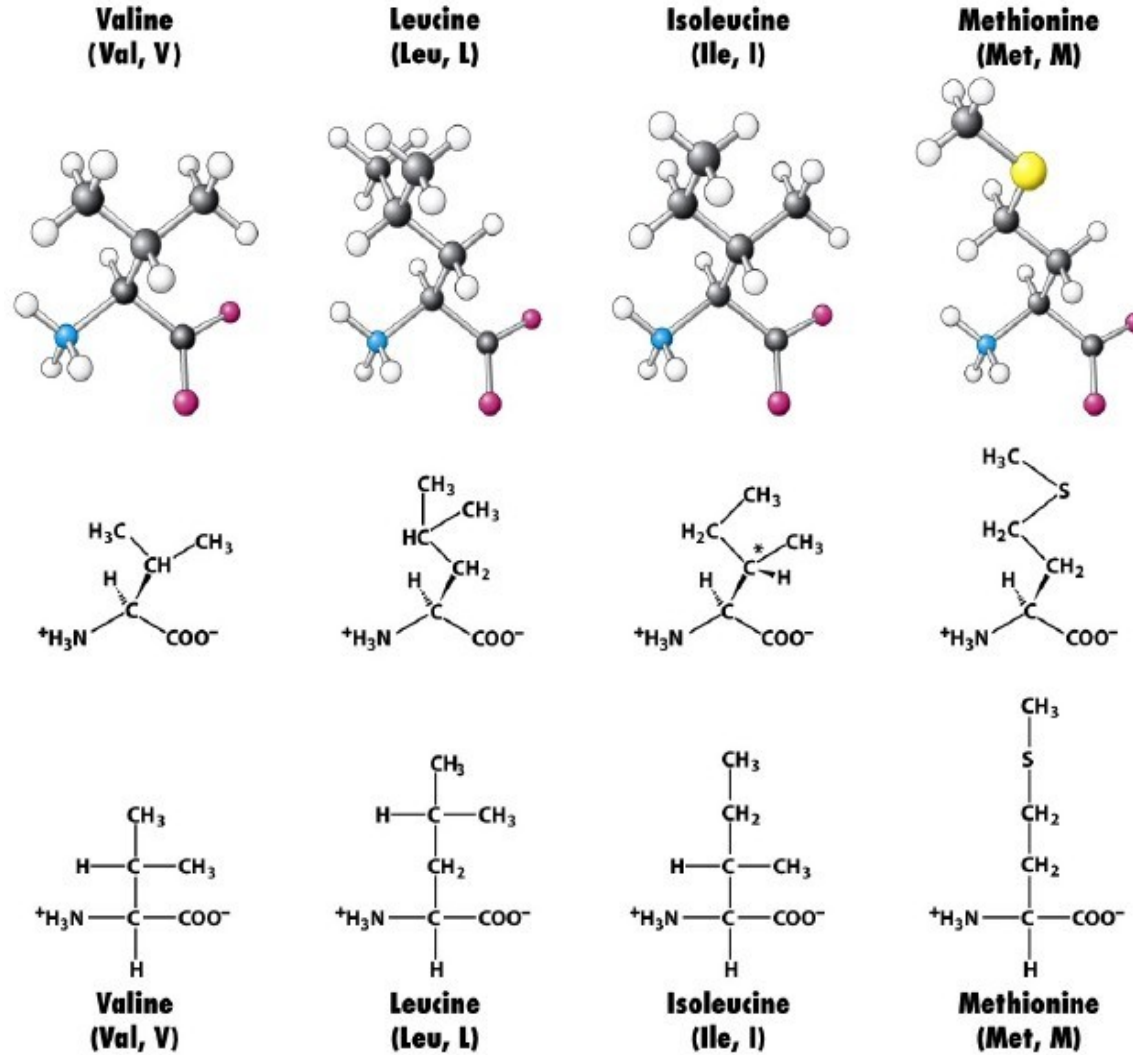


Figure 2-8
Biochemistry, Sixth Edition
 © 2007 W. H. Freeman and Company

Amino acids: Cyclic and Sulfhydryl

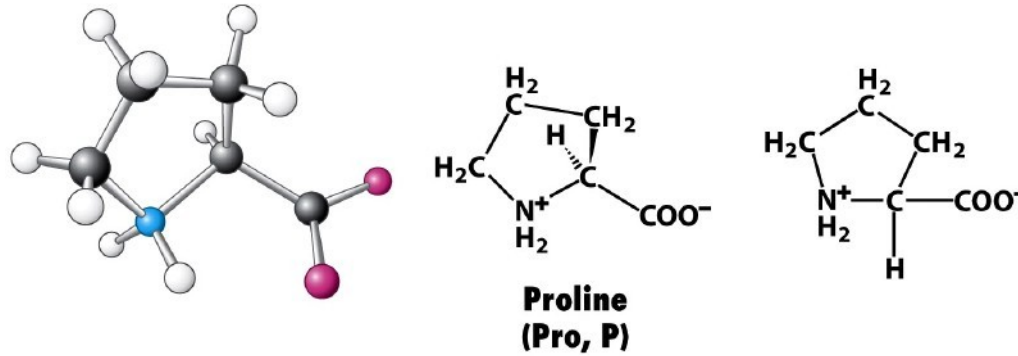


Figure 2-9
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

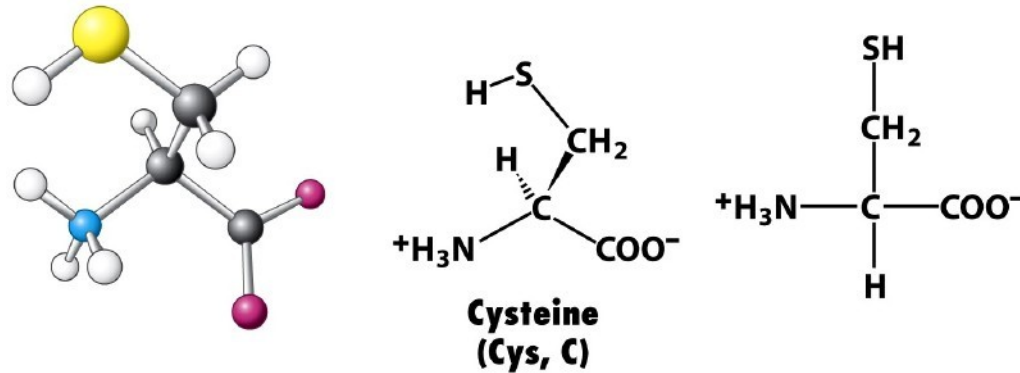


Figure 2-13
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Amino acids: Aromatics

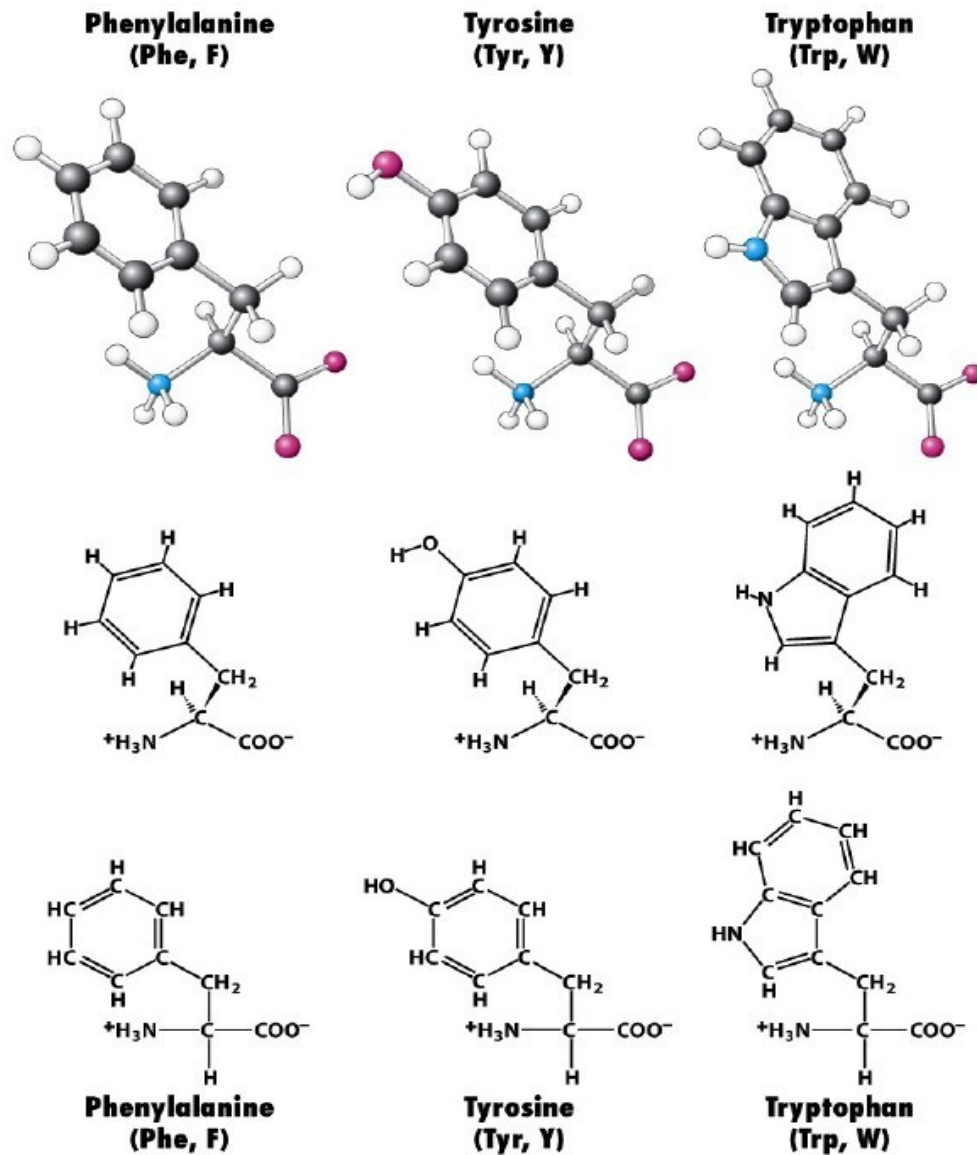
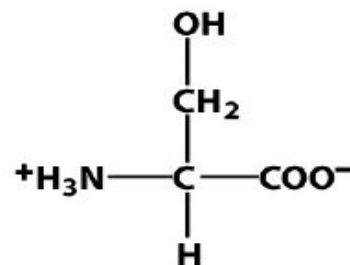
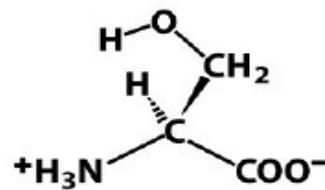
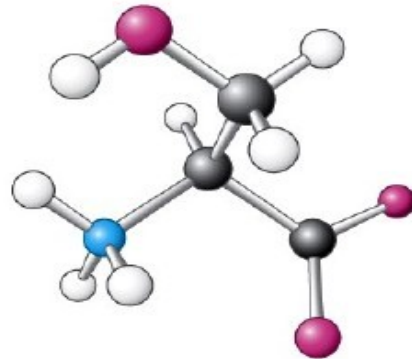


Figure 2-10
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

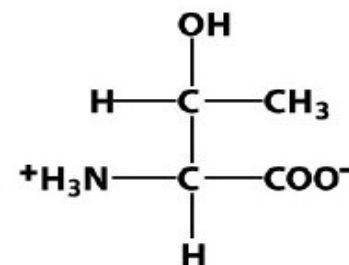
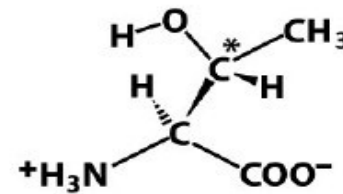
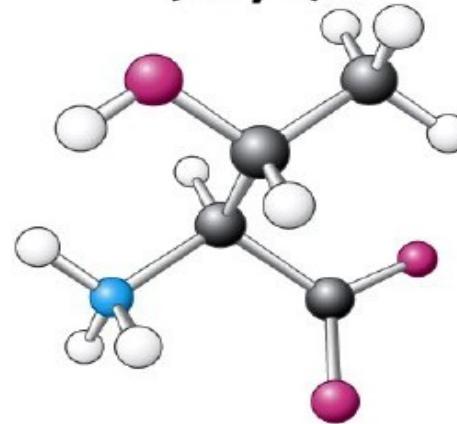
Amino acids: Aliphatic hydroxyl

Serine
(Ser, S)



Serine
(Ser, S)

Threonine
(Thr, T)

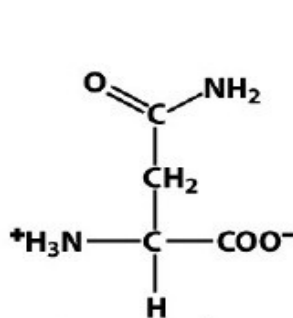
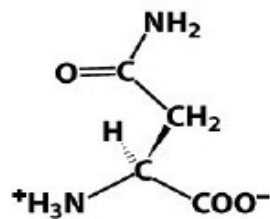
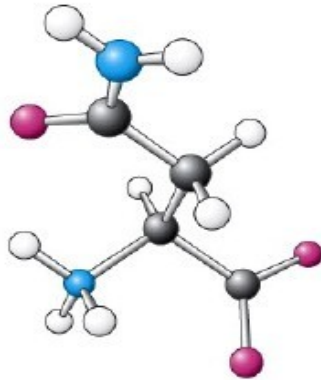


Threonine
(Thr, T)

Figure 2-11
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

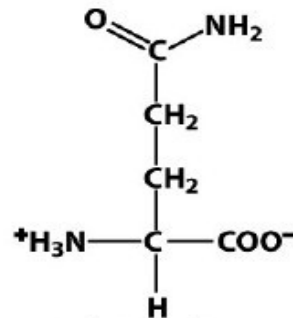
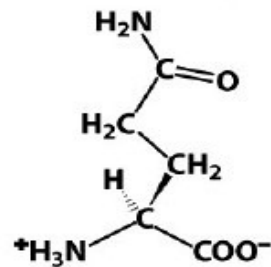
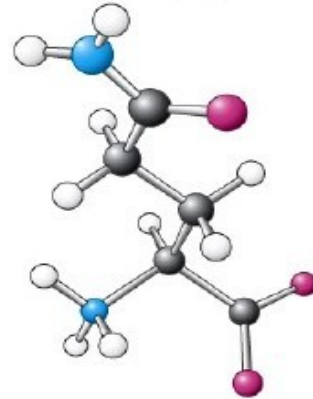
Amino acids: Carboxamides & Carboxylates

**Asparagine
(Asn, N)**



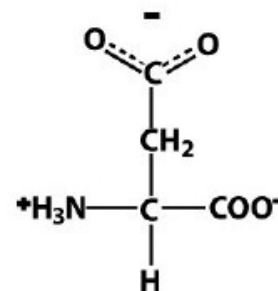
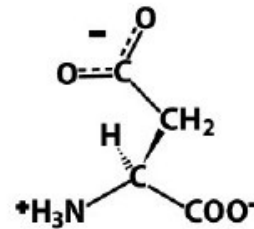
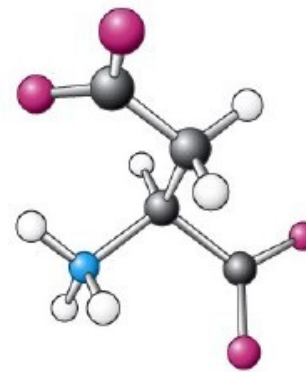
**Asparagine
(Asn, N)**

**Glutamine
(Gln, Q)**



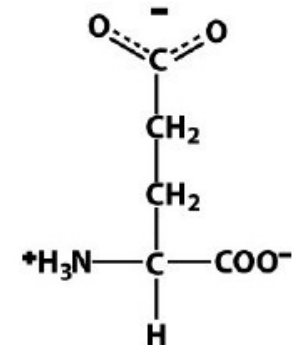
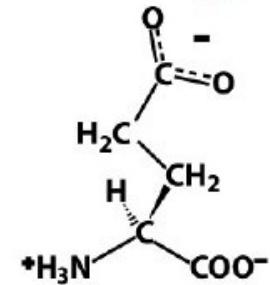
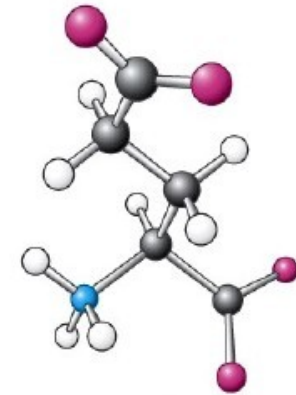
**Glutamine
(Gln, Q)**

**Aspartate
(Asp, D)**



**Aspartate
(Asp, D)**

**Glutamate
(Glu, E)**



**Glutamate
(Glu, E)**

Figure 2-12
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

Figure 2-16
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

Amino acids: Basics

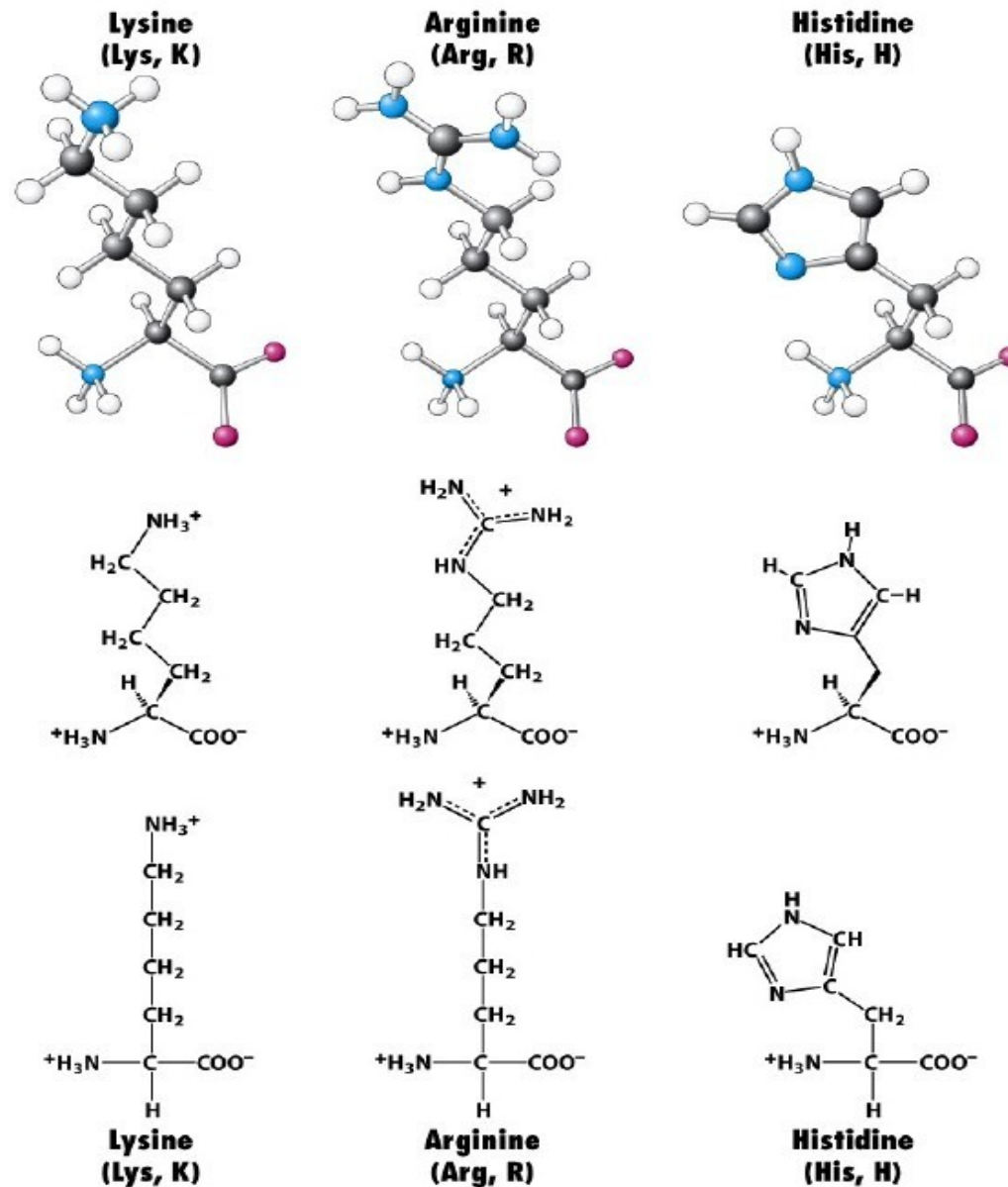


Figure 2-14
Biochemistry, Sixth Edition
 © 2007 W.H. Freeman and Company

Histidine ionisation

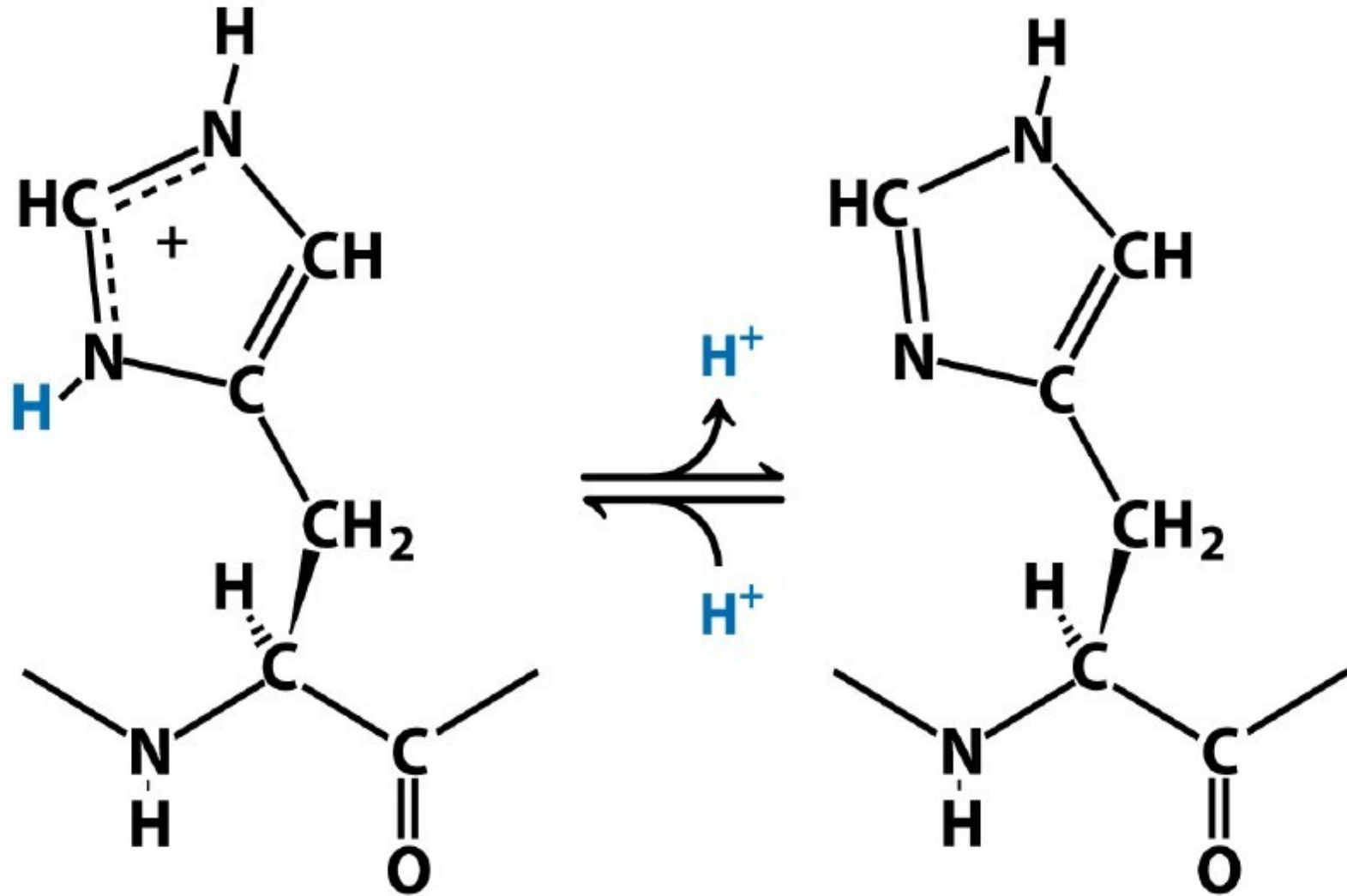


Figure 2-15
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Primary structure

A peptide bond assemble two amino acids together:

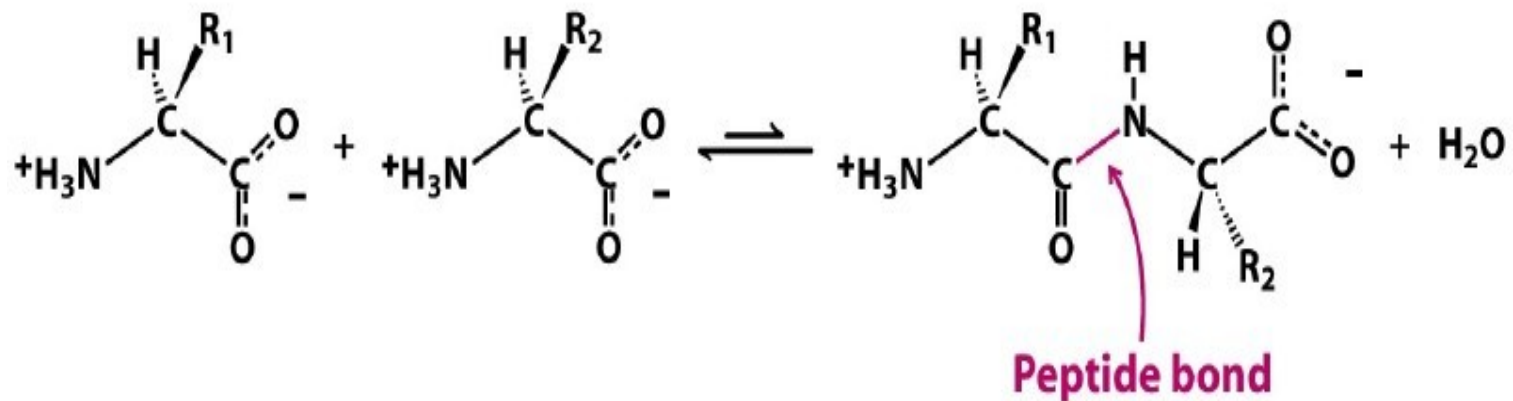


Figure 2-18
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

A chain is obtained through the concatenation of several amino acids:

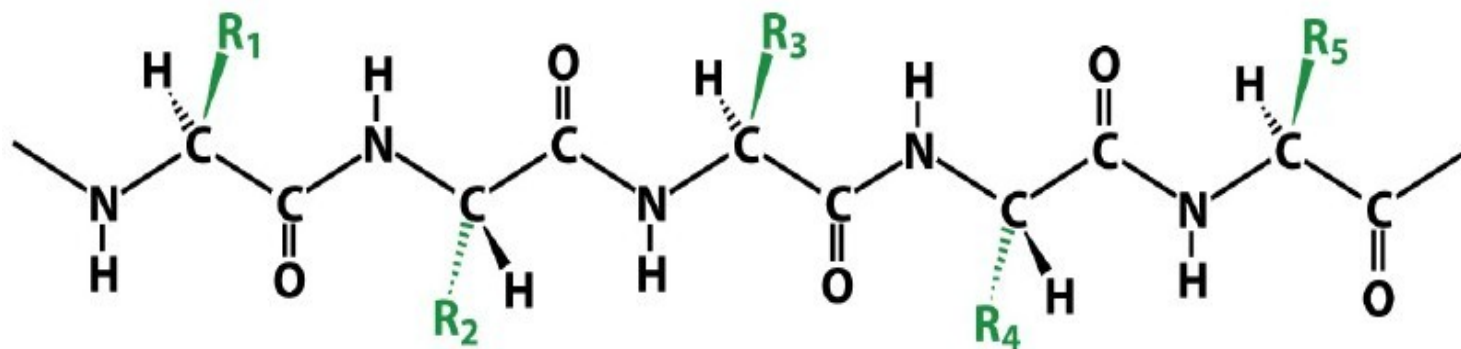


Figure 2-20
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

Peptide bond is pH dependent

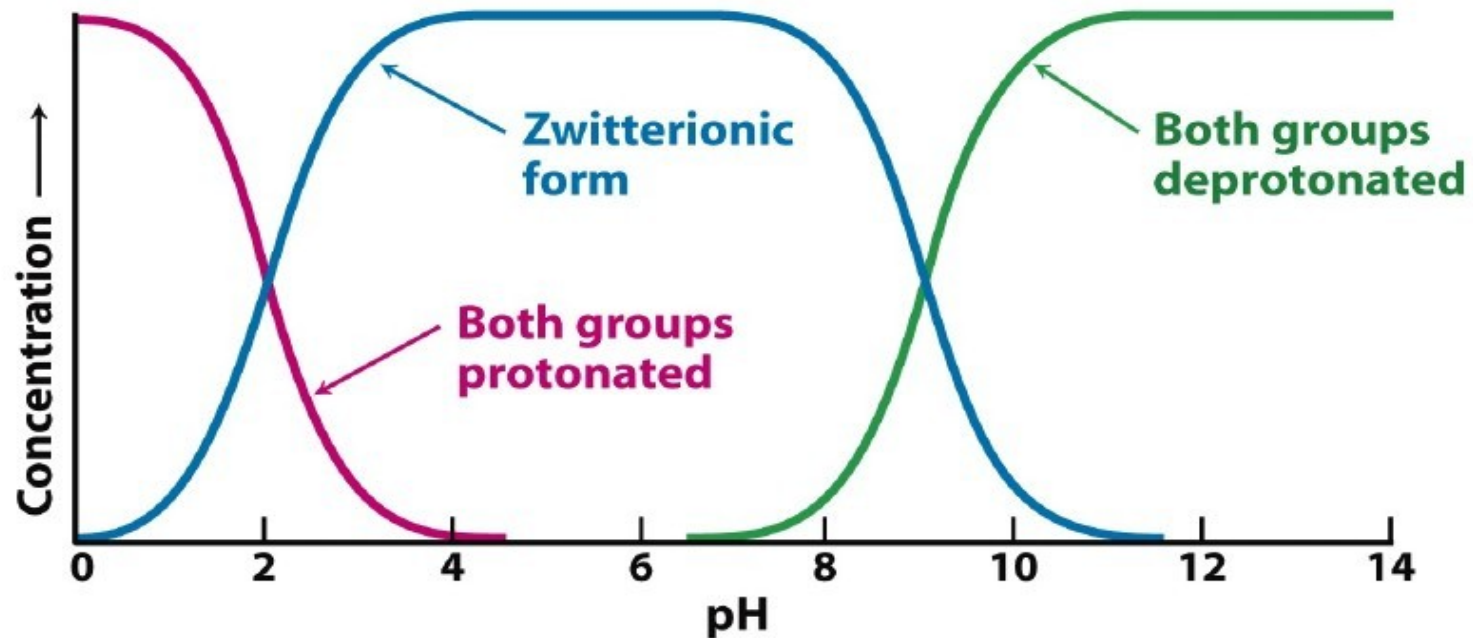
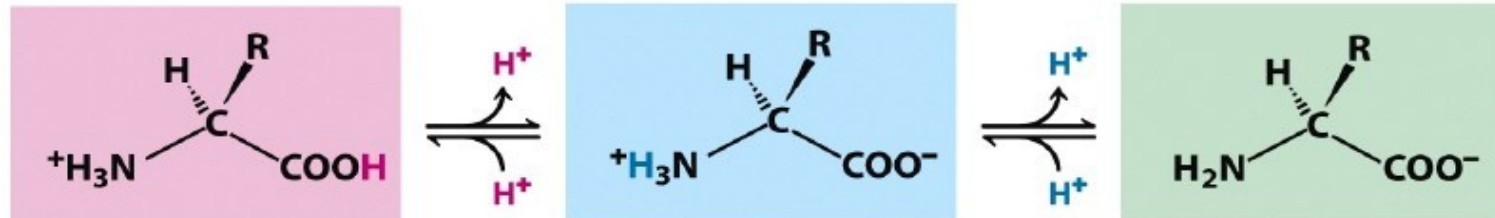
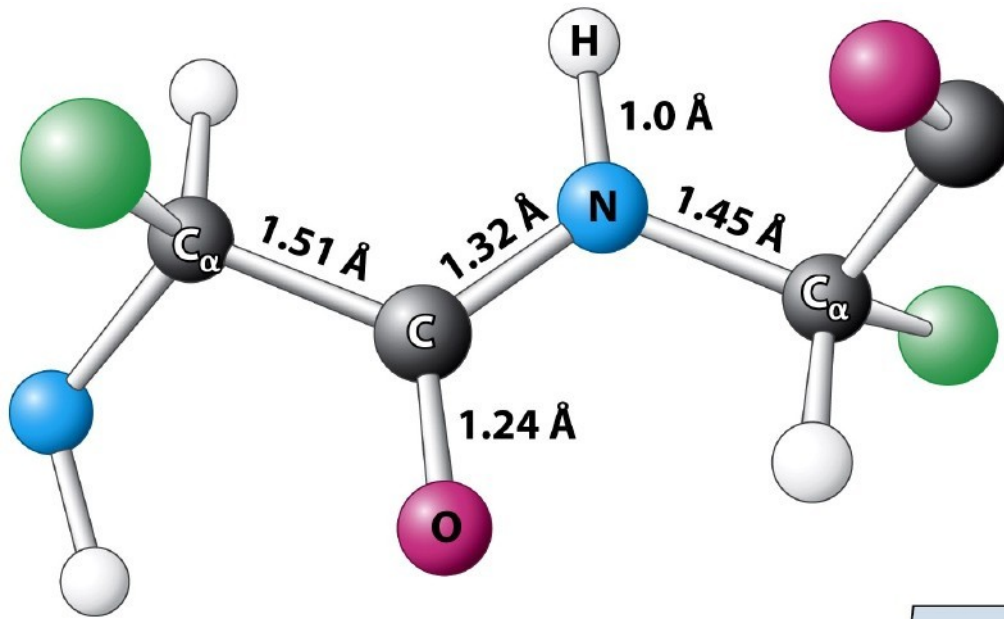


Figure 2-6
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Peptide bond features (1)



Bond lengths

Figure 2-24
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Peptide bond lies on a plane

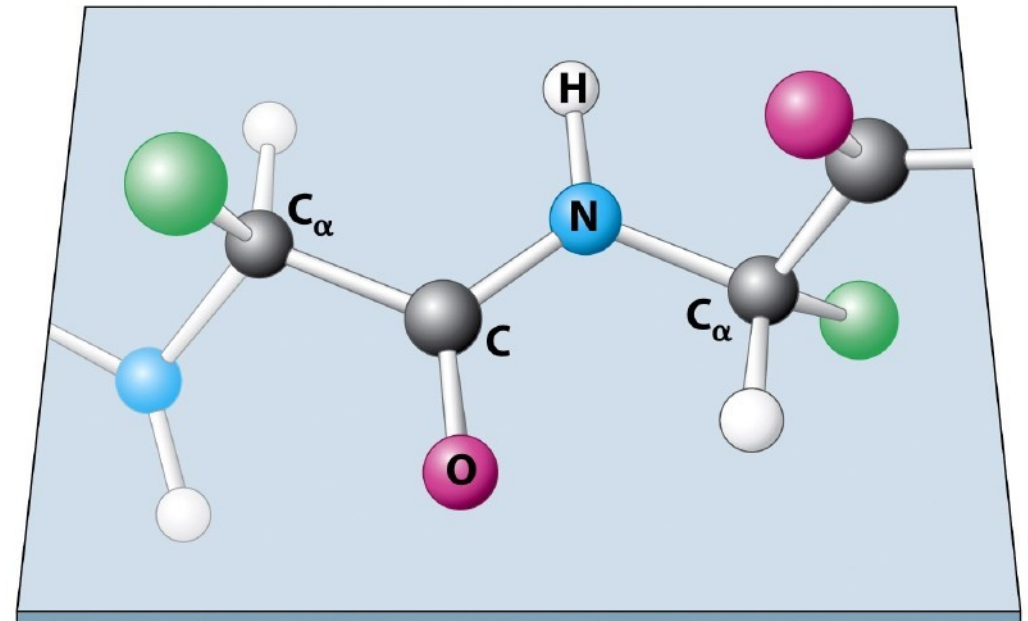


Figure 2-23
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Peptide bond features (2)

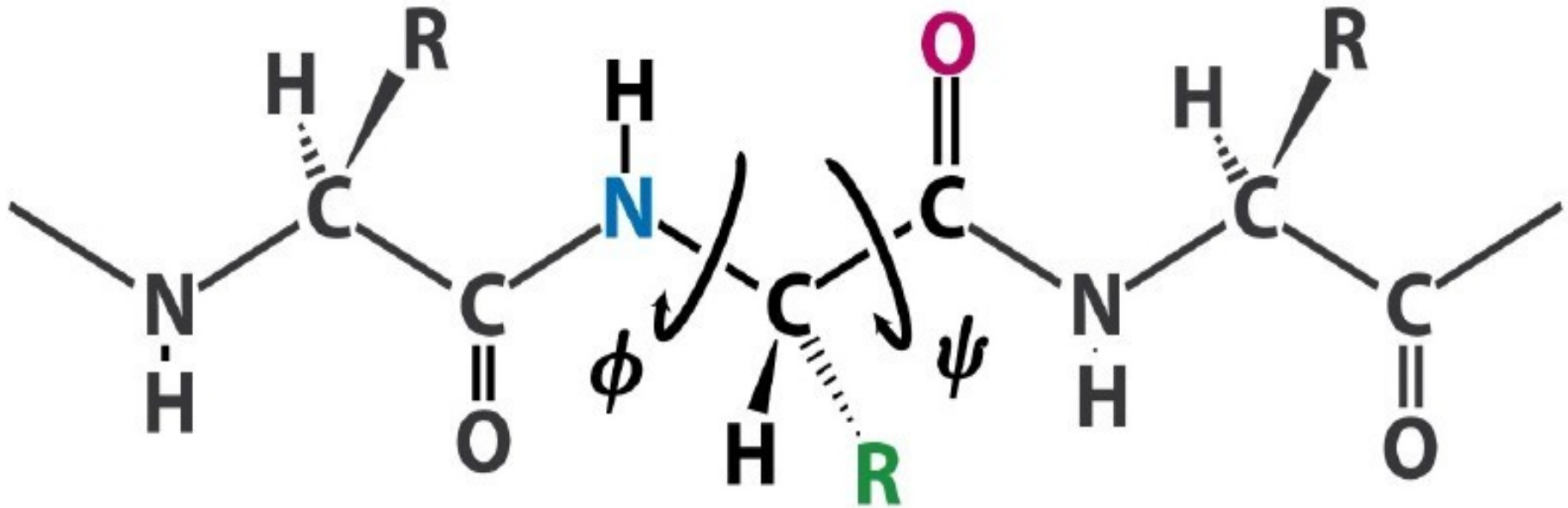


Figure 2-27a
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

The chain has 2 degrees of liberty given by the dihedral angles Φ and Ψ .
The geometry of the chain can be characterized though Φ and Ψ .

Peptide bond features (3)

Cis/trans isomers of the peptide group

Trans configuration is preferred versus Cis (ratio ~1000:1)

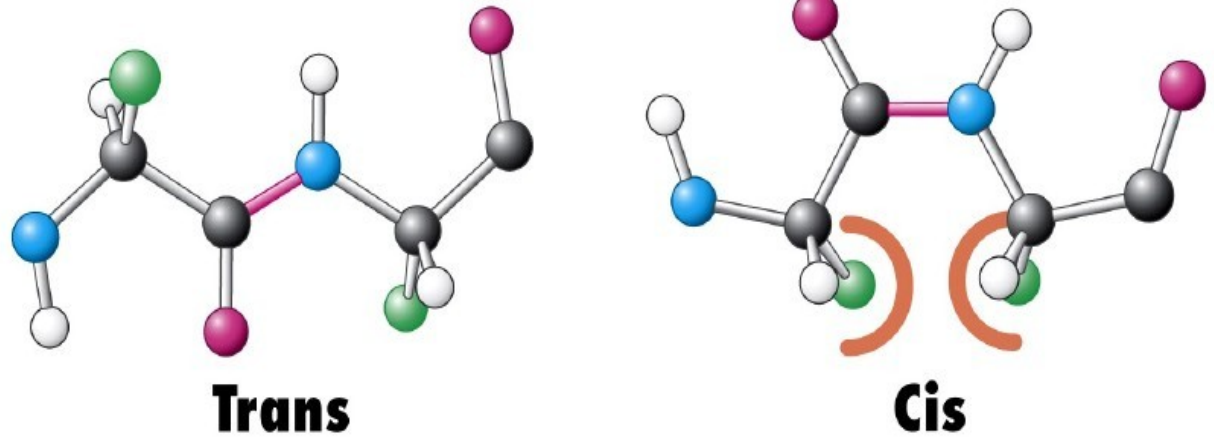


Figure 2-25
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

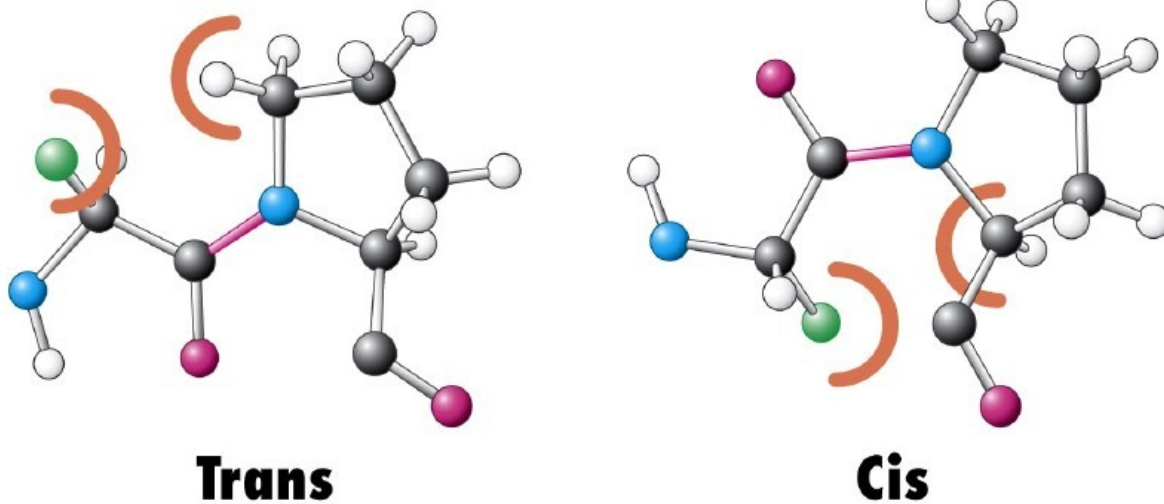


Figure 2-26
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

An exception is the Proline with a preference ratio of ~3:1

Ramachandran diagram gives the values which can be adopted by Φ and Ψ

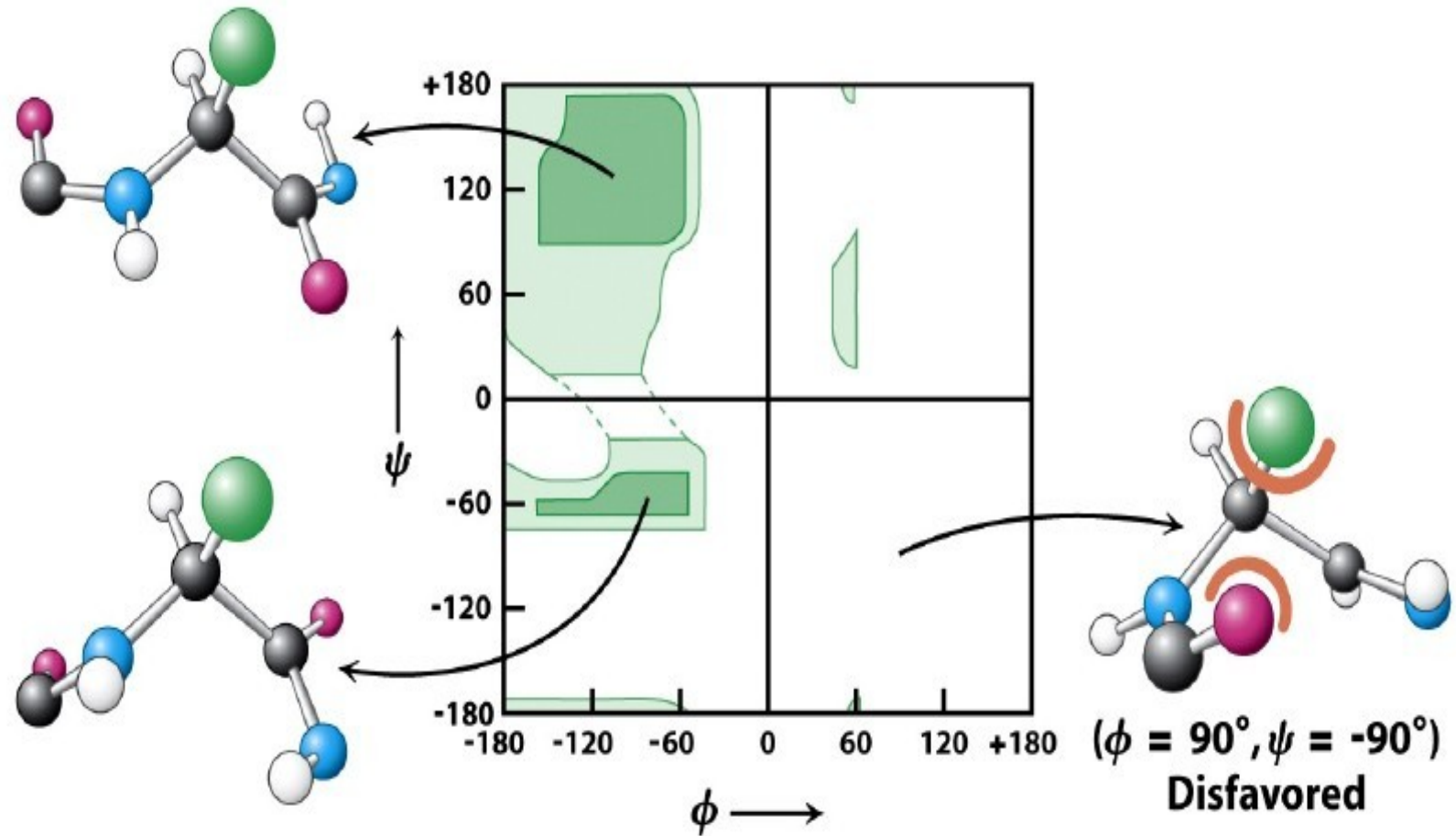
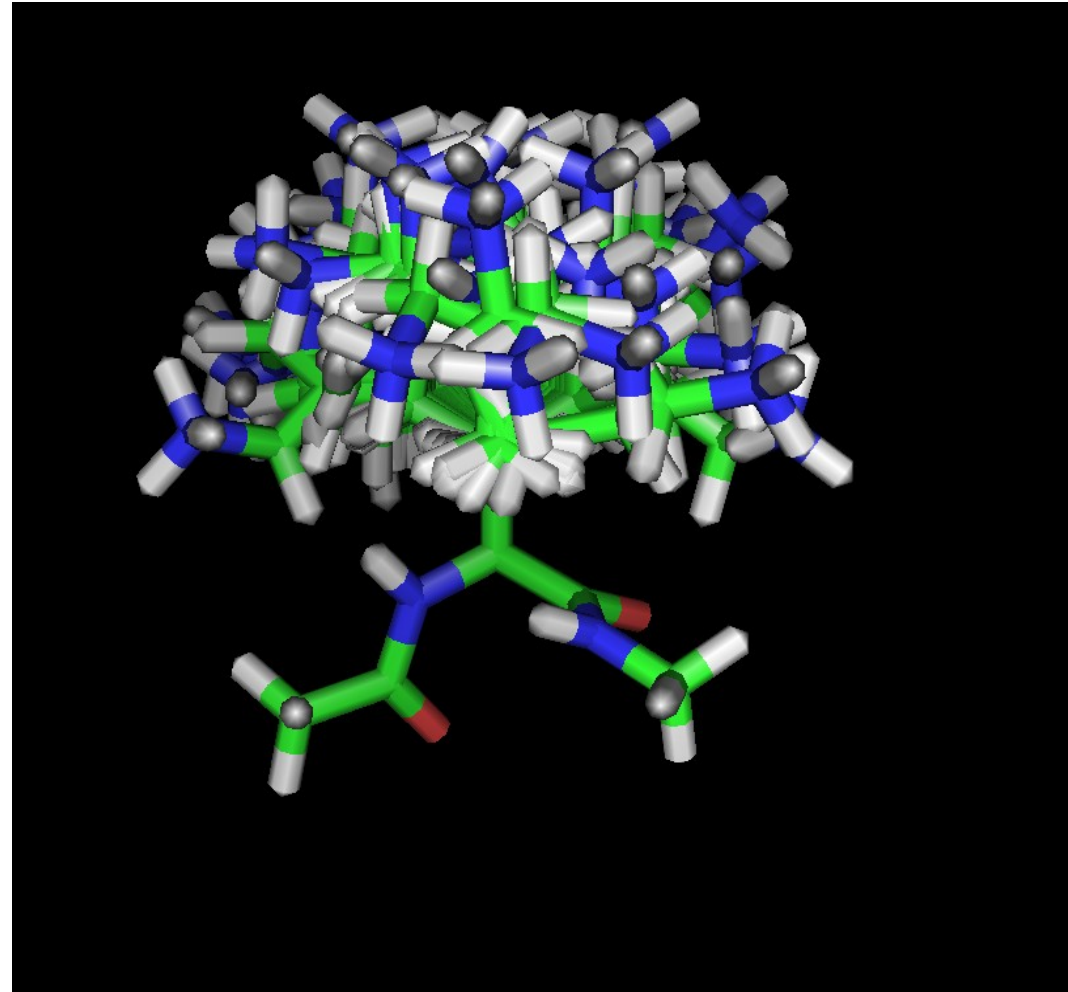
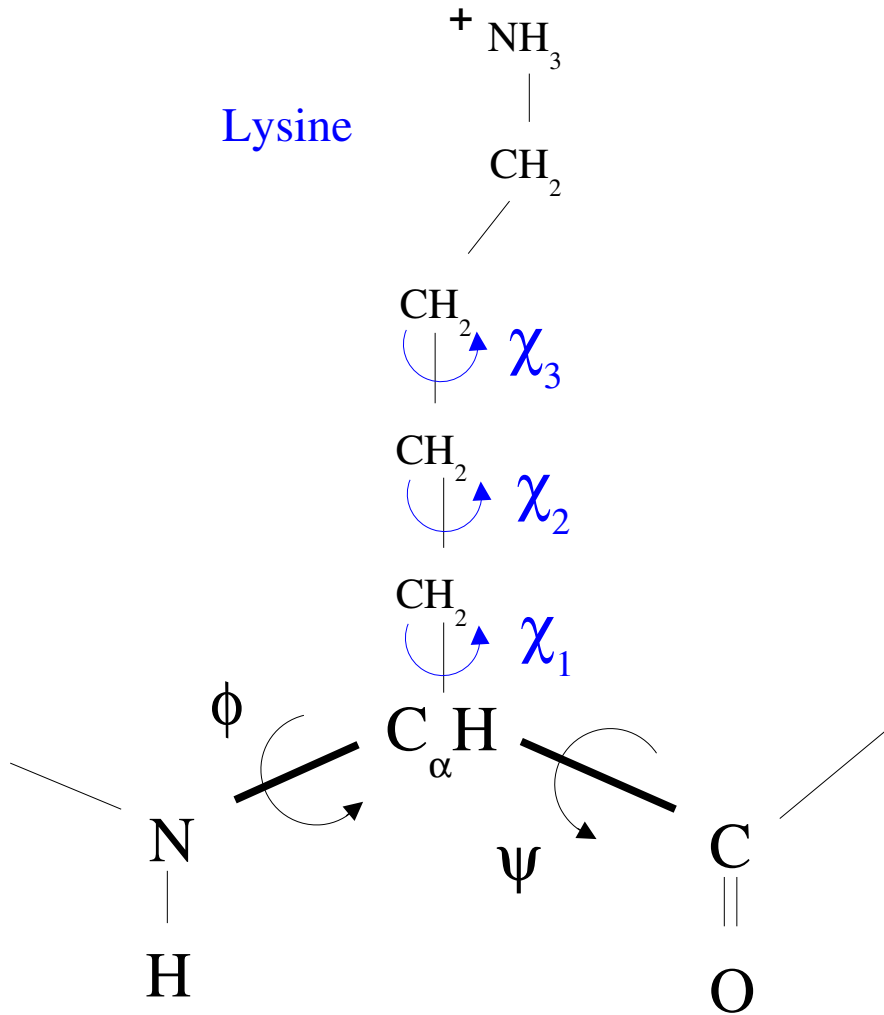


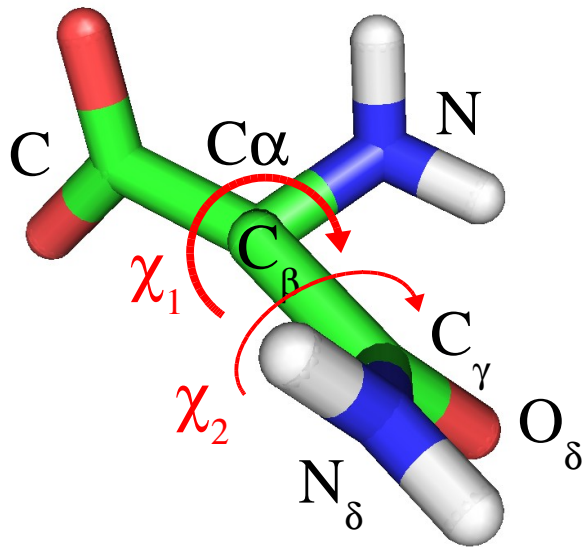
Figure 2-28
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

The side chains also have flexible torsion angles

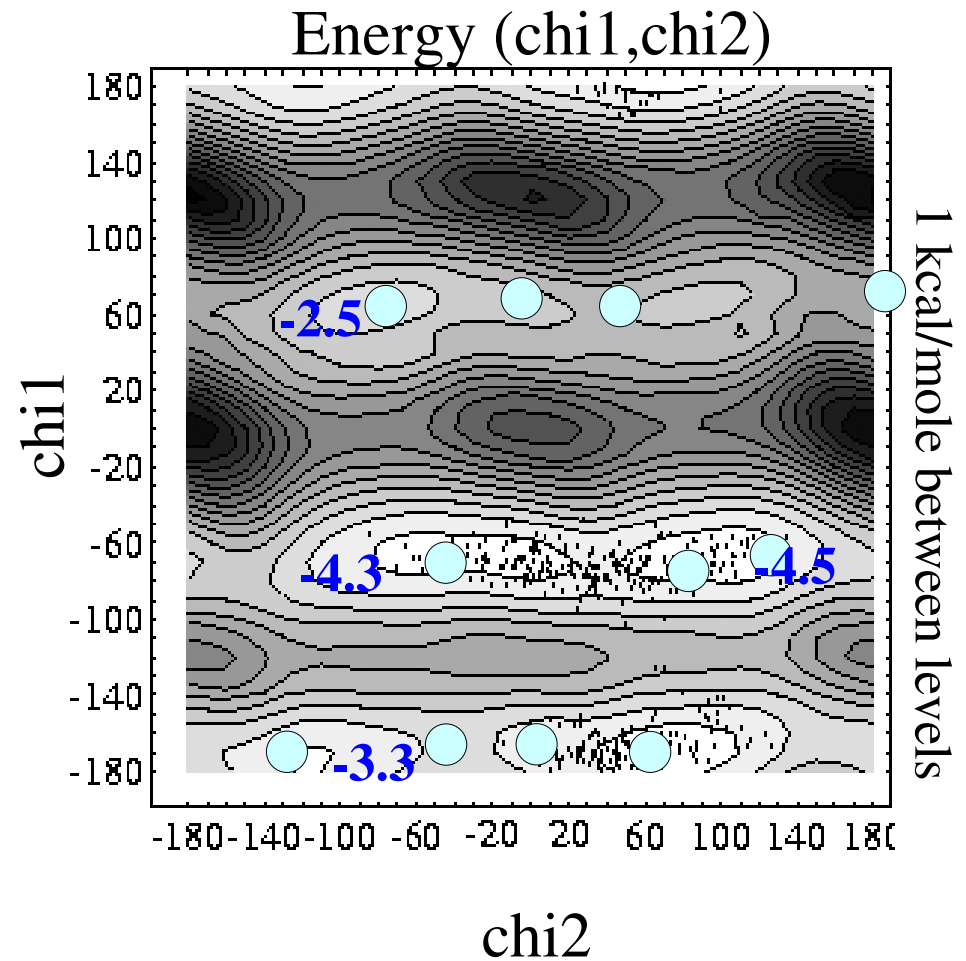
Lysine



The preferred side-chains conformations are called “rotamers”



Example: Asparagine

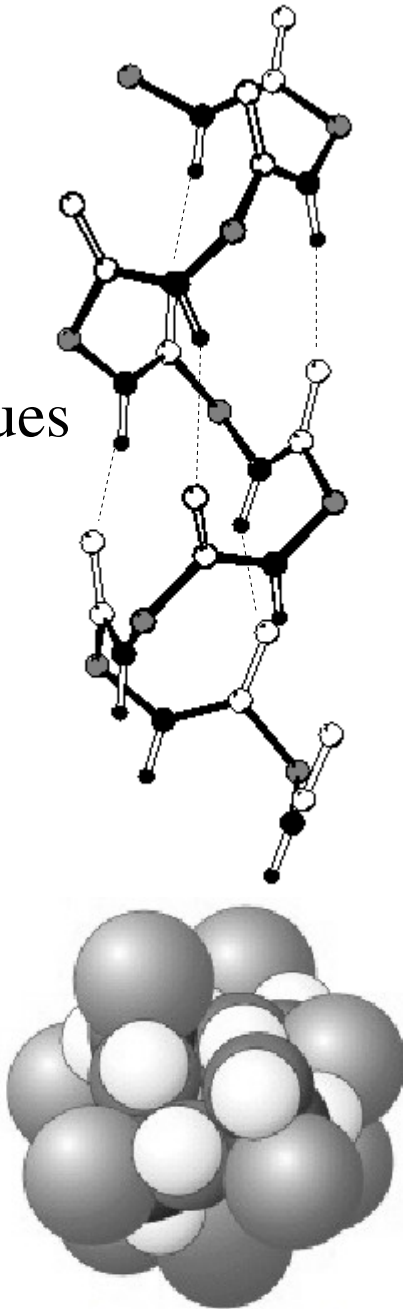


- Typical conformations experimentally observed
- conformations observed by simulation

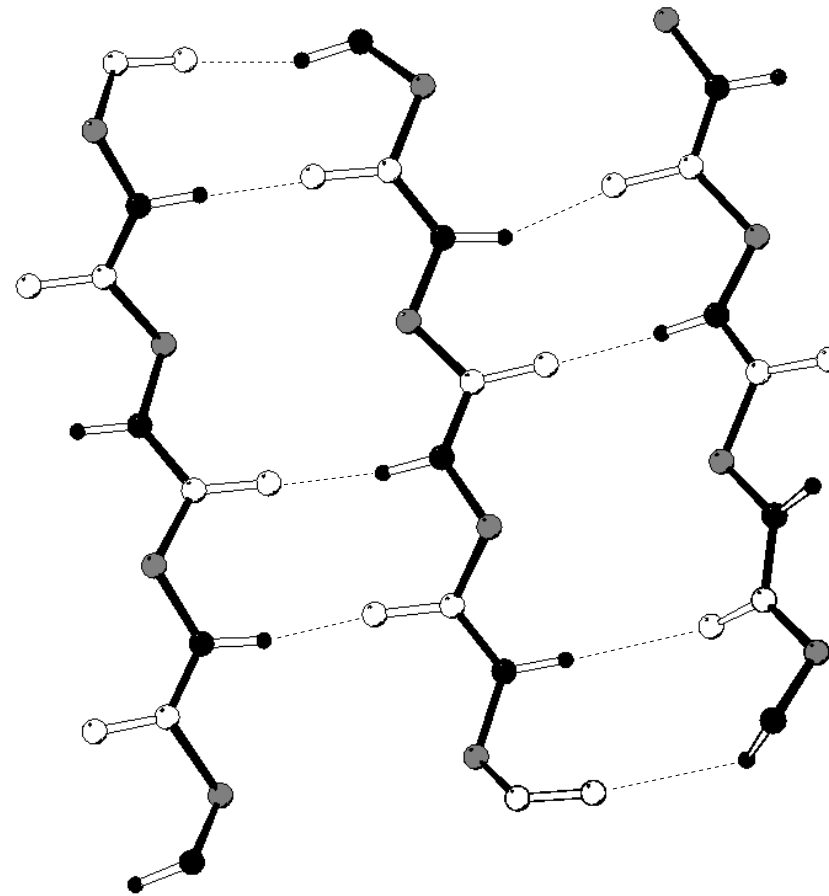
In helices and sheets, polar groups are involved into hydrogen bonds

α helix

3.6 residues
per turn



β -sheet



Pseudo-periodicity of 2

α -helix

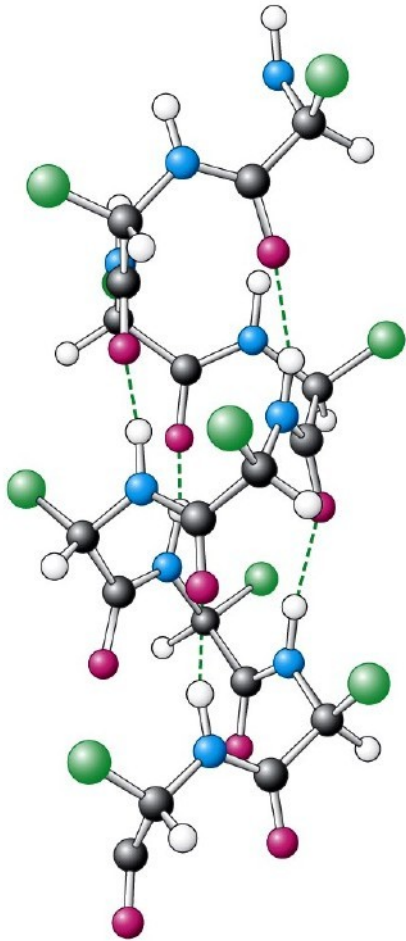


Figure 2-29b
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

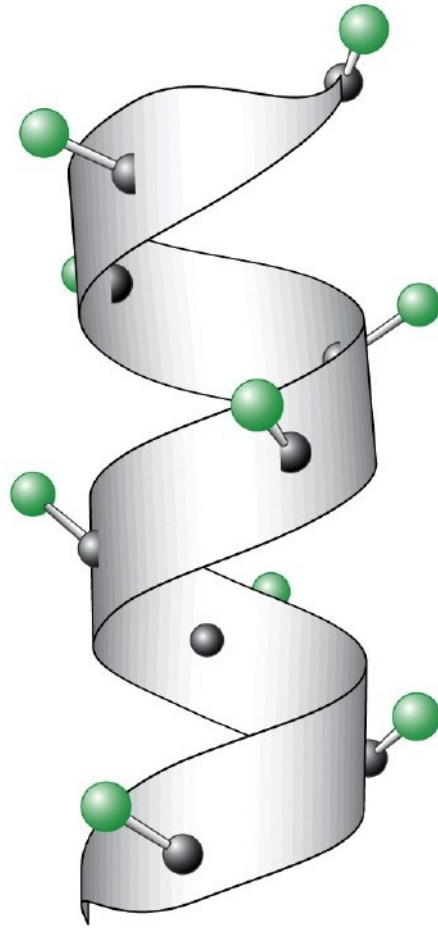


Figure 2-29a
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

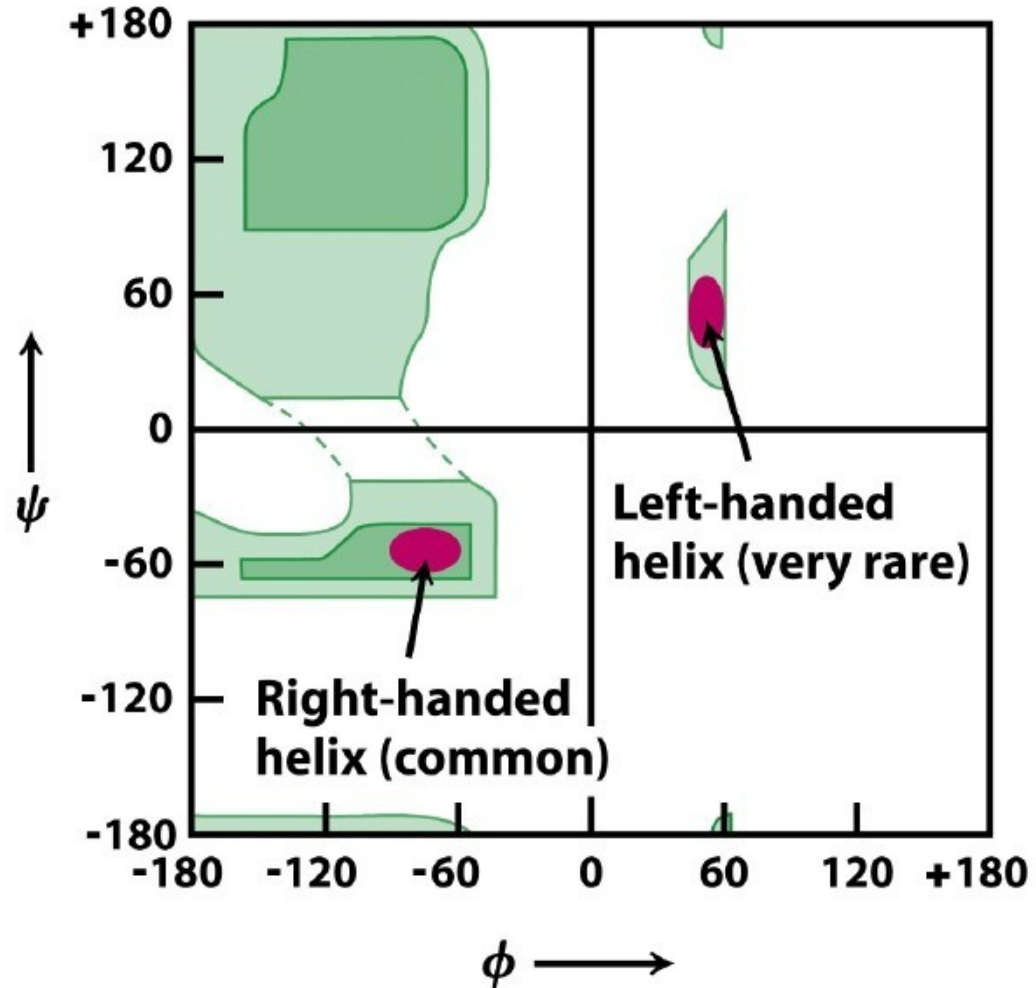


Figure 2-31
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

3.6 residues per turn, H-bond between residue n and n+4

Although other (rare) helices are observed: π -helices, 3.10-helices...

β -sheets

β -strand (elementary blocks) :

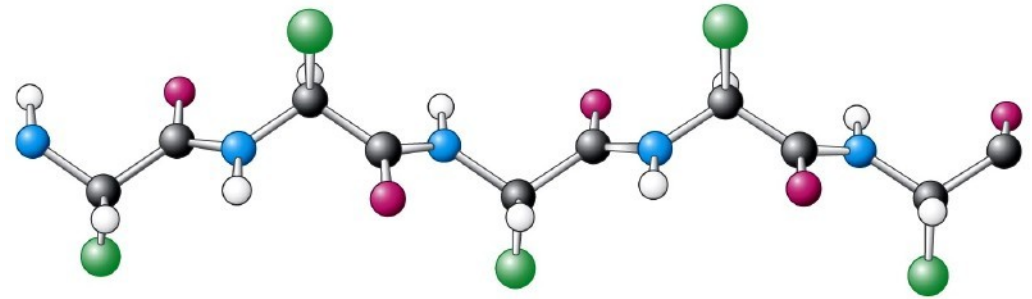
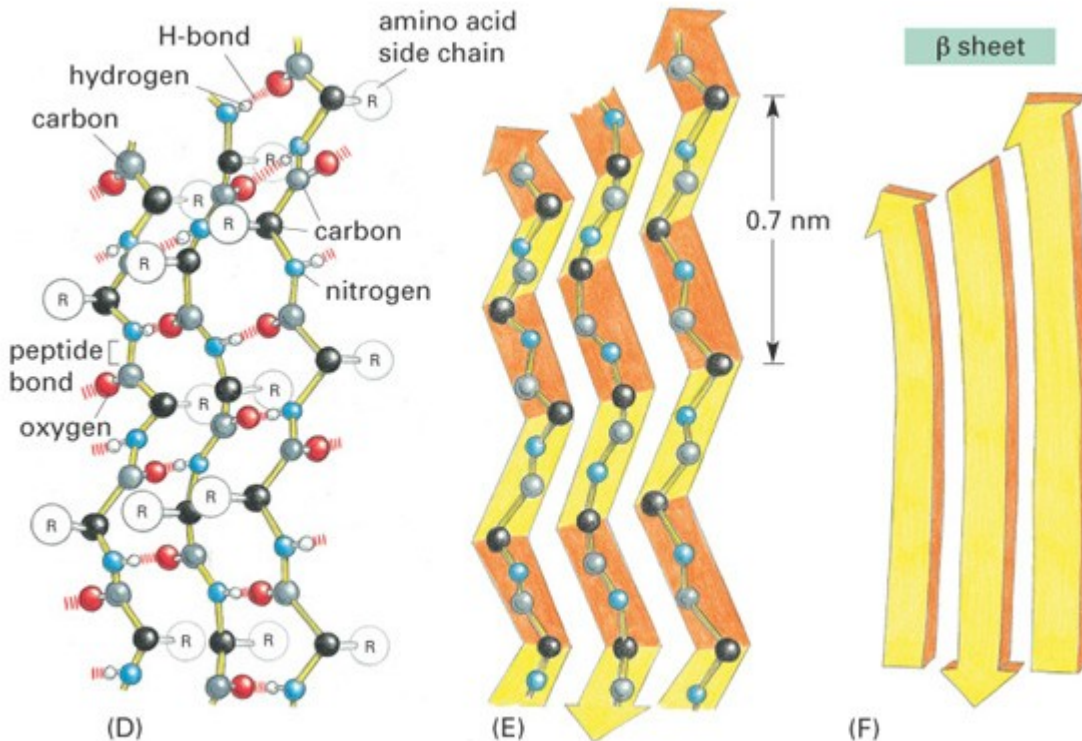


Figure 2-35
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company



β -strands are assembled into (parallel, anti-parallel) β -sheets.

β -sheets

Anti-parallel β -sheets

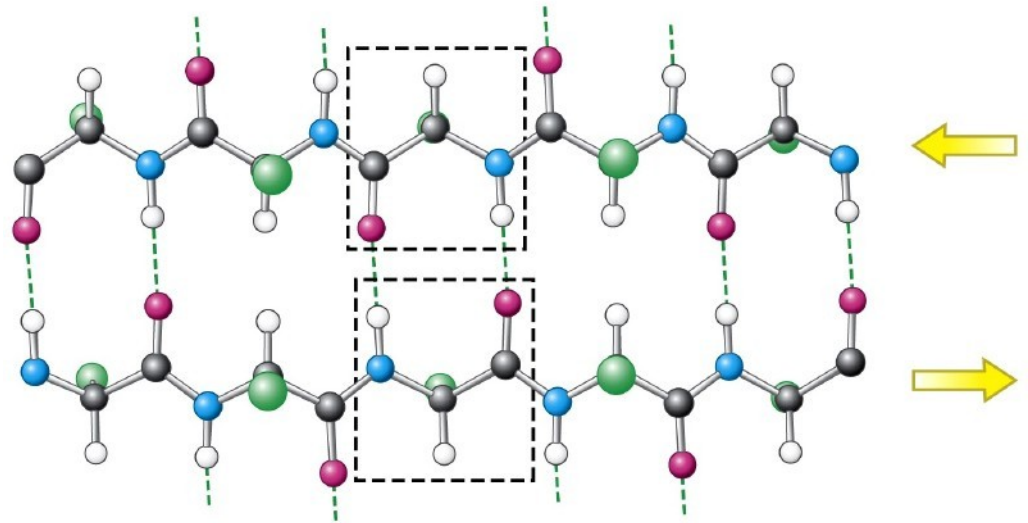
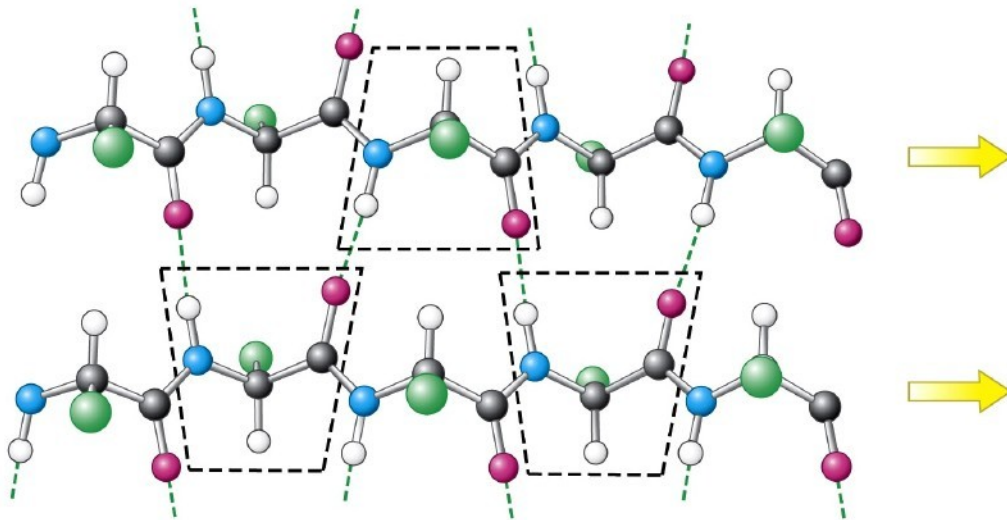


Figure 2-36
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company



Parallel β -sheets

Figure 2-37
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

β -sheets

Various shapes of β structures

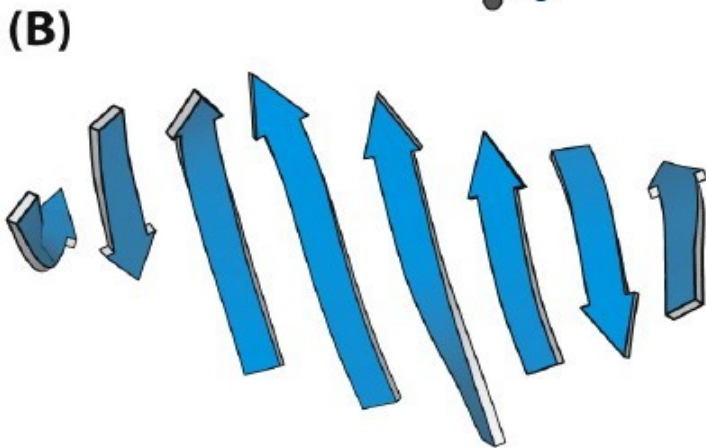
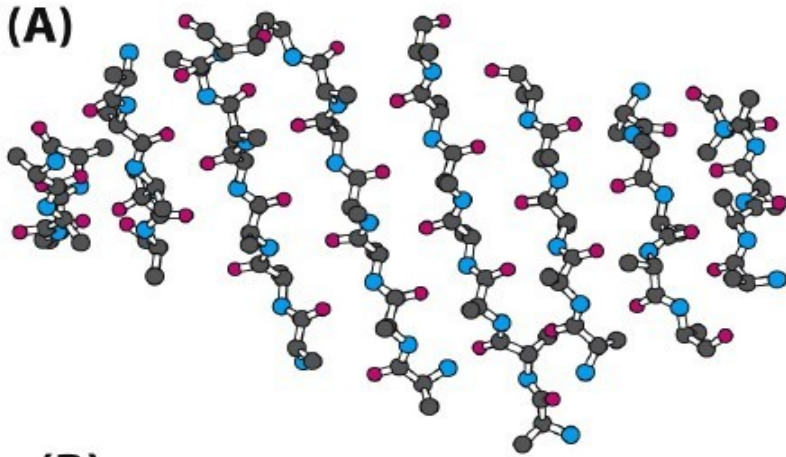


Figure 2-39
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Twisted β -sheets

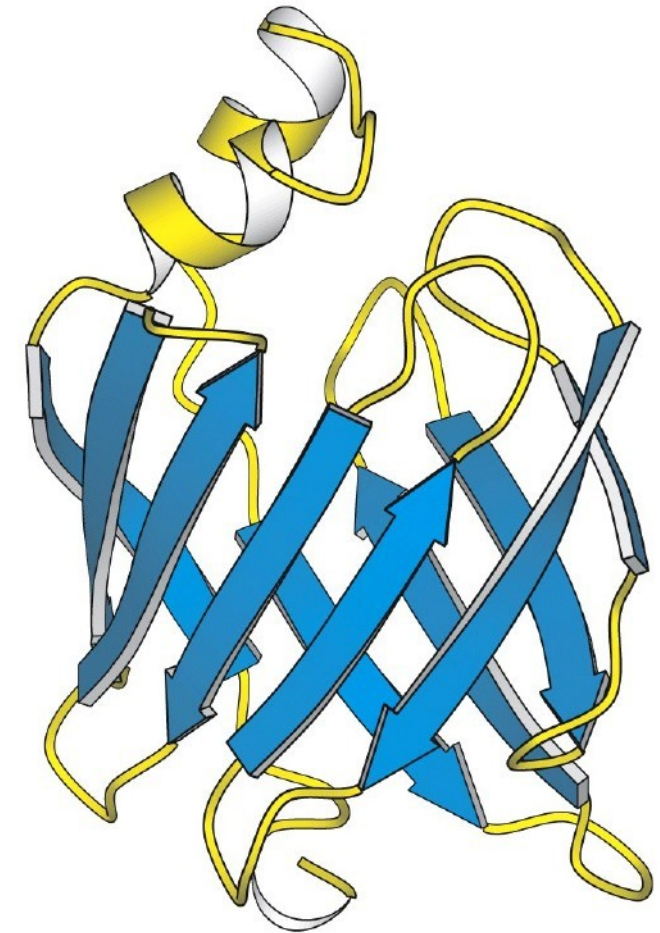
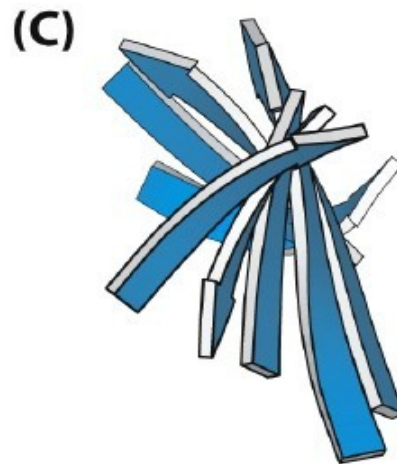
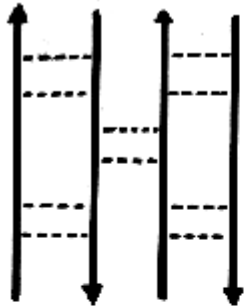


Figure 2-40
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

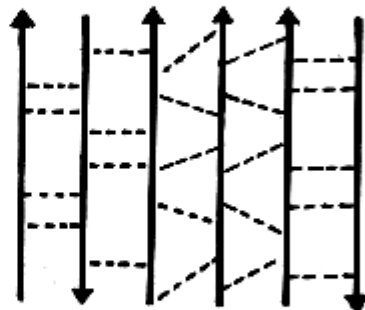
β -barrel

β -sheets

Antiparallel beta-sheet



The different types of beta-sheet. Dashed lines indicate main chain hydrogen bonds.



Mixed beta-sheet

Parallel beta-sheet

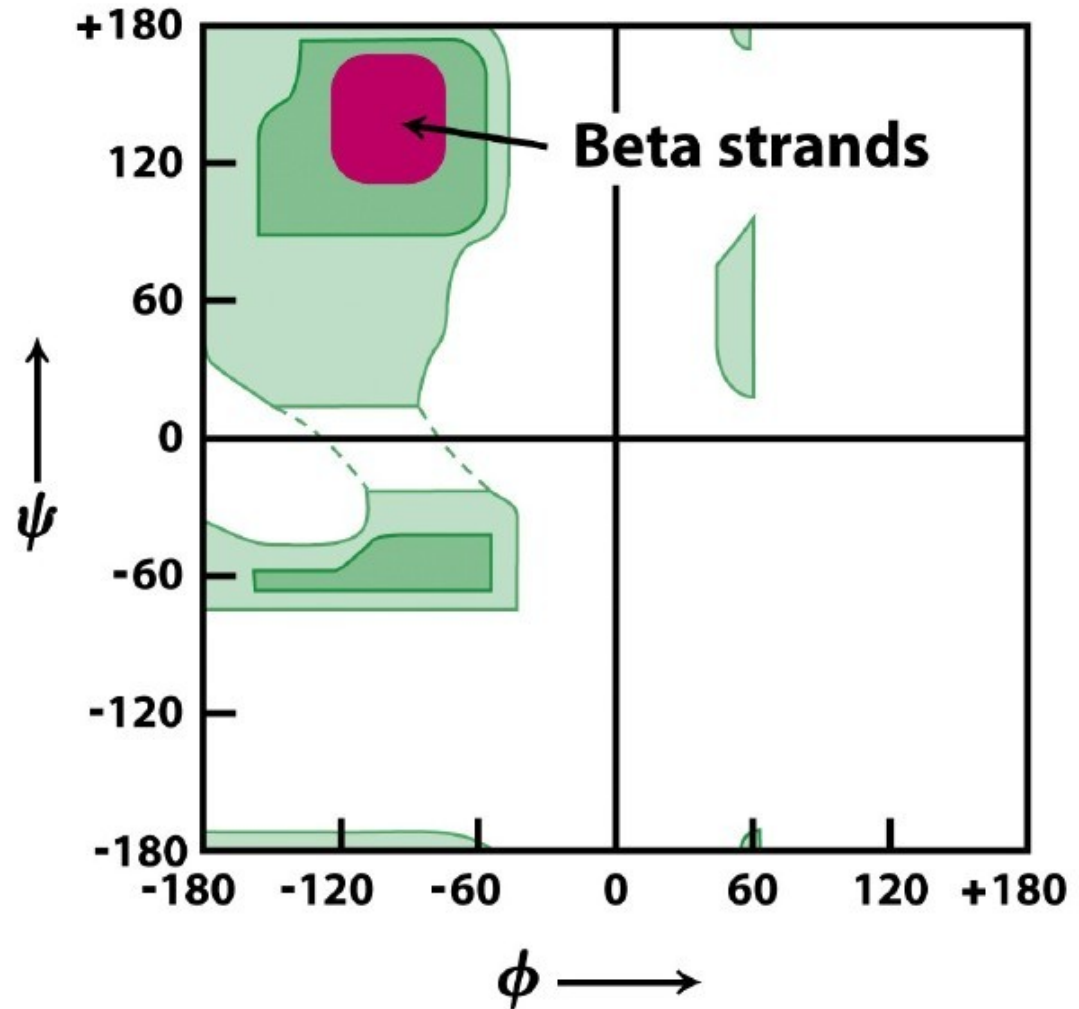
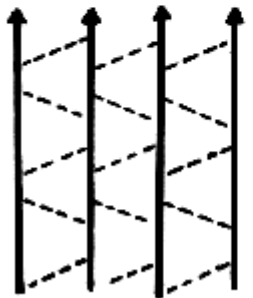
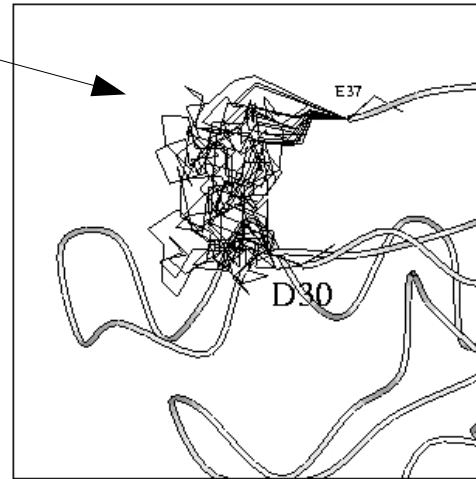
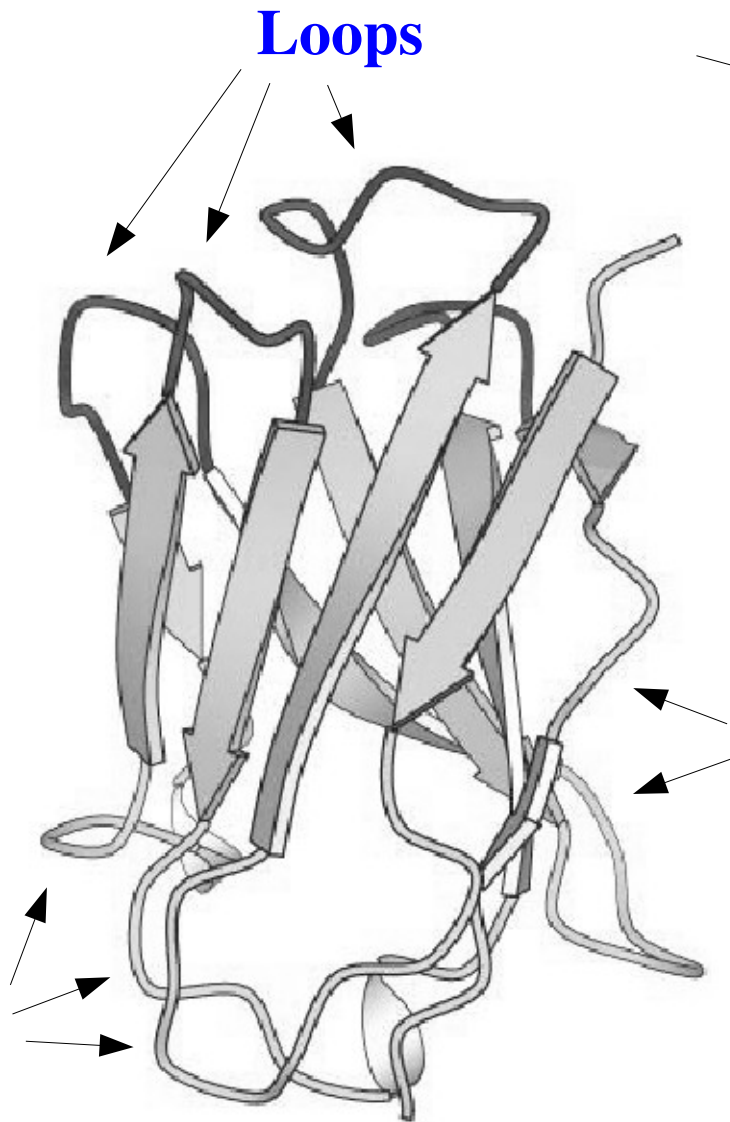
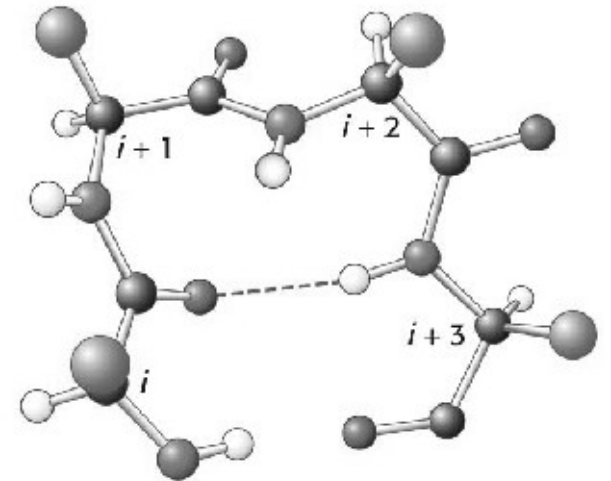


Figure 2-34
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Loops



turn



~ 1/3 of amino acids

Super-secondary & Tertiary structure

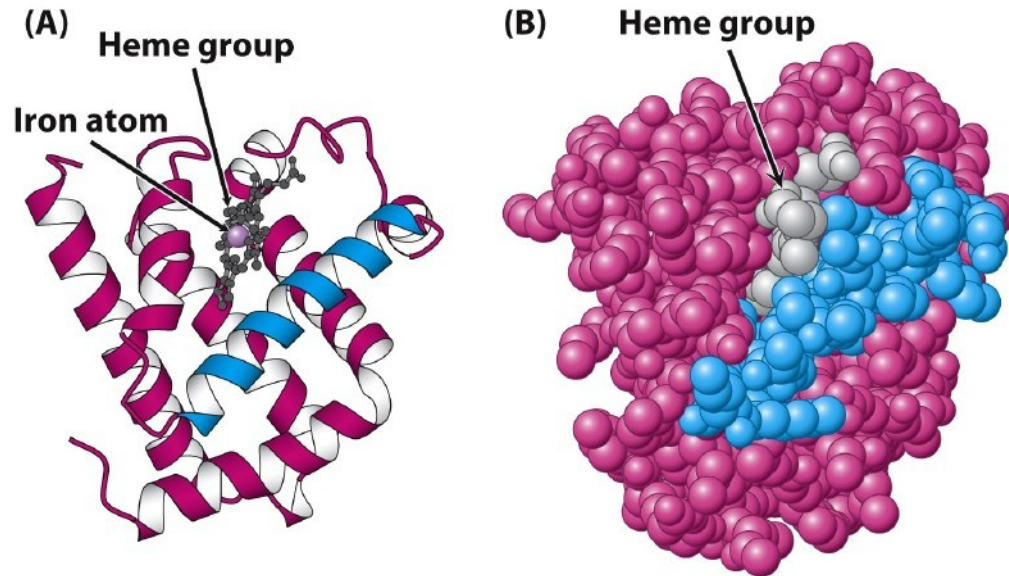


Figure 2-48
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

The tertiary structure is the set of 3D coordinates of atoms of a single amino acid chain

Secondary structure elements can be assembled into super-secondary motifs.

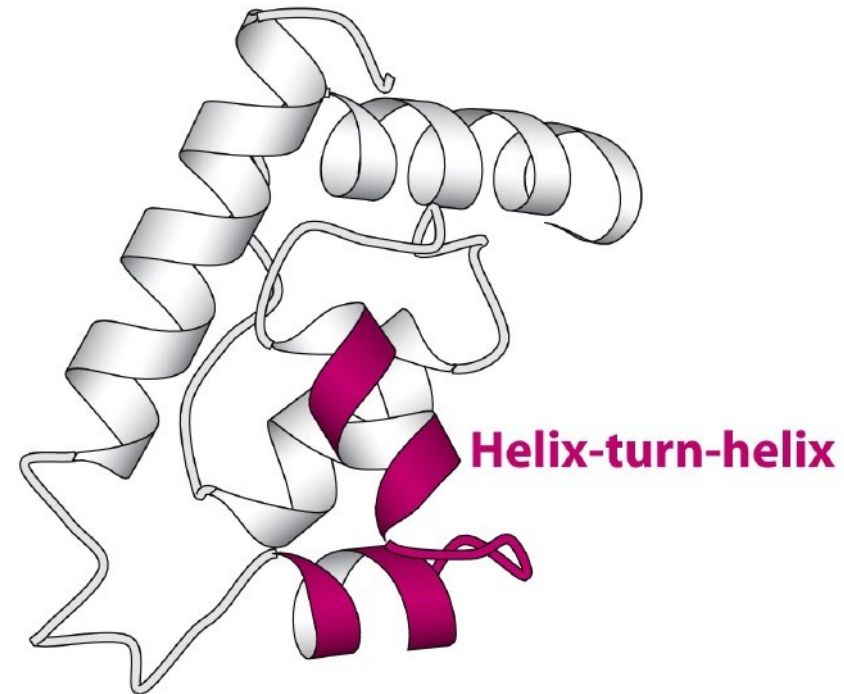


Figure 2-51
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

Quaternary structure

A protein can be composed of multiple chains with interacting subunits.

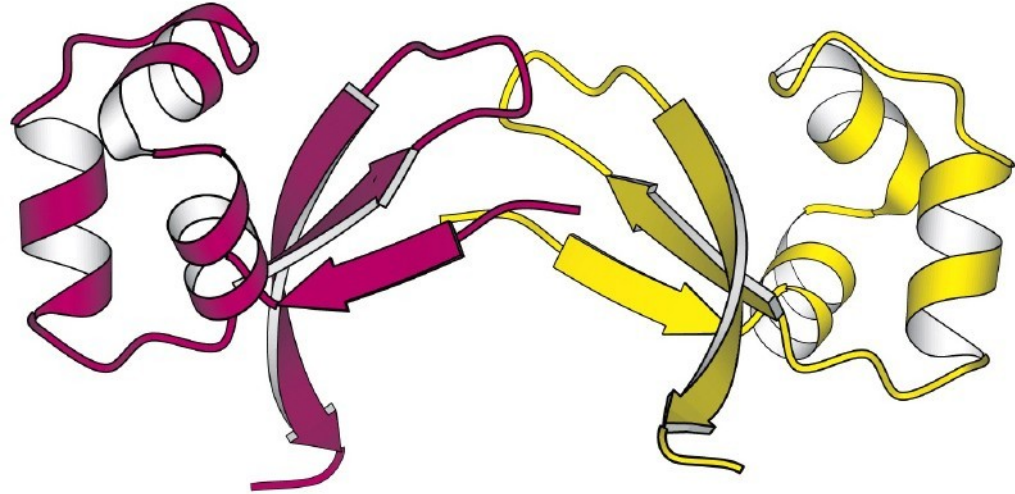


Figure 2-53
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

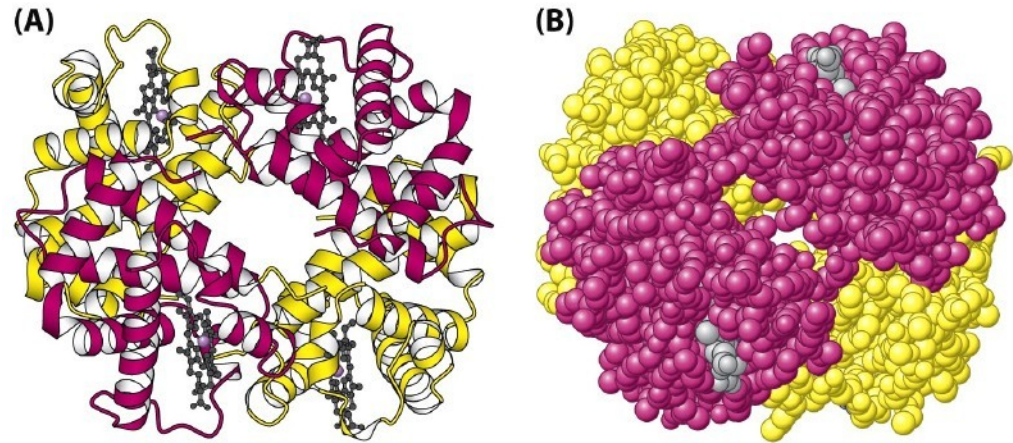


Figure 2-54
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Protein can interact with molecules

Example: Hemoglobin

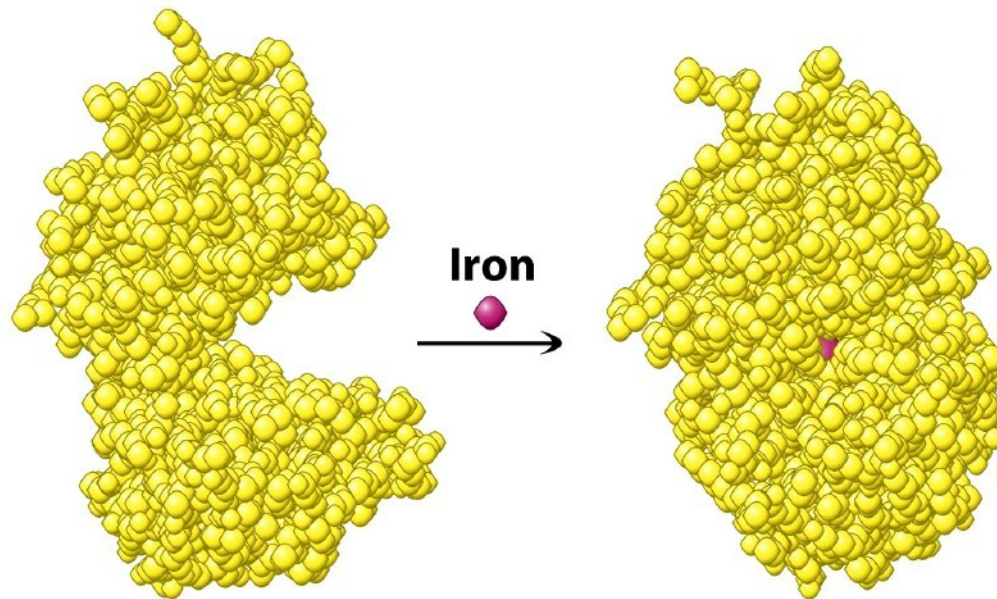


Figure 2-3
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

An Heme (iron + organic ring) binds to the protein, and allow the capture of oxygen atoms.

Disulfide bond

Two cysteines can interact and create a disulfide bond.

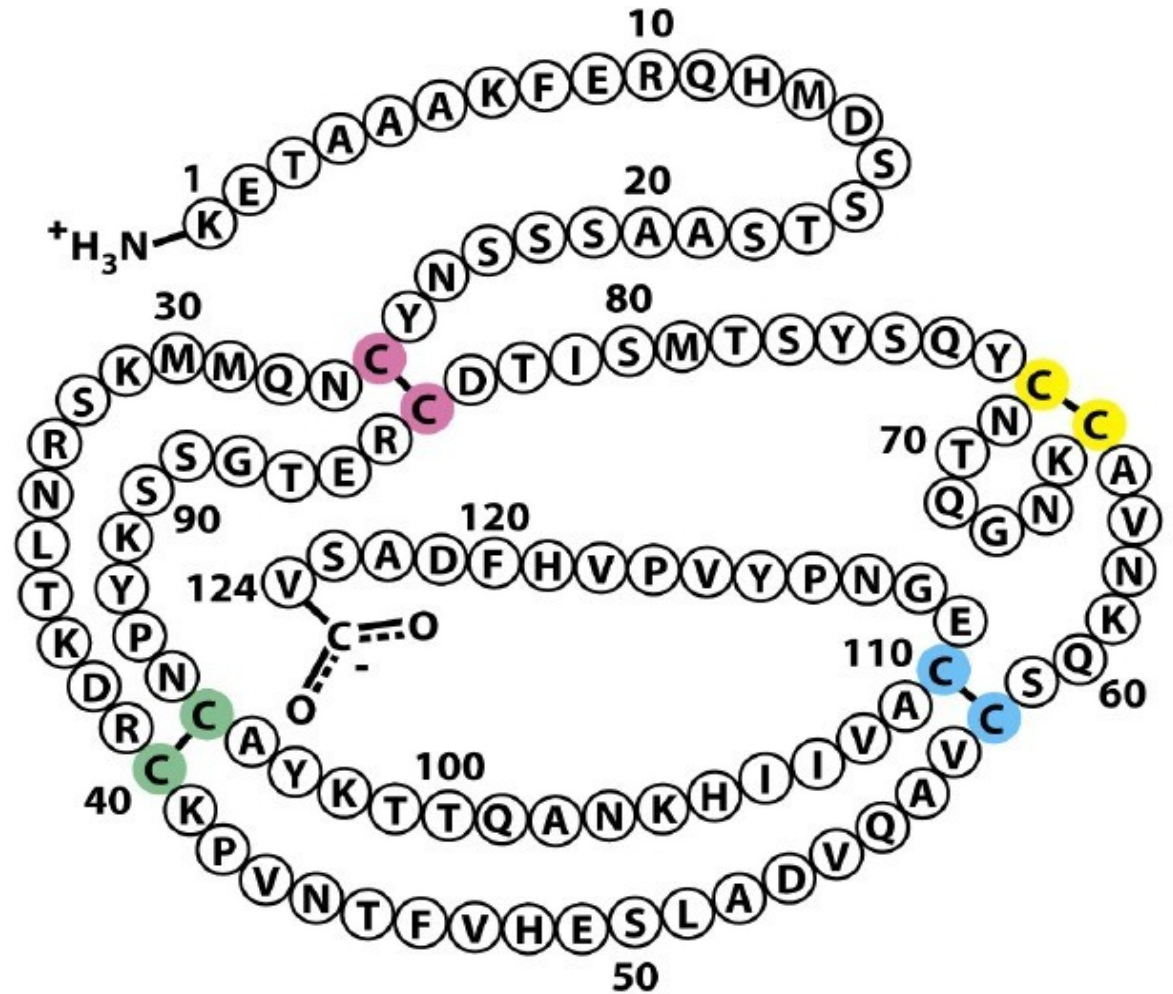
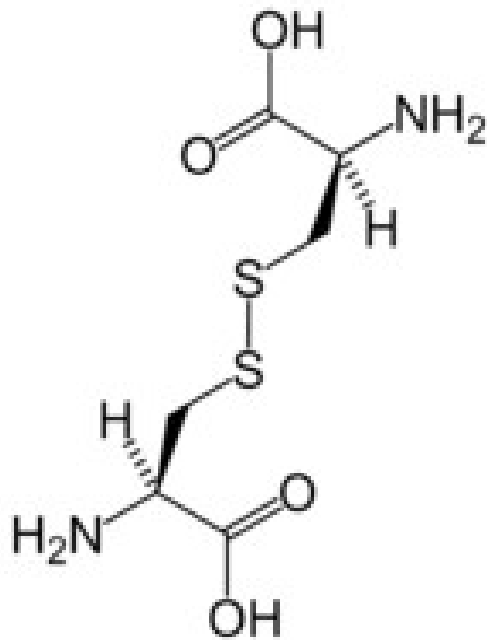
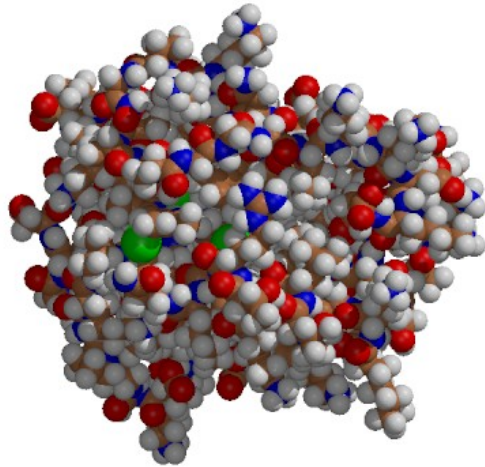


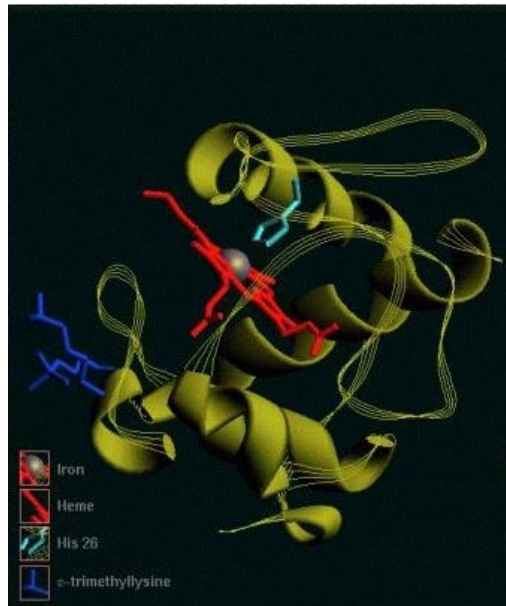
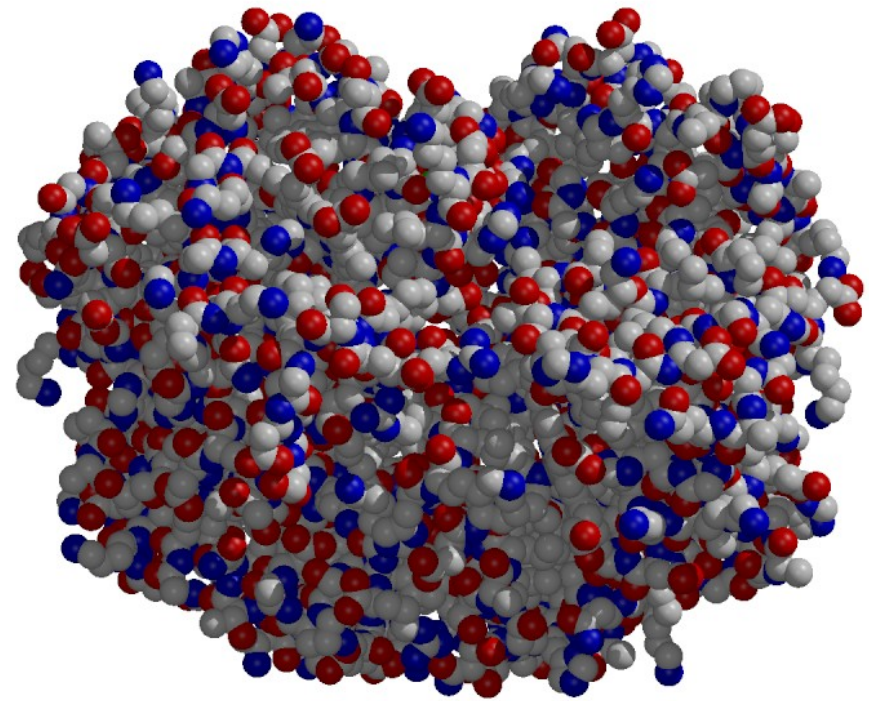
Figure 2-56
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

The tertiary structure is globular, with a preference for polar residues on its surface but rather apolar in its interior

Cytochrom c

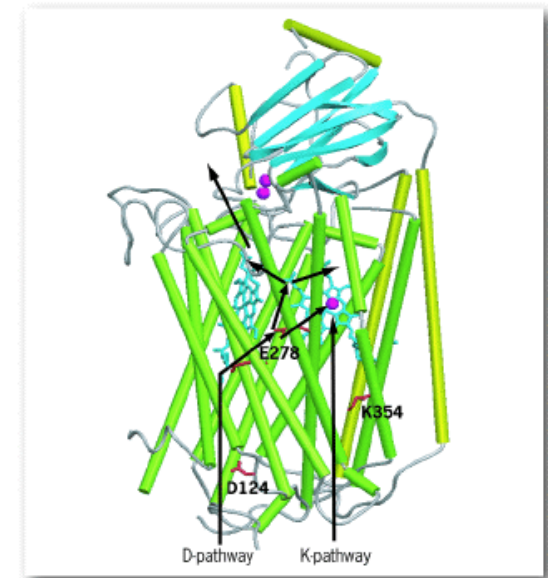
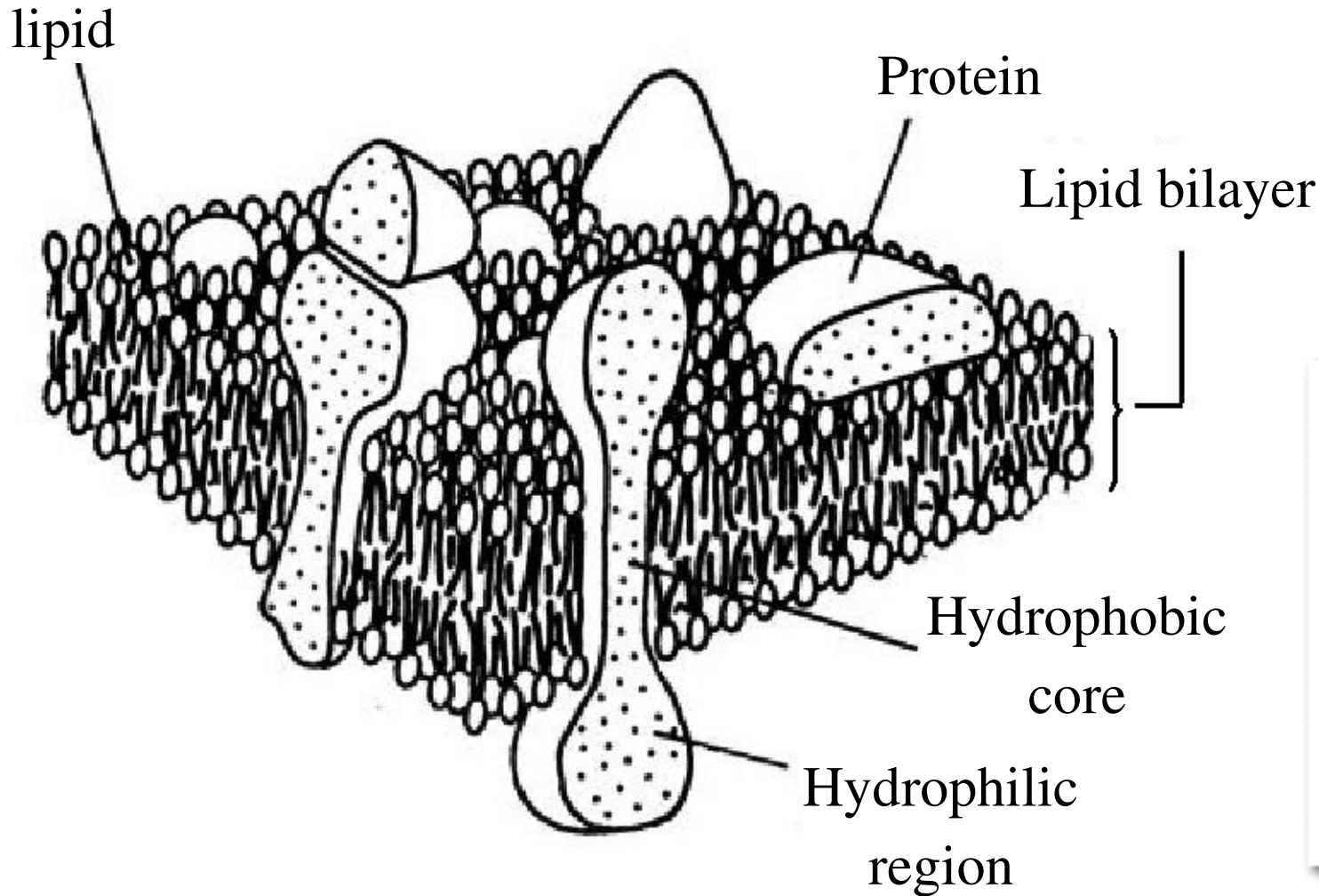


Hemoglobine



water

Membrane proteins are an exception



Cytochrom oxidase

~ 30% of human genome, ~ 50% of antibiotics

Proteins folds into a native structure

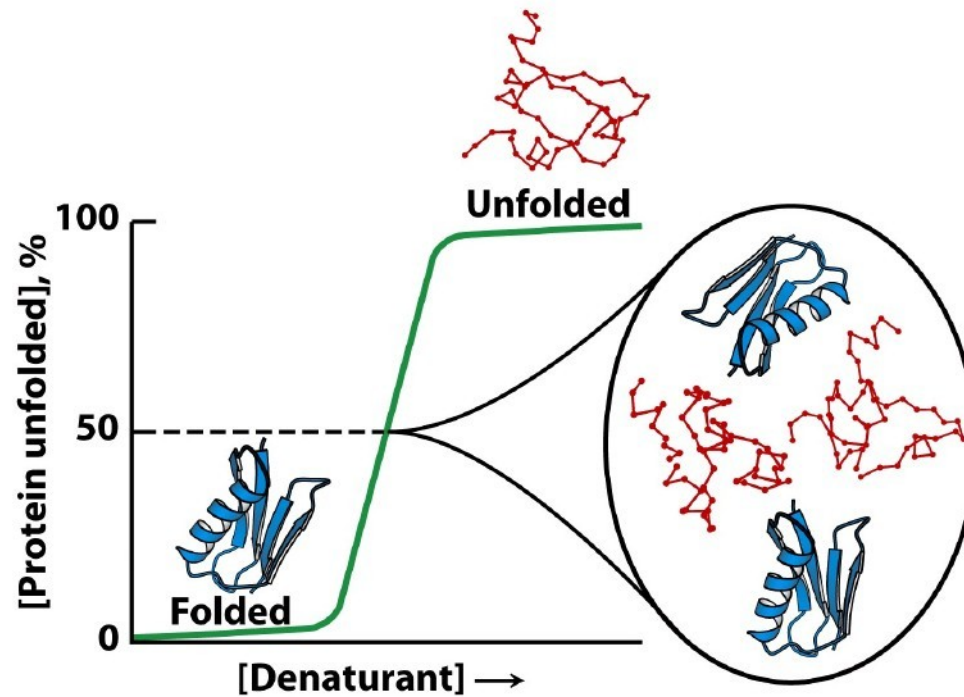


Figure 2-64
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

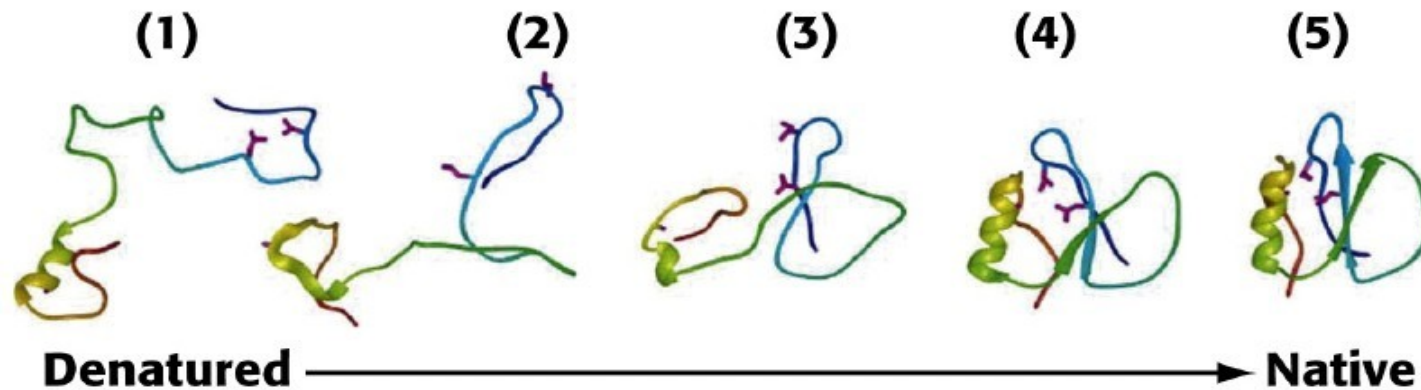


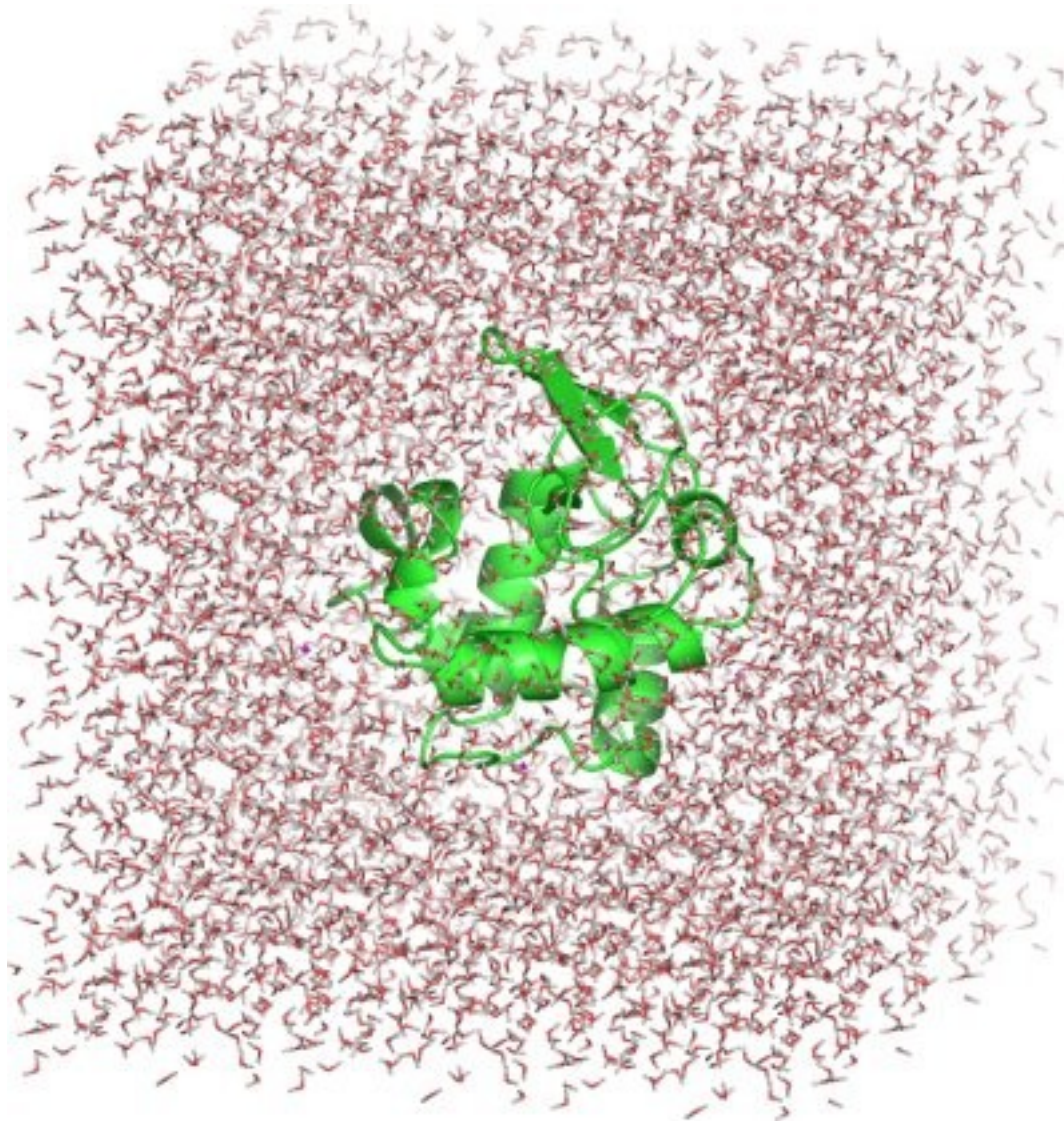
Figure 2-66
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Overview of the methods used to predict the protein structure

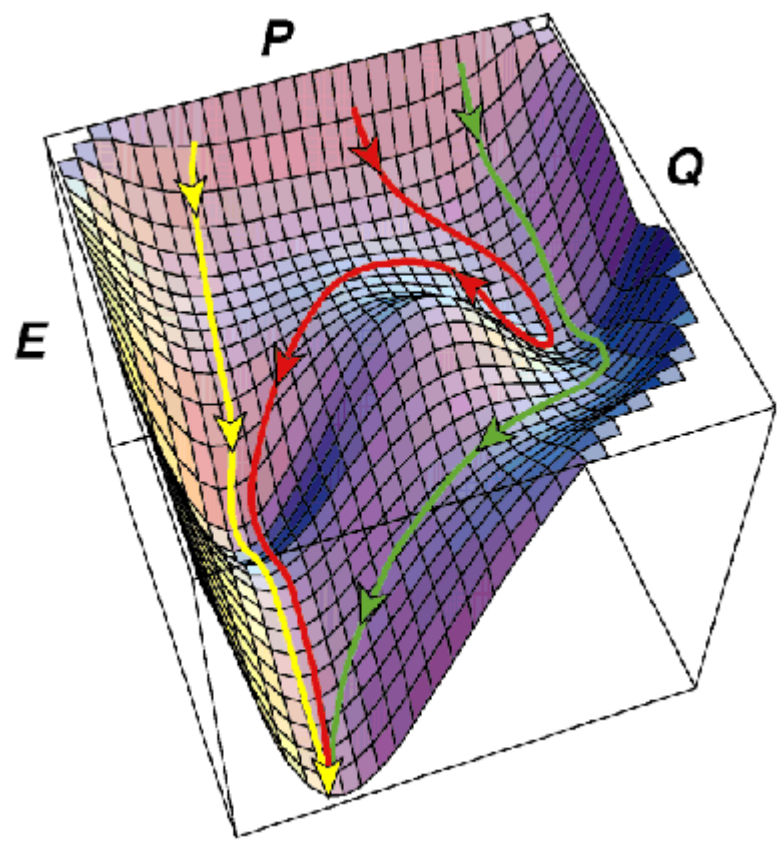
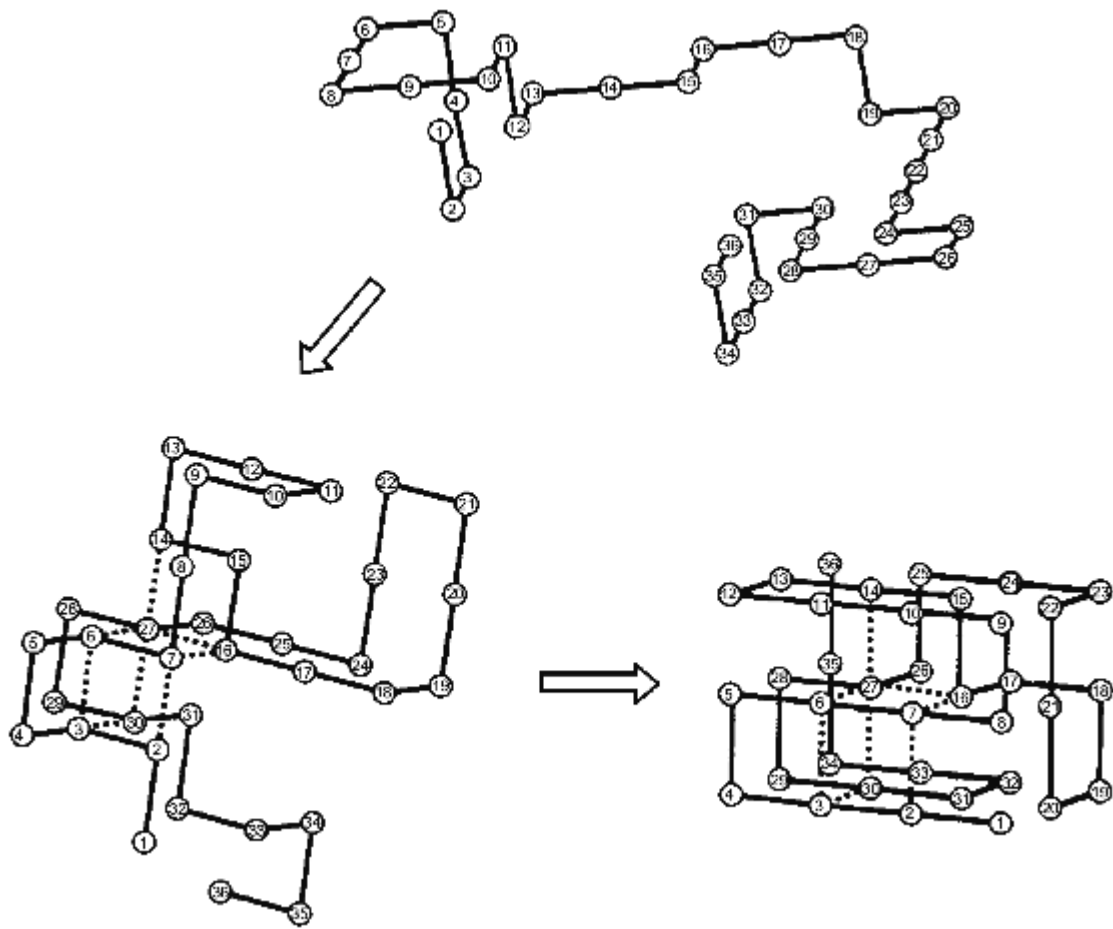
Several issue must be addressed first:

- Which degree of definition?
- What's the length of the sequence?
- Which representation/modeling suits the best?
- Should we simulate the folding or predict the structure?
- Do we want a single prediction or a set of candidates?
- Machine learning approach or physical model?

Molecular Dynamics

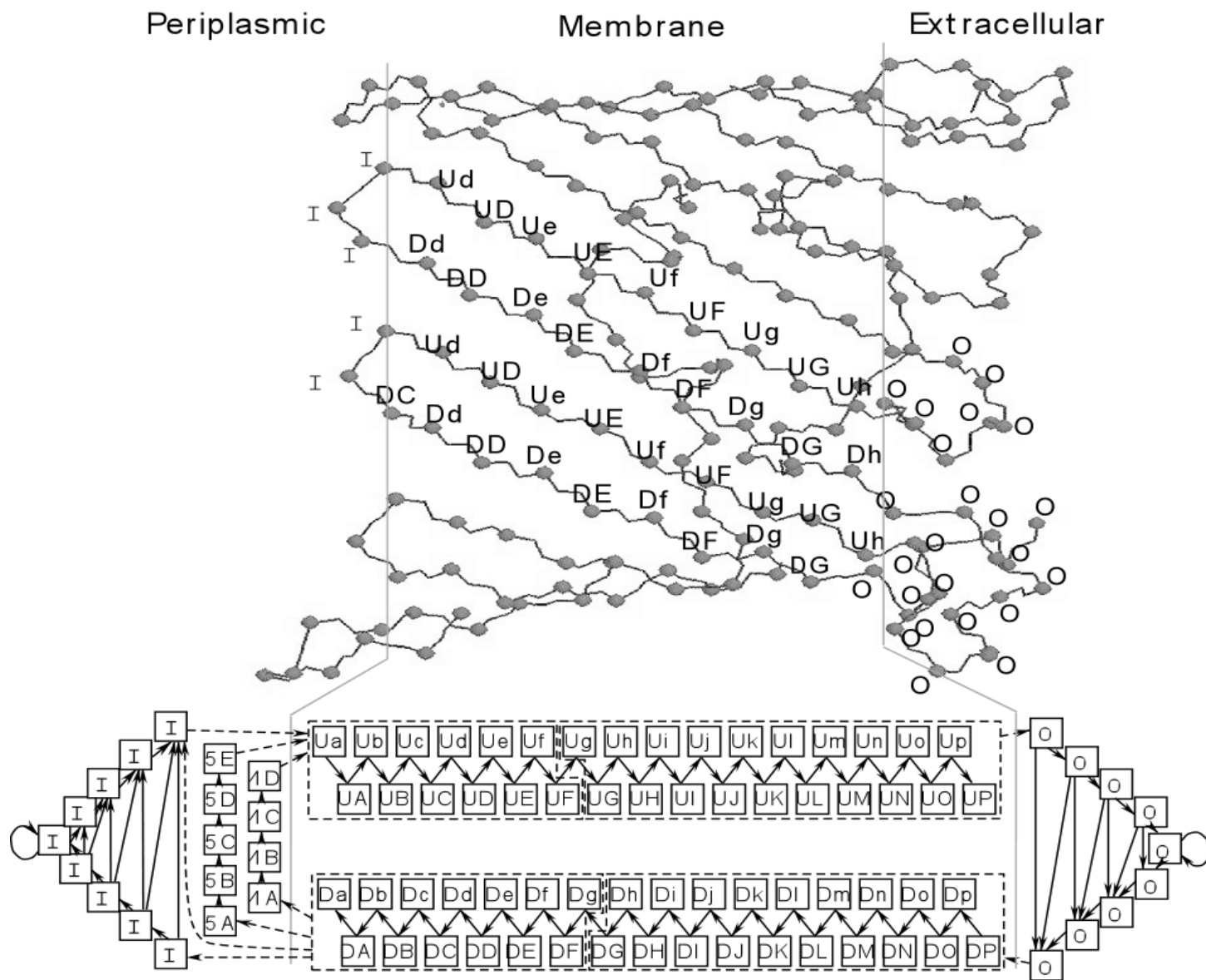


HP lattice model

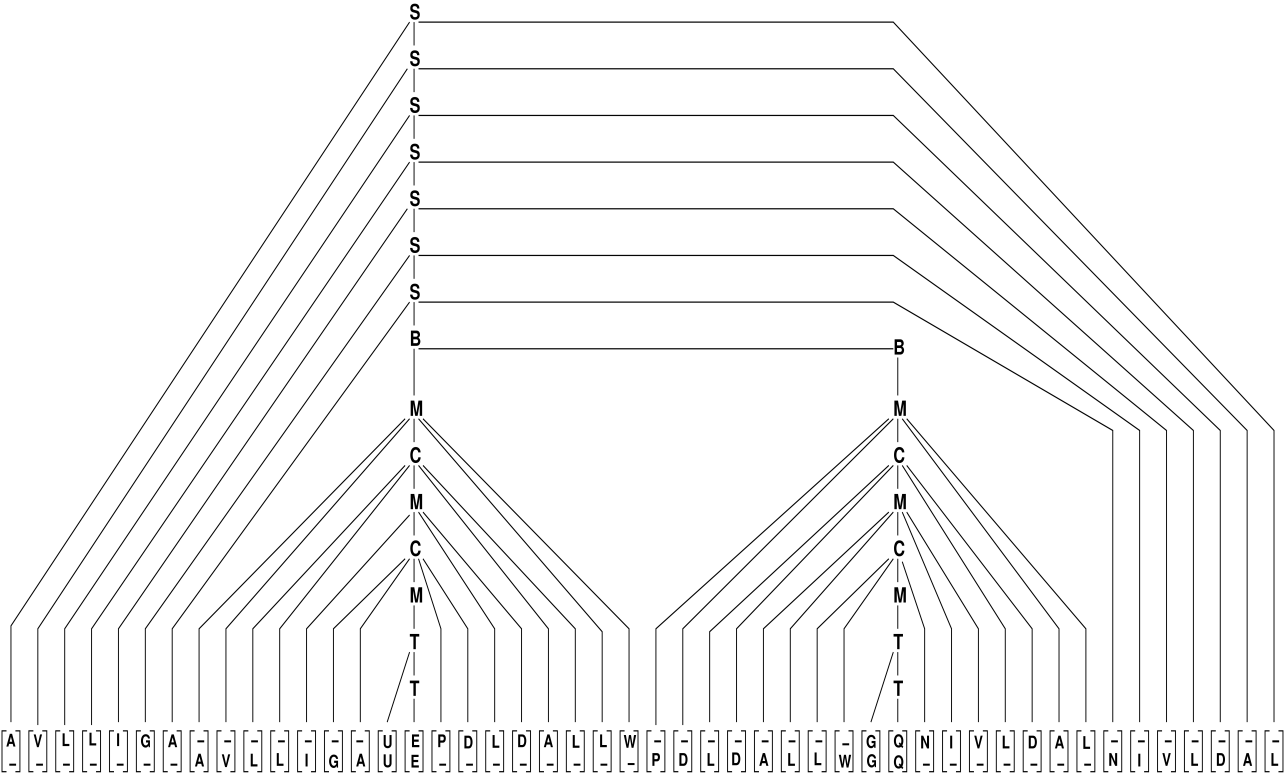


Hidden Markov models

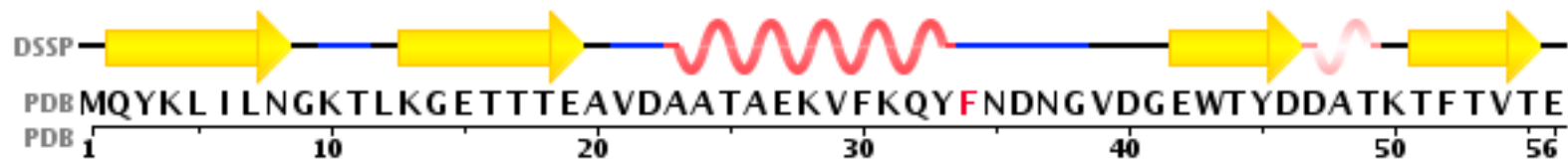
(and other machine learning approaches)








Structural template methods



Protein Secondary Structure



-  E: beta strand
-  empty: no secondary structure assigned
-  G: 3/10-helix
-  S: bend
-  H: alpha helix

Protein Secondary Structure Prediction Using Statistical Models

- Sequences determine structures
- Proteins fold into minimum energy state.
- Structures are more conserved than sequences. Two proteins with 30% identity likely share the same fold.

How to evaluate a prediction?

In 2D: The Q_3 test.

$$Q_3 = \frac{\text{correctly predicted residues}}{\text{number of residues}}$$

In 3D: The Root Mean Square Deviation (RMSD)

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2}$$


Old methods

- First generation – single residue statistics

Fasman & Chou (1974) :

Some residues have particular secondary structure preference.

Examples: Glu  α -Helix

Val  β -strand

- Second generation – segment statistics

Similar, but also considering adjacent residues.

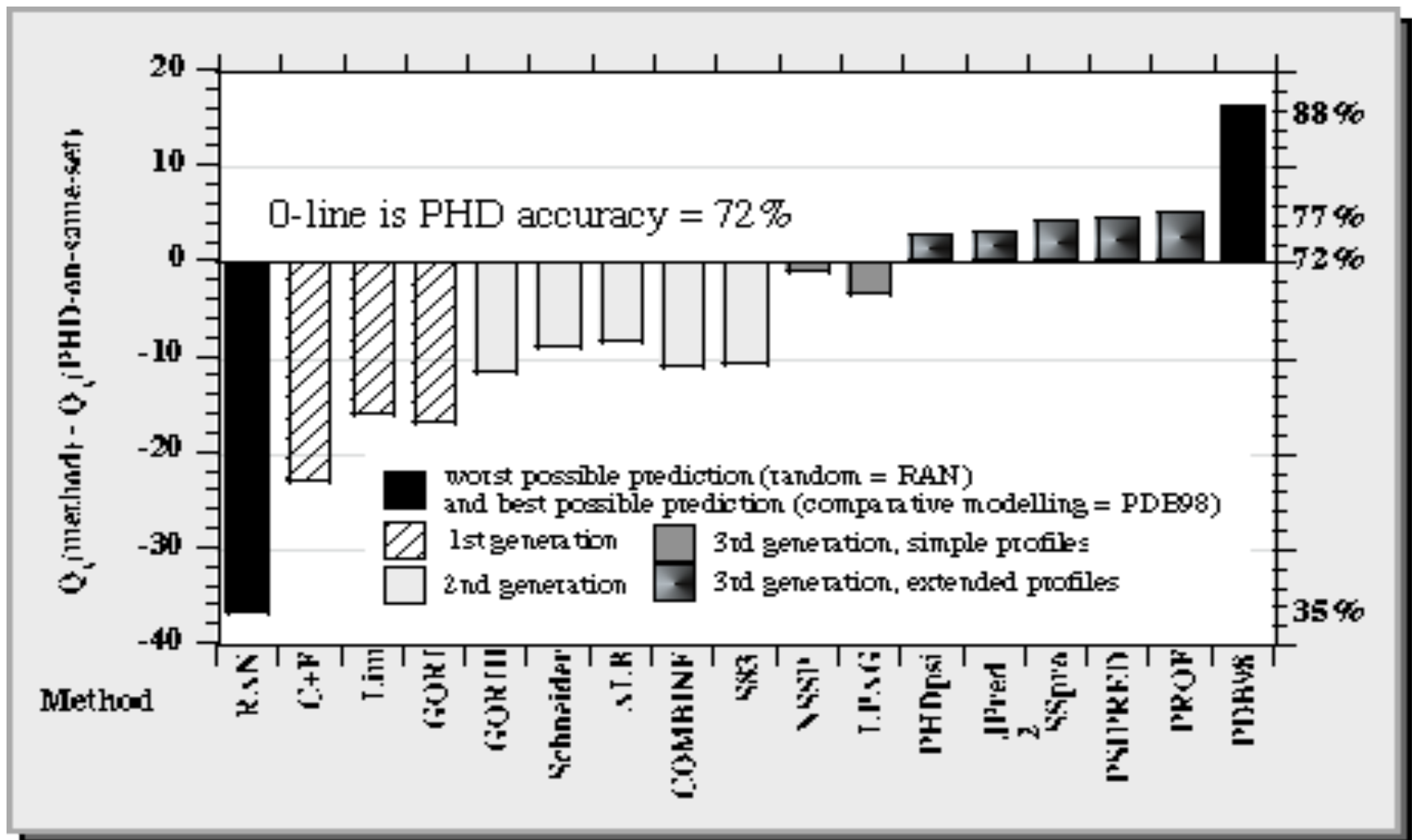
Difficulties

Bad accuracy - below 66% (Q3 results).

Q3 of strands (E) : 28% - 48%.

Predicted structures were too short.


Methods Accuracy Comparison



3rd generation methods

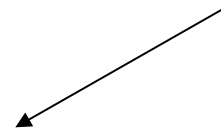
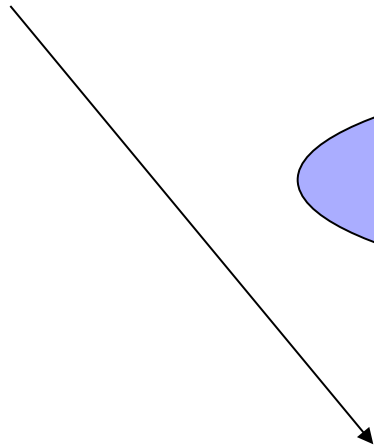
- Third generation methods reached 77% accuracy.
- They consist of two new ideas:
 1. **A biological idea** –
Using evolutionary information.
 2. **A technological idea** –
Using neural networks.

How can evolutionary information help us?

Homologues  similar structure

But sequences change up to 85%

Sequence would vary differently - depends on structure



How can evolutionary information help us?

Where can we find high sequence conservation?

Some examples:

- ➡ In defined secondary structures.
- ➡ In protein core's segments (more hydrophobic).
- ➡ In amphipatic helices (cycle of hydrophobic and hydrophilic residues).

How can evolutionary information help us?

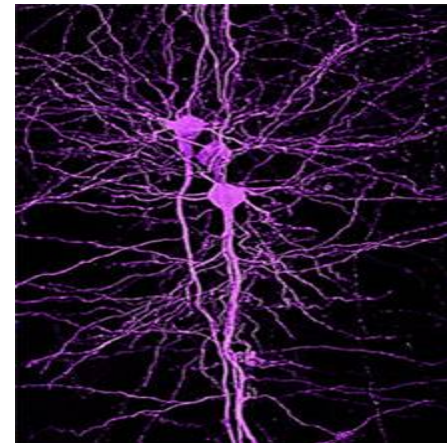
- Predictions based on multiple alignments were made manually.

Problem:

- There isn't any well defined algorithm!

Solution:

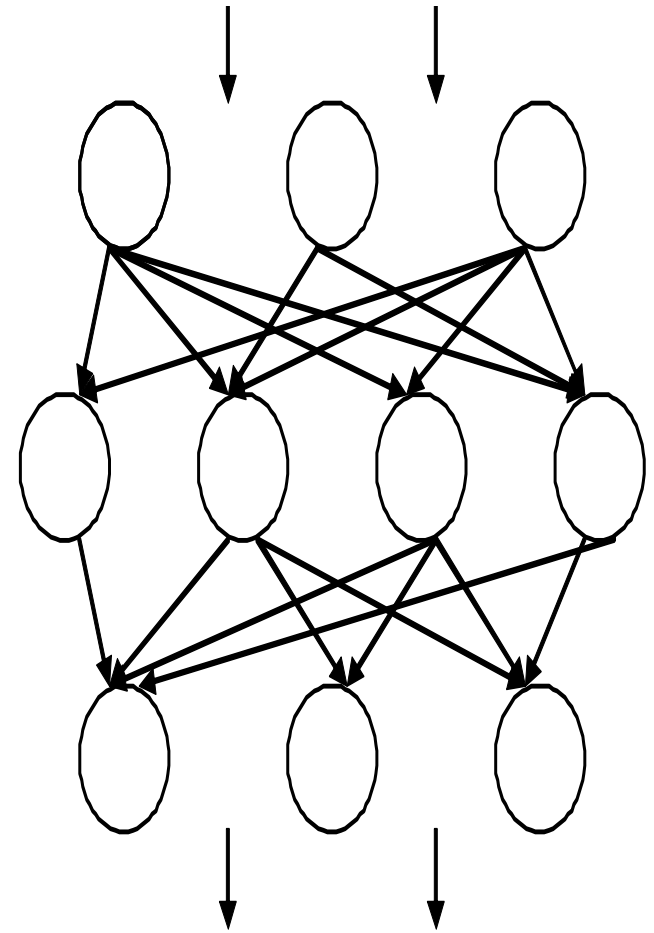
- Use **Neural Networks** .



Artificial Neural Network

The neural network basic structure :

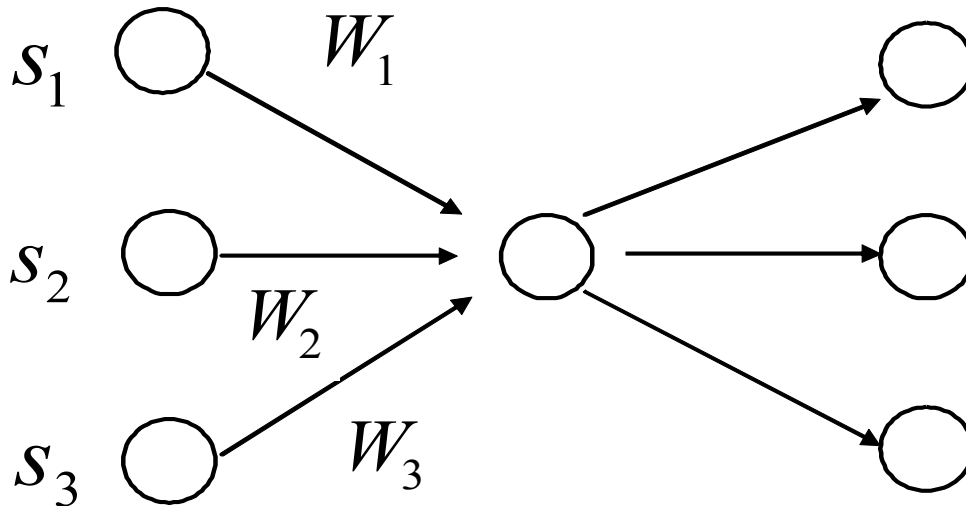
- Big amount of processors – “neurons” .
- Highly connected.
- Working together.



Artificial Neural Network

What does a neuron do?

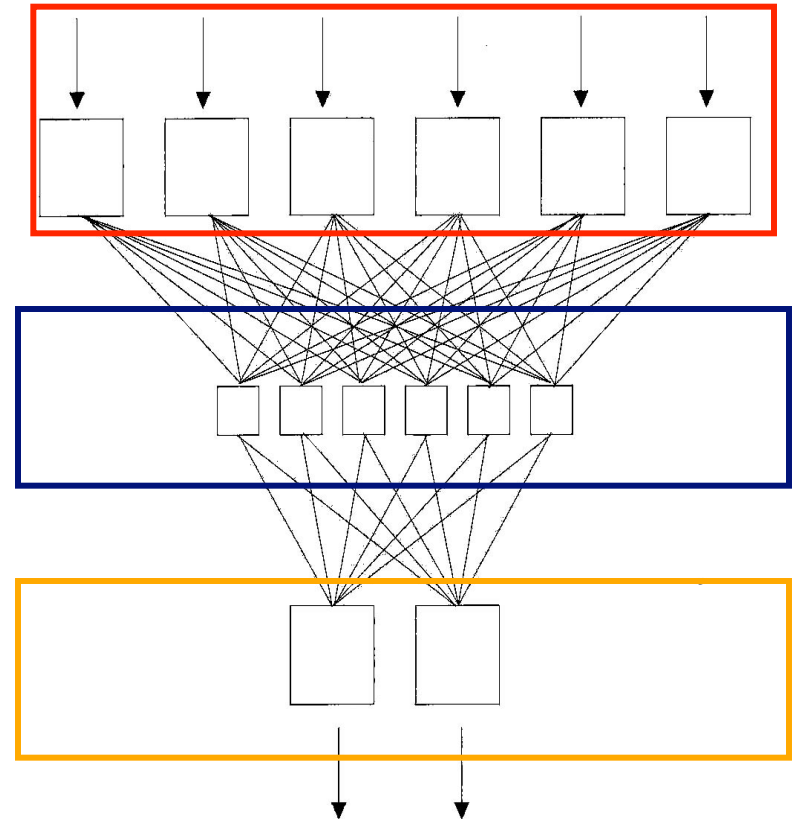
- Gets “signals” from its neighbors.
- Each signal has different weight.
- When achieving certain threshold - sends signals.



Artificial Neural Network

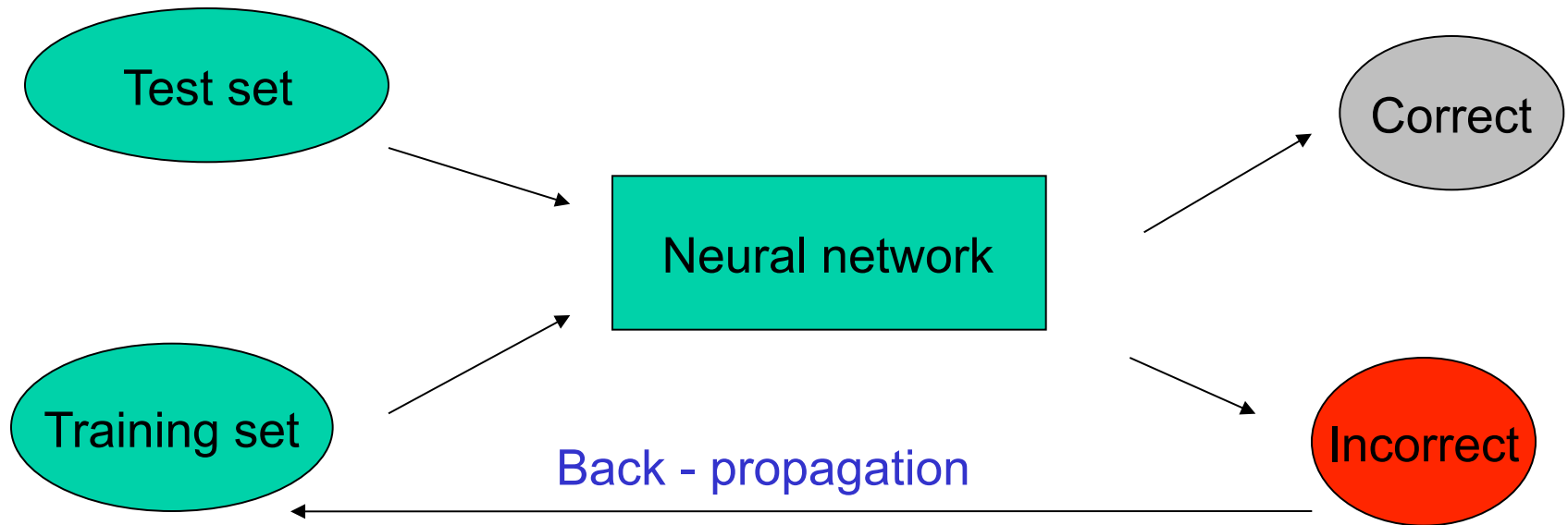
General structure of ANN :

- One input layer.
- Some hidden layers.
- One output layer.
- Our ANN have one-direction flow !



Artificial Neural Network

Network training and testing :

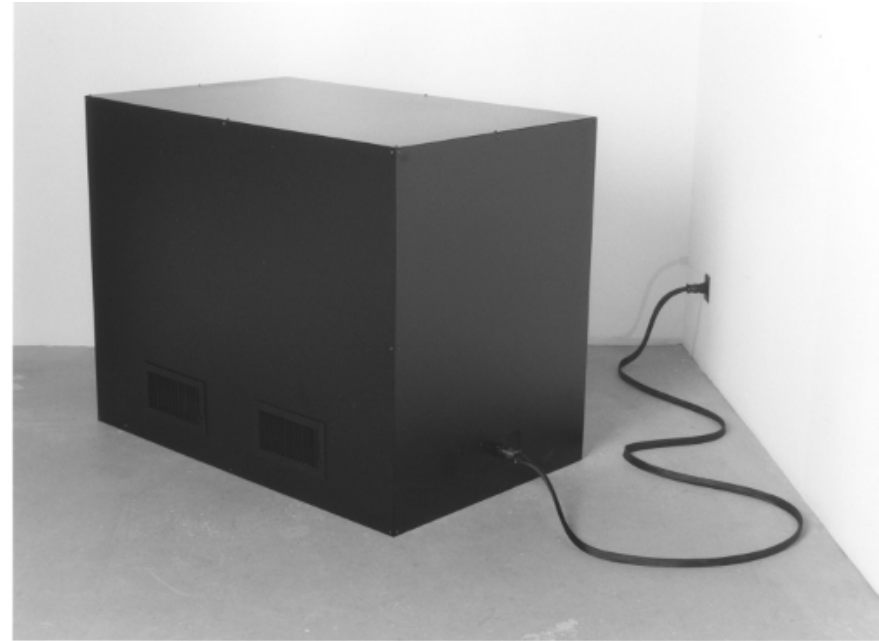


- Training set - inputs for which we know the wanted output.
- Back propagation - algorithm for changing neurons pulses “power”.
- Test set - inputs used for final network performance test.

Artificial Neural Network

The Network is a 'black box' :

- Even when it succeeds it's hard to understand how.
- It's difficult to conclude an algorithm from the network.
- It's hard to deduce new scientific principles.

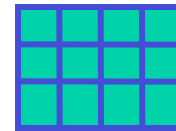


Structure of 3rd generation methods

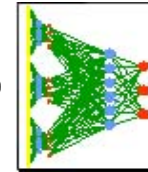
Find homologues using large data bases.



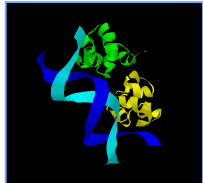
Create a profile representing the entire protein family.



Give sequence and profile to ANN.



Output of the ANN:
2nd structure prediction.



Structure of 3rd generation methods

The ANN learning process:

Training & testing set:

- Proteins with known sequence & structure.

Training:

- Insert training set to ANN as input.
- Compare output to known structure.
- Back propagation.

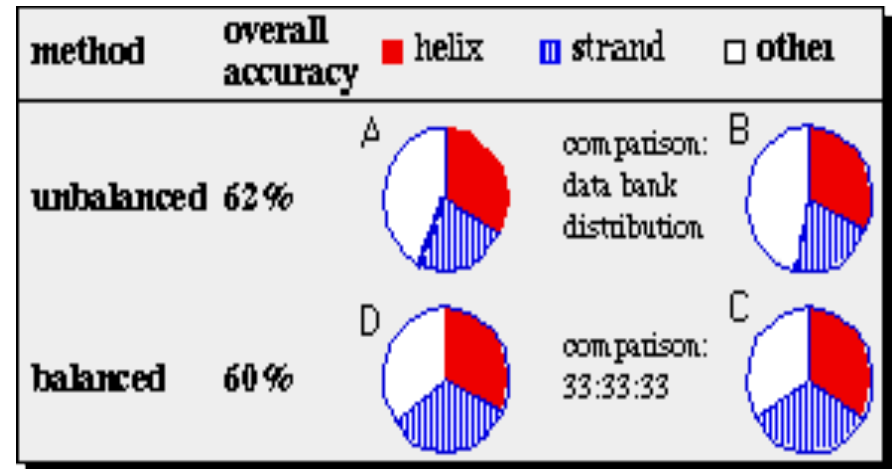
3rd generation methods - difficulties

Main problem - unwise selection of training & test sets for ANN.

- First problem – unbalanced training

Overall protein composition:

- Helices - 32%
- Strands - 21%
- Coils – 47%



What will happen if we train the ANN with random segments ?

3rd generation methods - difficulties

- Second problem – unwise separation between training & test proteins

What will happen if homology / correlation exists between test & training proteins?

Above 80% accuracy in testing.



- Third problem – similarity between test proteins.

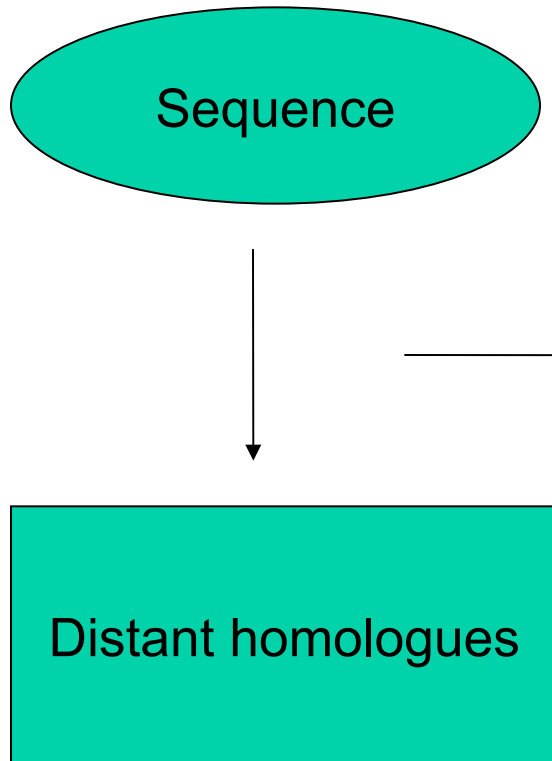
Protein Secondary Structure Prediction Based on Position – specific Scoring Matrices

David T. Jones

**PSI - PRED : 3RD generation method based on the iterated
PSI – BLAST algorithm.**

PSI - BLAST

PSSM - position specific scoring matrix



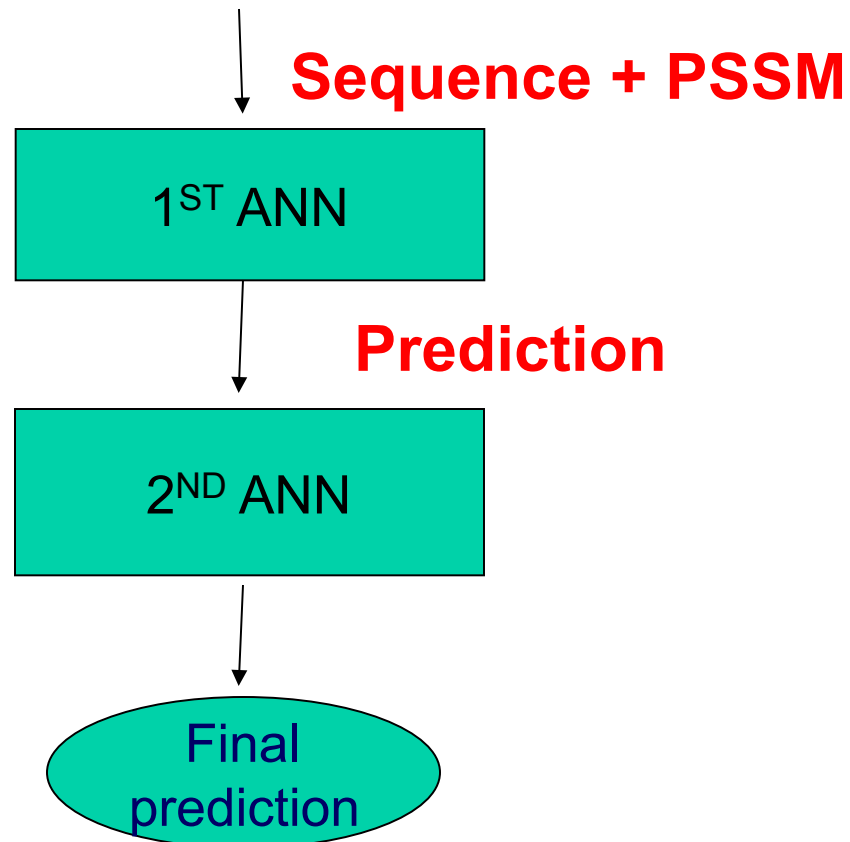
	A	C	D	E	F	G	H	I	L	K	M	P	S	T	W	Y	V		
A	-3	-4	-4	-4	-3	-4	-4	-2	-1	-1	-4	-1	8	-5	-3	-3	0	2	-2
C	0	-1	-1	3	-4	3	4	1	-1	-4	-4	0	-3	-4	-2	-1	-2	-4	-3
D	0	-1	2	1	-3	4	0	-1	-2	-4	-3	1	-2	-4	-2	2	0	-4	-3
E	-2	-3	-4	-5	-2	-3	-4	-6	-4	0	6	0	0	-1	-4	-3	-2	-4	-2
F	0	-3	-1	-2	-3	0	-2	4	-3	-3	0	-2	-2	-4	-3	3	1	-4	-4
G	0	2	0	4	-4	1	2	1	-2	-4	-4	0	-3	-4	-3	1	-2	-5	-4
H	1	5	3	2	4	1	1	1	-2	1	4	1	-3	-4	-3	1	-2	-5	-4
I	-2	-3	-4	-5	-3	-3	-4	-5	-4	3	4	-1	1	2	-4	-3	-2	-3	-1
L	-2	3	2	-2	-4	2	1	-3	-2	-3	-3	1	1	-4	-3	2	1	-4	-3
K	0	2	3	1	-4	0	0	0	-2	-4	-4	1	-3	-4	-3	2	0	-5	-4
M	5	-3	-3	-3	-2	-3	-3	-2	-3	1	-2	-3	-2	1	-3	0	1	-4	-2
P	-1	-4	-5	-5	-3	-4	-4	-5	-4	3	3	-4	2	3	-5	-3	-2	5	-1
S	0	3	3	0	-4	3	0	1	-2	-4	-4	1	-3	-4	-3	1	-1	-4	-3
T	-1	0	1	0	-4	1	-1	-1	-2	-4	-3	5	-2	0	-3	0	-2	-4	0
W	-2	-3	-1	-5	-3	-3	-4	-5	-4	3	4	0	4	2	-4	-3	-2	-3	-2
Y	0	3	0	-2	-3	-1	0	0	-2	0	0	1	0	-1	-3	2	0	-4	-3
V	-1	1	3	-2	-4	0	-2	4	-2	-4	-4	0	-3	0	-3	0	0	-3	0

- PSI – BLAST finds distant homologues.
(It exists now alternatives such as HMMER 3.0 or HHblits)
- **PSSM – input for PSI - PRED.**

PSI - PRED

ANN' s architecture:

- Two ANNs working together.



PSI - PRED

Step 1:

- Create PSSM from sequence - 3 iterations of PSI – BLAST.

Step 2: 1ST ANN

- Sequence + PSSM  1st ANN' s input.

ADCQEIL**H**TSTTWYV
15 RESIDUES



output: central amino acid
secondary state prediction.

ADCQEIL**H**TSTTWYV

PSI - PRED

Using PSI - BLAST brings up PSI – BLAST difficulties:

Iteration - extension of proteins family

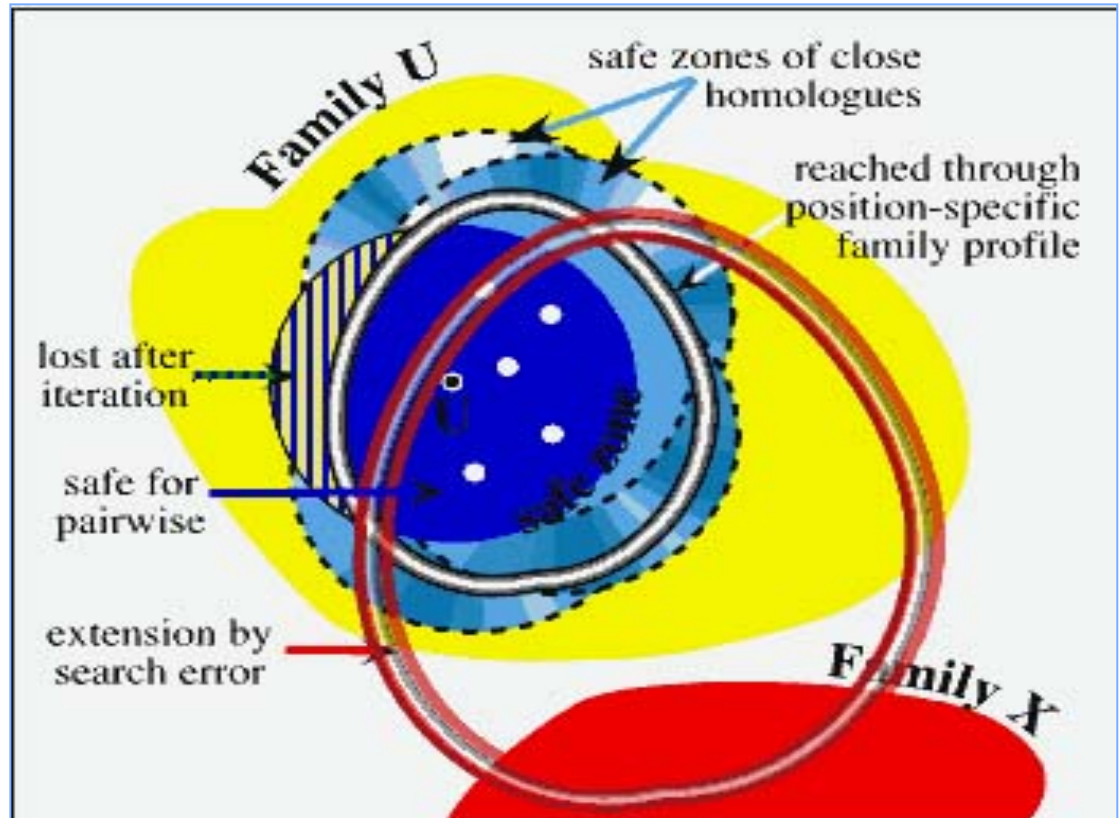


Updating PSSM

Inclusion of non – homologues



“Misleading” PSSM



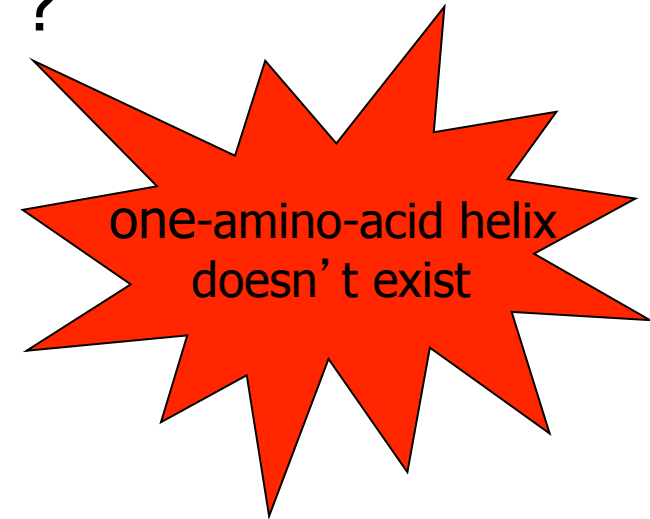
PSI - PRED

Step 3: 2nd ANN

- So why do we need a second ANN ?

possible output for 1st ANN:

seq	A	A	P	L	L	L	M	M	M	S	I	M	R	R	I	M
pred	E	E	E	E	E	C	C	C	C	H	C	C	C	C	E	E



what's wrong with that ?

Solution: ANN that “looks” at the whole context !

Input: output of 1st ANN.

Output: final prediction.

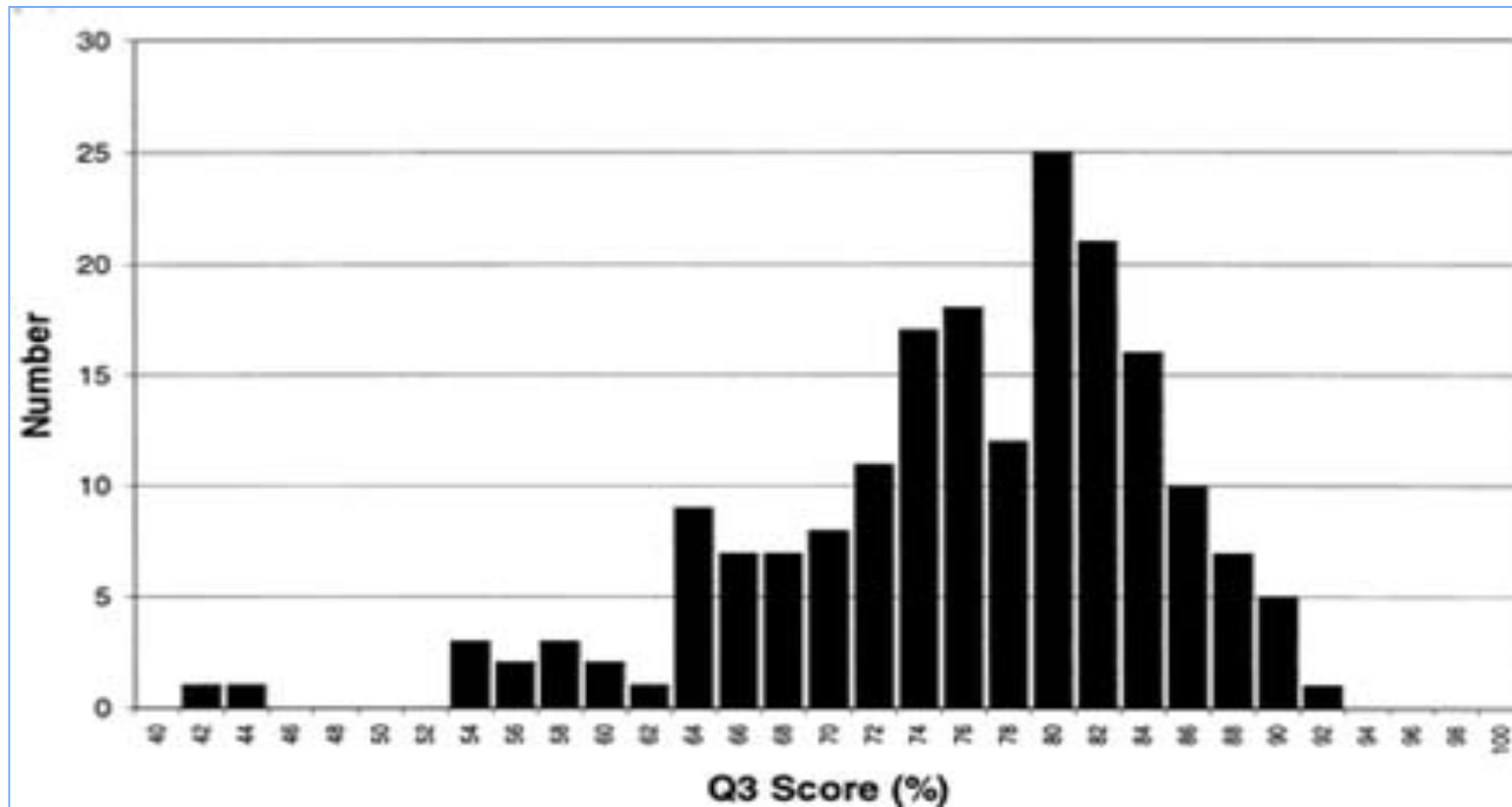
PSI - PRED

Training : Balanced training.

- Testing :**
- 187 proteins, Highly resolved structure.
 - PSI – BLAST was used for removing homologues.
 - Without structural similarities.

PSI - PRED

Jones' s reported results : Q₃ results : 76% - 77%



PSI - PRED

Reliability numbers:

- The way the ANN tells us how much it is sure about the assignment.
- Used by many methods.
- **Correlates with accuracy.**

PSIPRED PREDICTION RESULTS

Key

Conf: Confidence (0=low, 9=high)

Pred: Predicted secondary structure (H=helix, E=strand, C=coil)


AA: Target sequence

Conf: 97898377188899998530367741489987089

Pred: CEEEEECCHHHHHHHHHHHHCCCCCEEEEEEC

AA: KVVIIKPPPLVVLVLRRRRAGAGALLILIKPP

Conf: 

Pred: 

Pred: CEEEEECCHHHHHHHHHHHHCCCCCEEEEEEC

AA: KVVIIKPPPLVVLVLRRRRAGAGALLILIKPP

10 20 30


Legend:

 = helix

 = strand

 = coil

Conf:  = confidence of prediction

Pred:  predicted secondary structure

AA: target sequence

Performance Evaluation

- Through 3rd generation methods accuracy jumped ~10%.

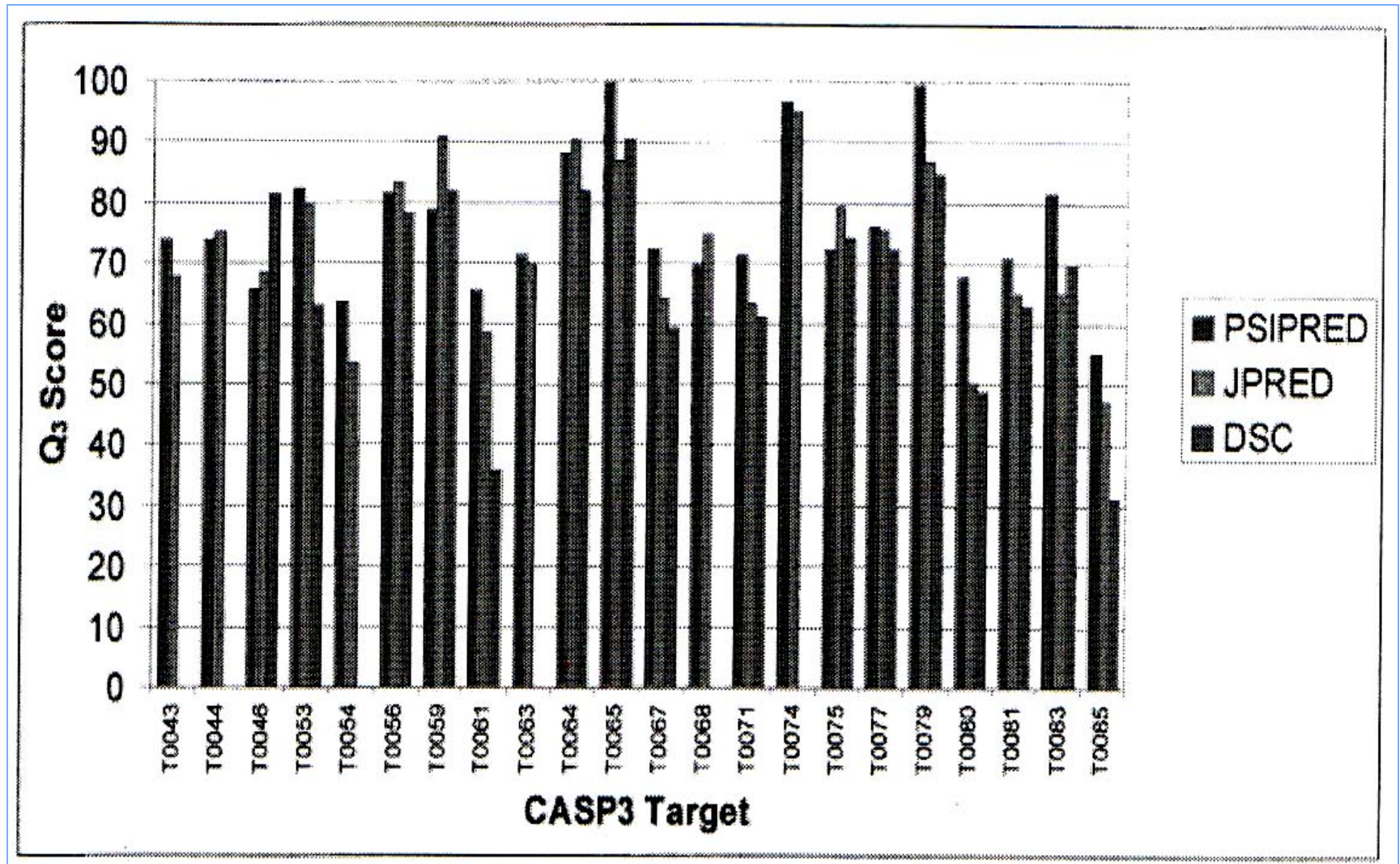
SEQ	KELVLALYDYQEKSPREVTMCKGDI LTL LN STN KDN NKVEVND RQGFVPAAYVKKLD									
OBS	EEEE		E--E	EEEEEE	EEEEEE	EEEEEE	EEEEEE	EEEEEE	EEEEEE	EEEEEE
1st C+F				EEEEEE			EEEEEE			EEEEEE
2nd GOR	H		H	HH	EEEEEE	EEEE			HH	HH
3rd PHD	EEEEEE		EEE	EEEEEEEE				EEEE	HH	EEEE
Rel	948999972587775211443884899847697314344045955111321221558									
	* *****			** *****				****		***

- Many 3rd generation methods exist today.

Which method is the best one ?

How to recognize “over-optimism” ?

Performance Evaluation



Performance Evaluation

Conclusion :

PSI-PRED seems to be one of the most reliable method today.

Reasons :

- The widest evolutionary information (PSI - BLAST profiles).
- Strict training & testing criterions for ANN.

Improvements

The first 3rd generation method **PHD**: ~72% in Q₃.

3rd generation methods best results: ~77% in Q₃.

Sources of improvement :

- Larger protein data bases.
- **PSI – BLAST**
PSI – PRED broke through, many followed...

Improvements

How can we do better than that ?

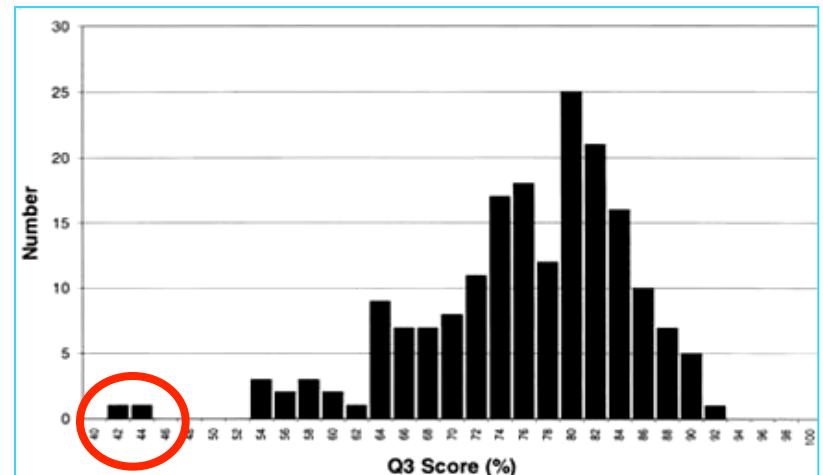
Through larger data bases (?).

- Combination of methods.

Example:

Combining 4 best methods \longrightarrow Q₃ of ~78% !

- Find why certain proteins predicted poorly.



Bibliography

- Jones DT. Protein secondary structure prediction based on position specific scoring matrices. *J Mol Biol.* 1999 292:195-202
- Rost B. Rising accuracy of protein secondary structure prediction 'Protein structure determination, analysis, and modeling for drug discovery' (ed. D Chasman), New York: Dekker, pp. 207-249