CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

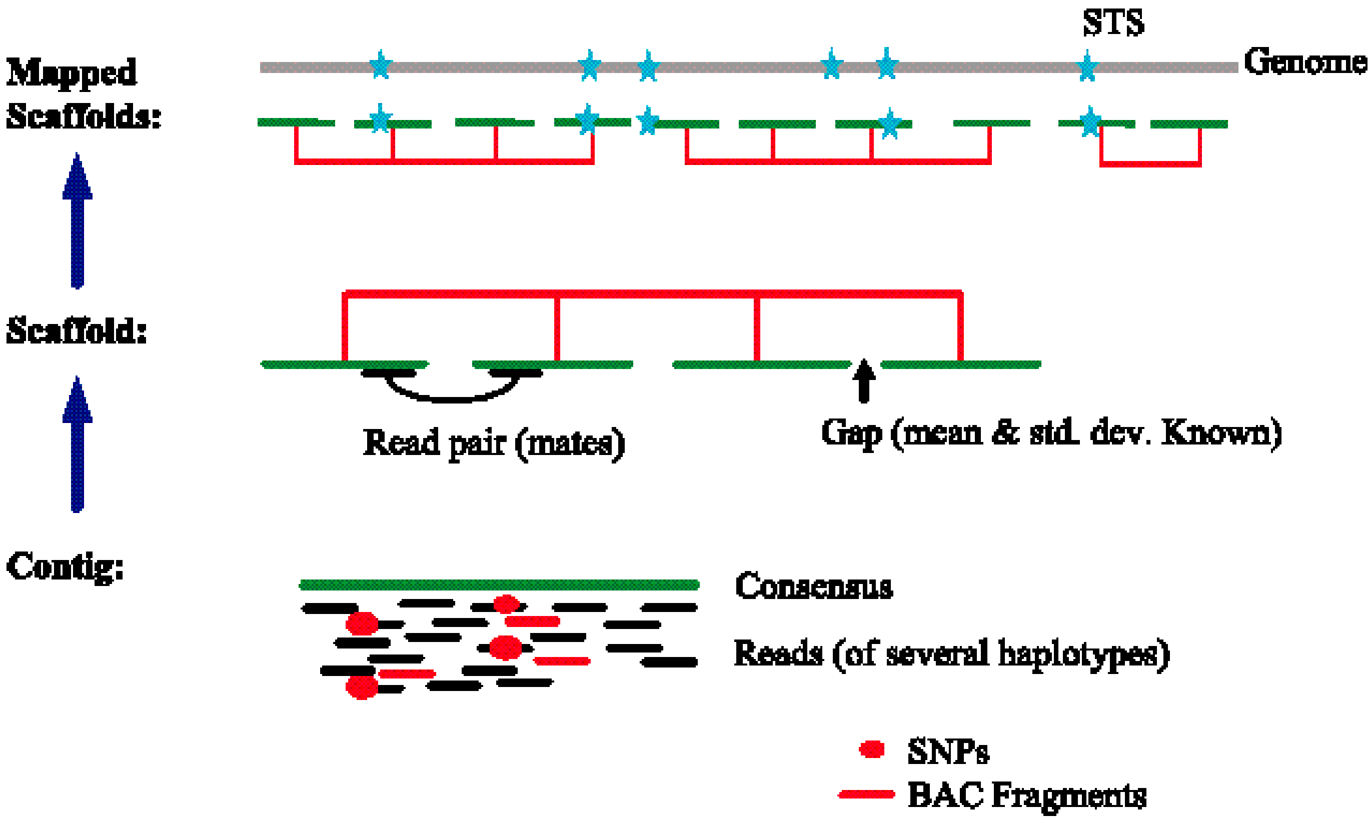
Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS13.html

Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

Human Genome Project

- ❑ Many videos available on youtube.com, dnatube.com, and elsewhere.
- ❑ Find some and watch them.

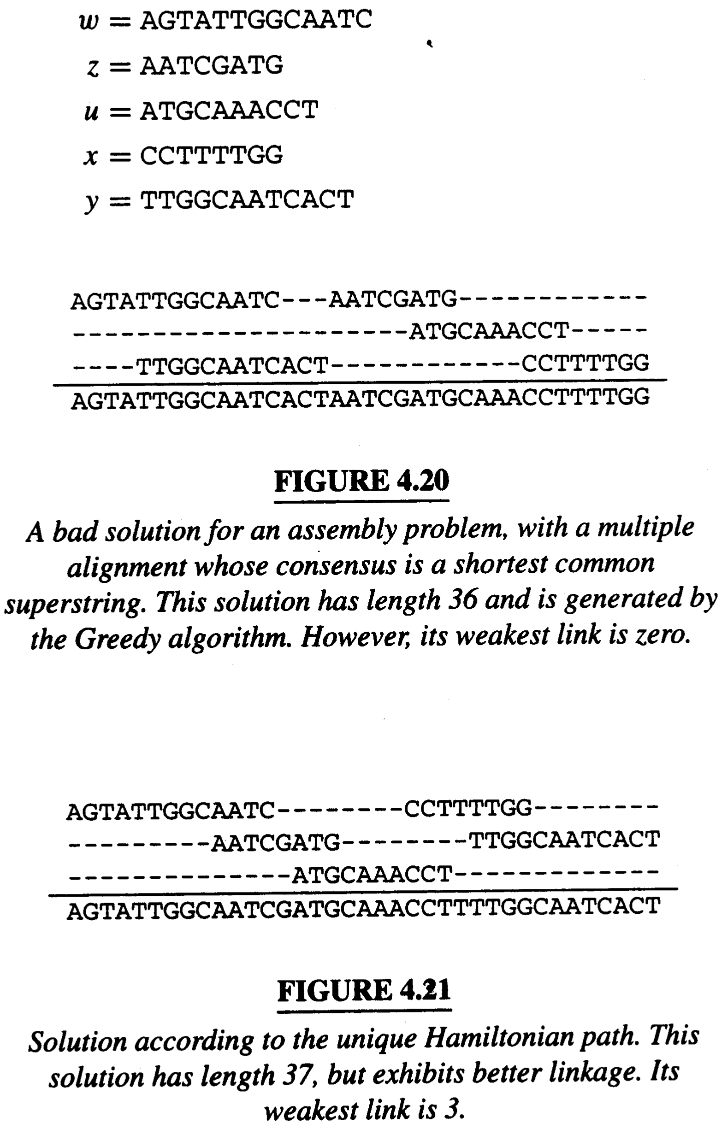
Assembly: Simple Example

- ACCGT, CGTGC, TTAC, TACCGT
- Total length = ~10
-
  - **--ACCGT--**
  - **----CGTGC**
  - **TTAC-----**
  - **-TACCGT–**
  - **TTACCGTGC**

Assembly: Complications

- Errors in input sequence fragments (~3%)
    - Indels or substitutions
- Contamination by host DNA
- Chimeric fragments (joining of non-contiguous fragments)
- Unknown orientation
- Repeats (long repeats)
    - Fragment contained in a repeat
    - Repeat copies not exact copies
    - Inherently ambiguous assemblies possible
    - Inverted repeats
- Inadequate Coverage

Assembly: Complications



$w =$ AGTATTGGCAATC

$z =$ AATCGATG

$u =$ ATGCAAACCT

$x =$ CCTTTTGG

$y =$ TTGGCAATCACT

```
AGTATTGGCAATC---AATCGATG------------
---------------------ATGCAAACCT-----
----TTGGCAATCACT------------CCTTTTGG
AGTATTGGCAATCACTAATCGATGCAAACCTTTTGG
```

FIGURE 4.20

A bad solution for an assembly problem, with a multiple alignment whose consensus is a shortest common superstring. This solution has length 36 and is generated by the Greedy algorithm. However, its weakest link is zero.

```
AGTATTGGCAATC--------CCTTTTGG--------
---------AATCGATG--------TTGGCAATCACT
--------------ATGCAAACCT-------------
AGTATTGGCAATCGATGCAAACCTTTTGGCAATCACT
```

FIGURE 4.21

Solution according to the unique Hamiltonian path. This solution has length 37, but exhibits better linkage. Its weakest link is 3.

Assembly: Complications

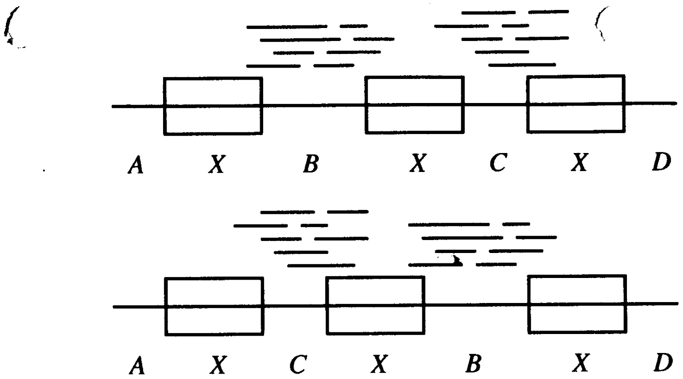


FIGURE 4.8

Target sequence leading to ambiguous assembly because of repeats of the form XXX.

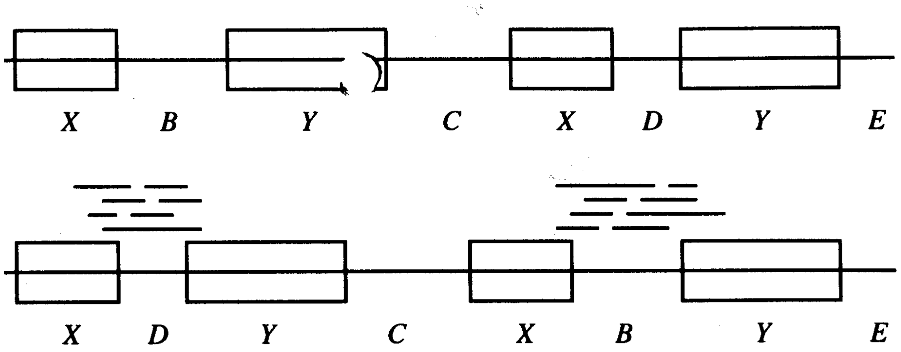


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form $XYXY$.

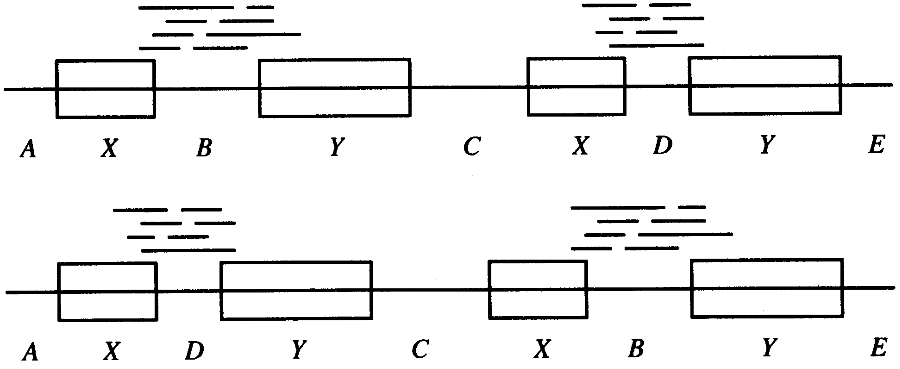


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form $XYXY$.

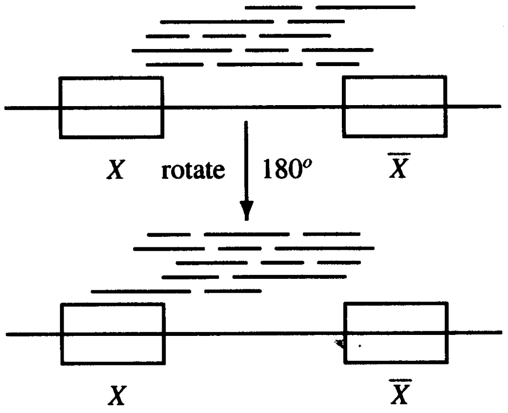


FIGURE 4.10

Target sequence with inverted repeat. The region marked $\overline{X}$ is the reverse complement of the region marked X.

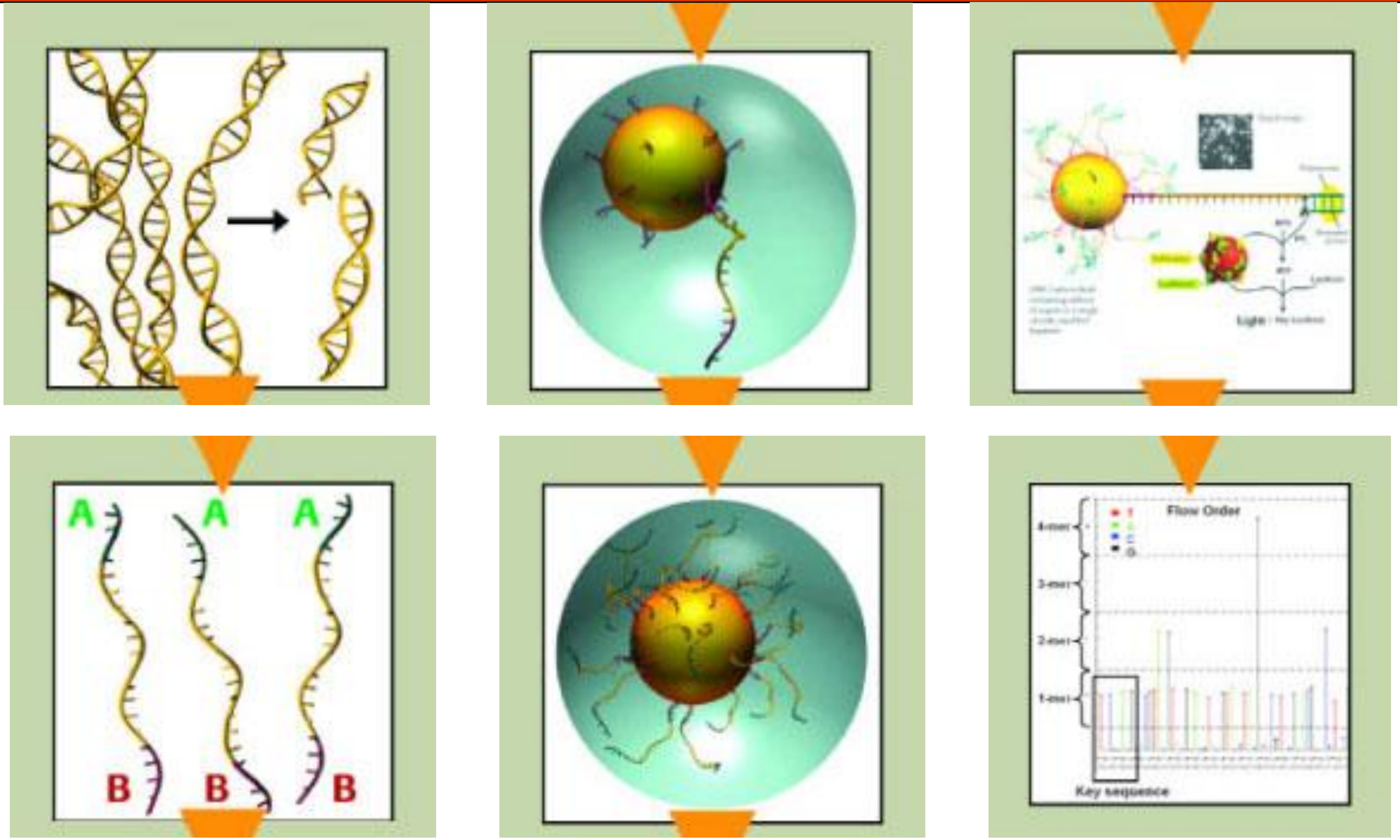
Other sequencing methods

- ❑ Sanger Method (70Kbp/run)
- ❑ Sequencing by Hybridization (SBH)
- ❑ Dual end sequencing
- ❑ Chromosome Walking (see page 5-6 of Pevzner's text)
- ❑ 454 Sequencing (60Mbp/run)
- ❑ Solexa Sequencing (600Mbp/run) [Illumina]

454 Sequencing: New Sequencing Technology

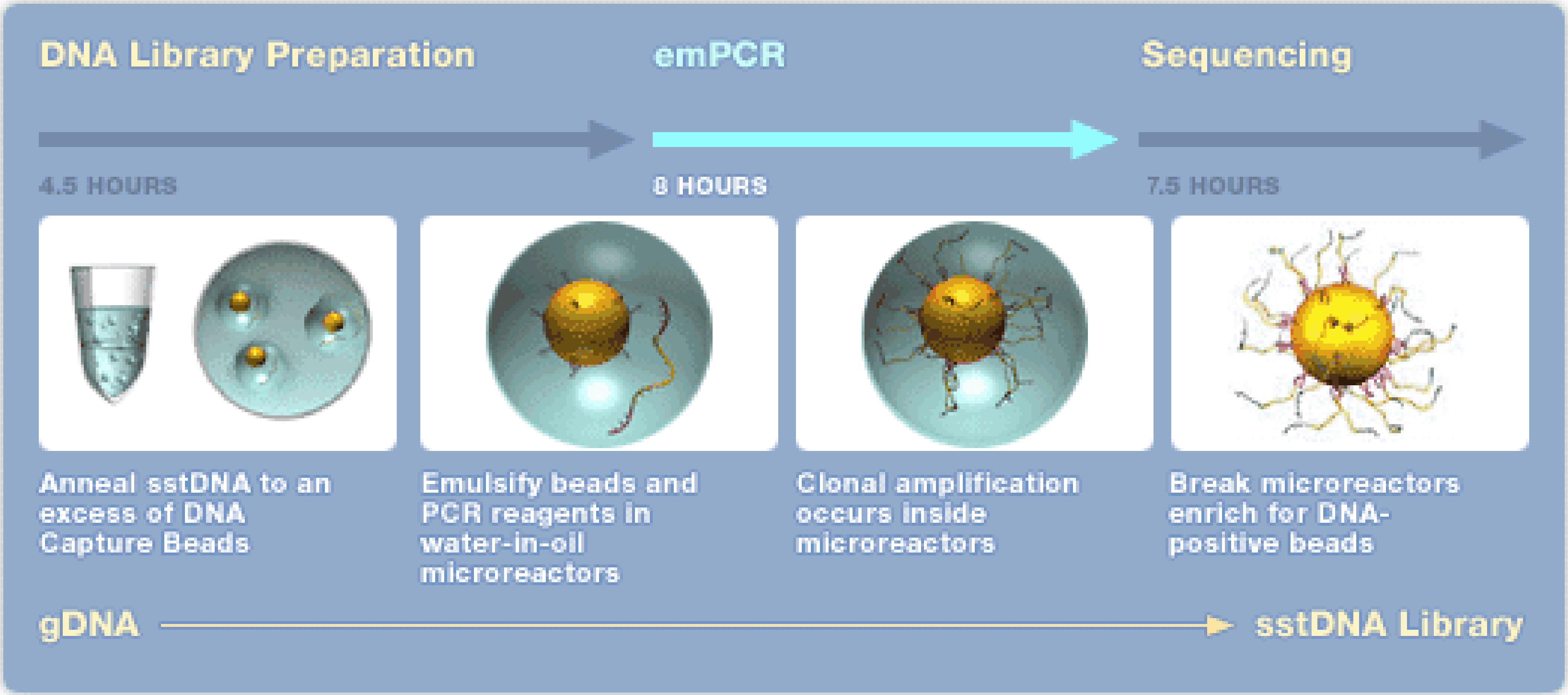
- 454 Life Sciences, Roche
- Fast (20 million bases per 4.5 hour run)
- Low cost (lower than Sanger sequencing)
- Simple (entire bacterial genome in days with one person -- without cloning and colony picking)
- Convenient (complete solution from sample prep to assembly)
- PicoTiterPlate Device
  - Fiber optic plate to transmit the signal from the sequencing reaction
- Process:
  - Library preparation: Generate library for hundreds of sequencing runs
  - Amplify: PCR single DNA fragment immobilized on bead
  - Sequencing: "Sequential" nucleotide incorporation converted to chemilluminscent signal to be detected by CCD camera.

(a) Fragment, (b) add adaptors, (c) "1 fragment, 1 bead", (d) emPCR on bead, (e) put beads in PicoTiterPlate and start sequencing: "1 bead, 1 read", and (f) analyze



emPCR

FIGURE 8



Anneal sstDNA to an excess of DNA Capture Beads

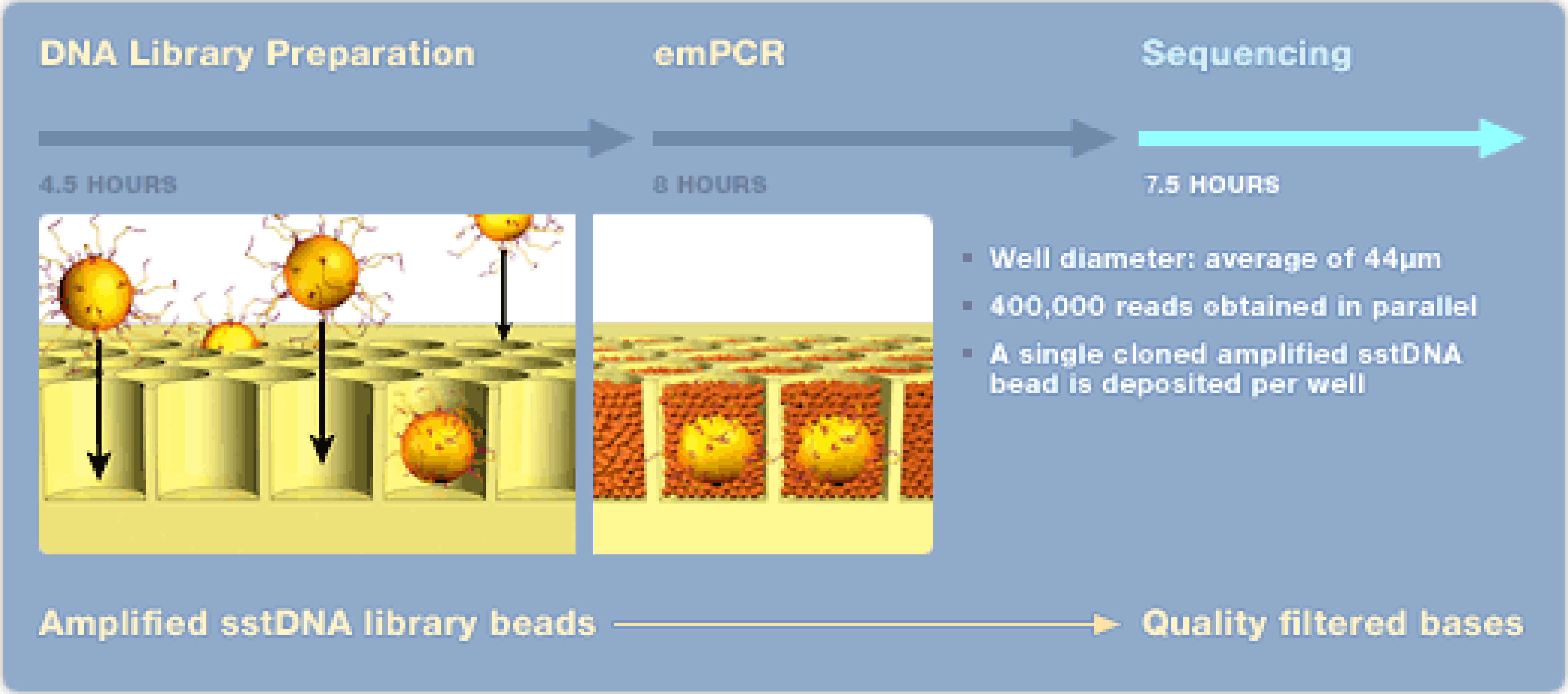
Emulsify beads and PCR reagents in water-in-oil microreactors

Clonal amplification occurs inside microreactors

Break microreactors enrich for DNA-positive beads

Sequencing

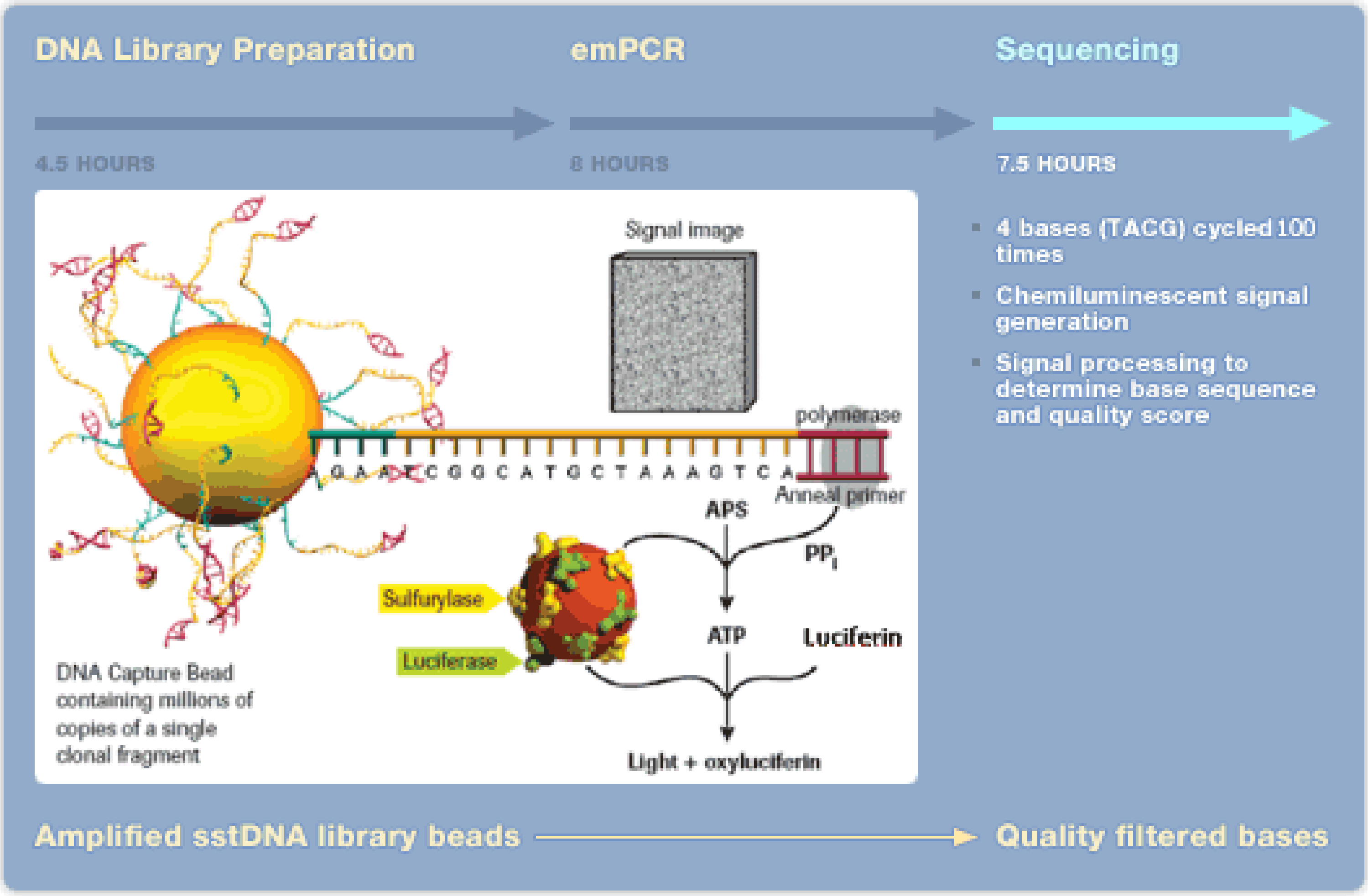
FIGURE 9



- Well diameter: average of 44µm
- 400,000 reads obtained in parallel
- A single cloned amplified sstDNA bead is deposited per well

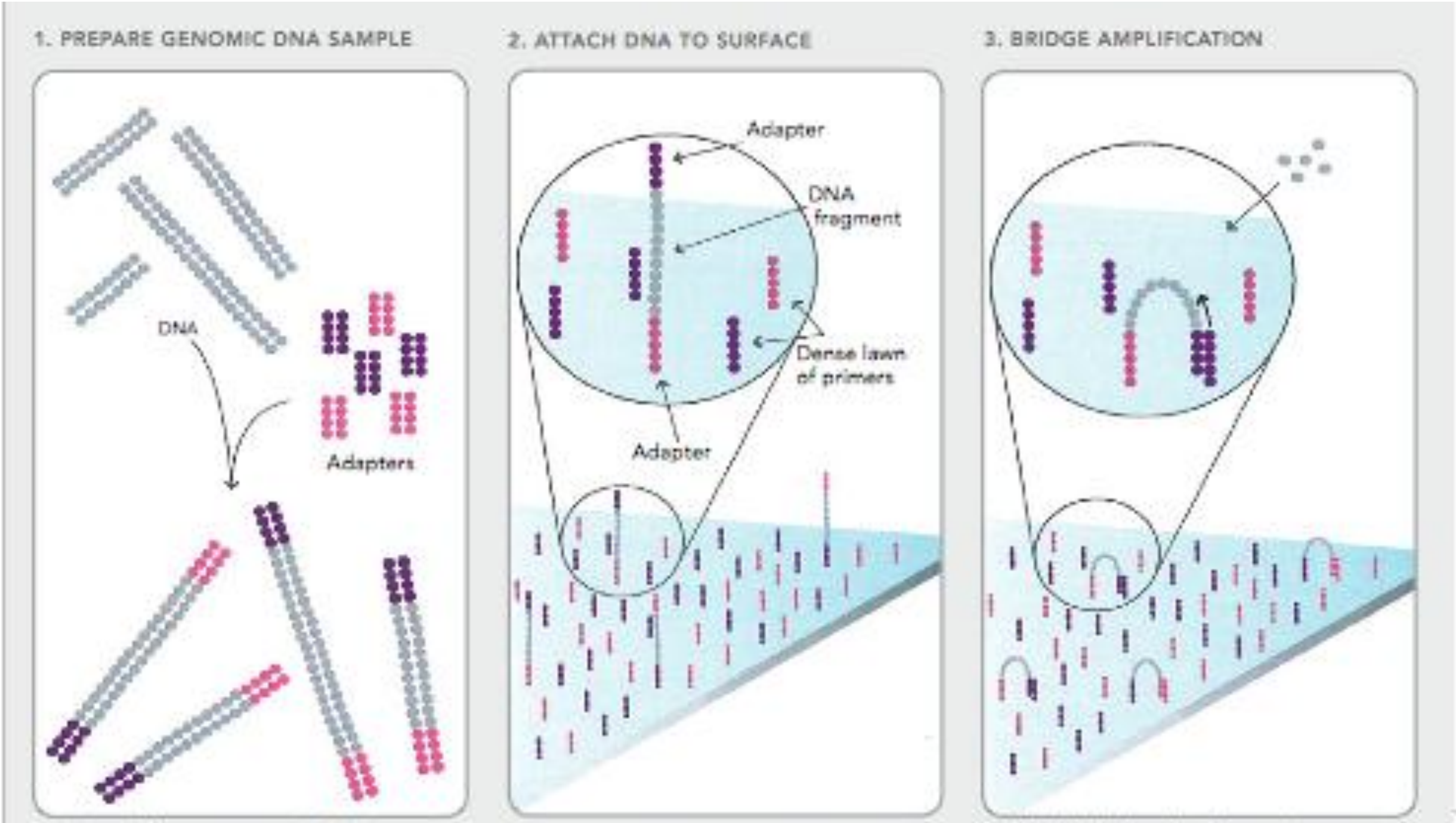
Sequencing

FIGURE 10



- 4 bases (TACG) cycled 100 times
- Chemiluminescent signal generation
- Signal processing to determine base sequence and quality score

Solexa Sequencing



1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

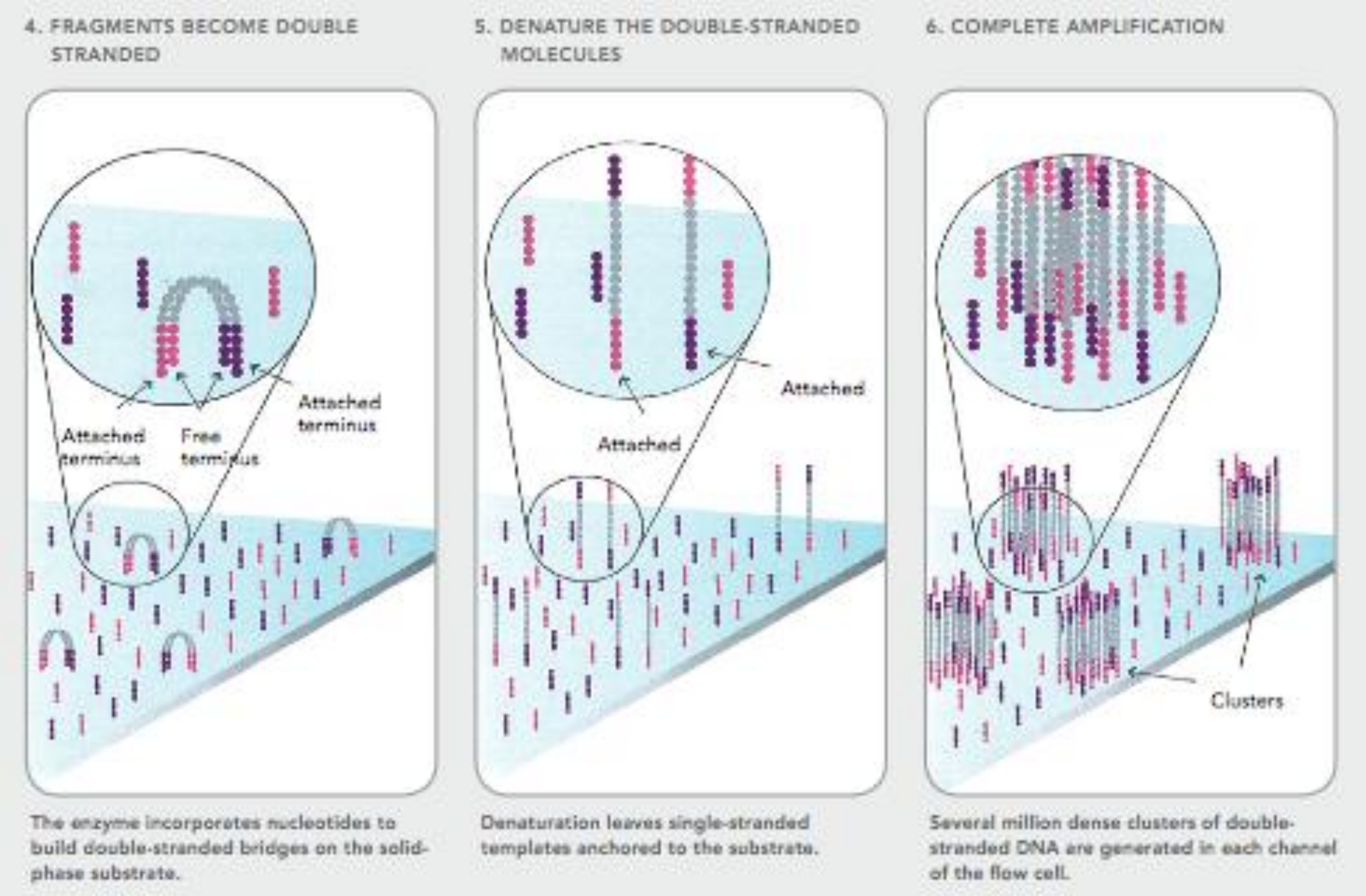
2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Solexa Sequencing



4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

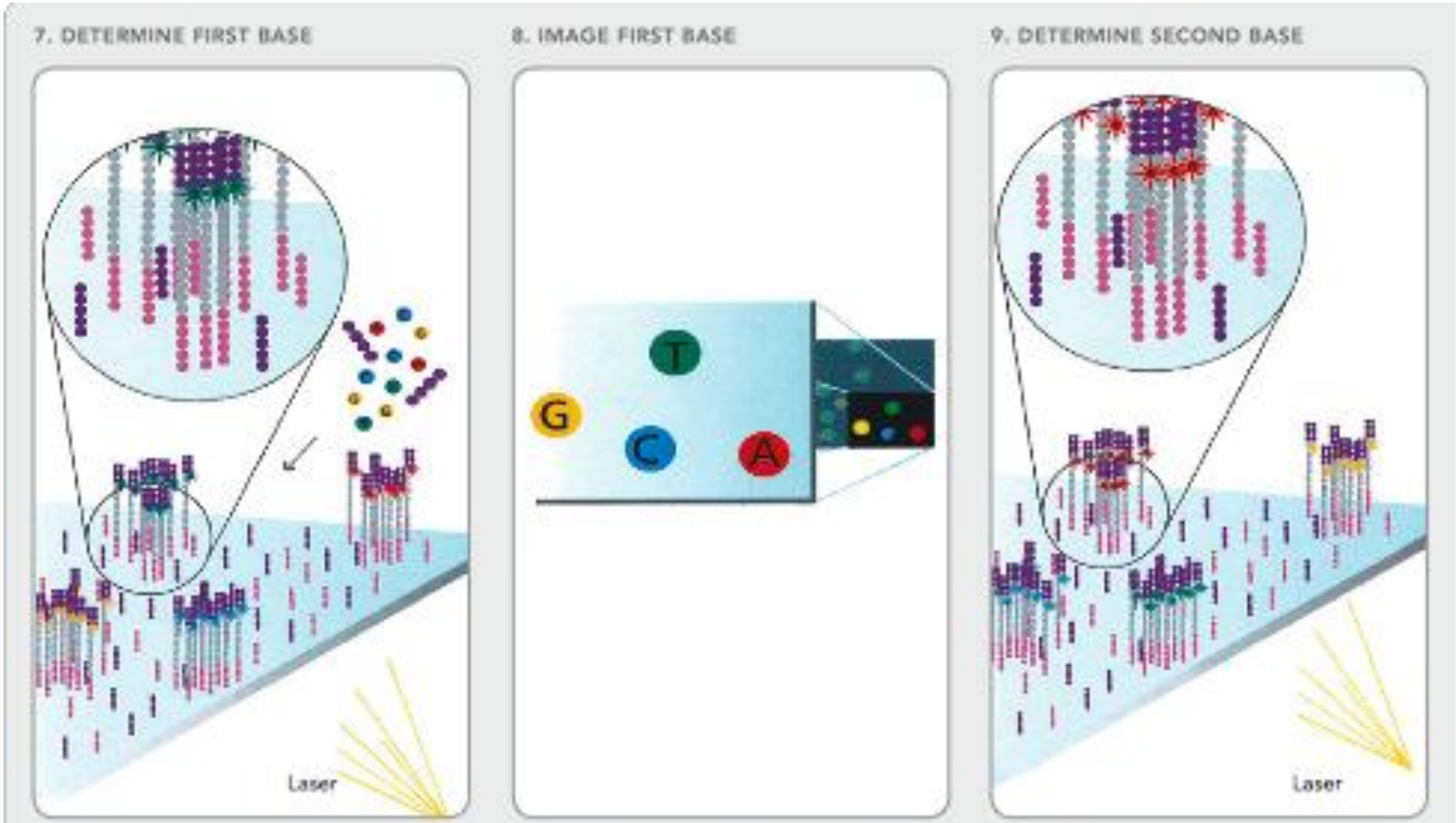
5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Solexa Sequencing



7. DETERMINE FIRST BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

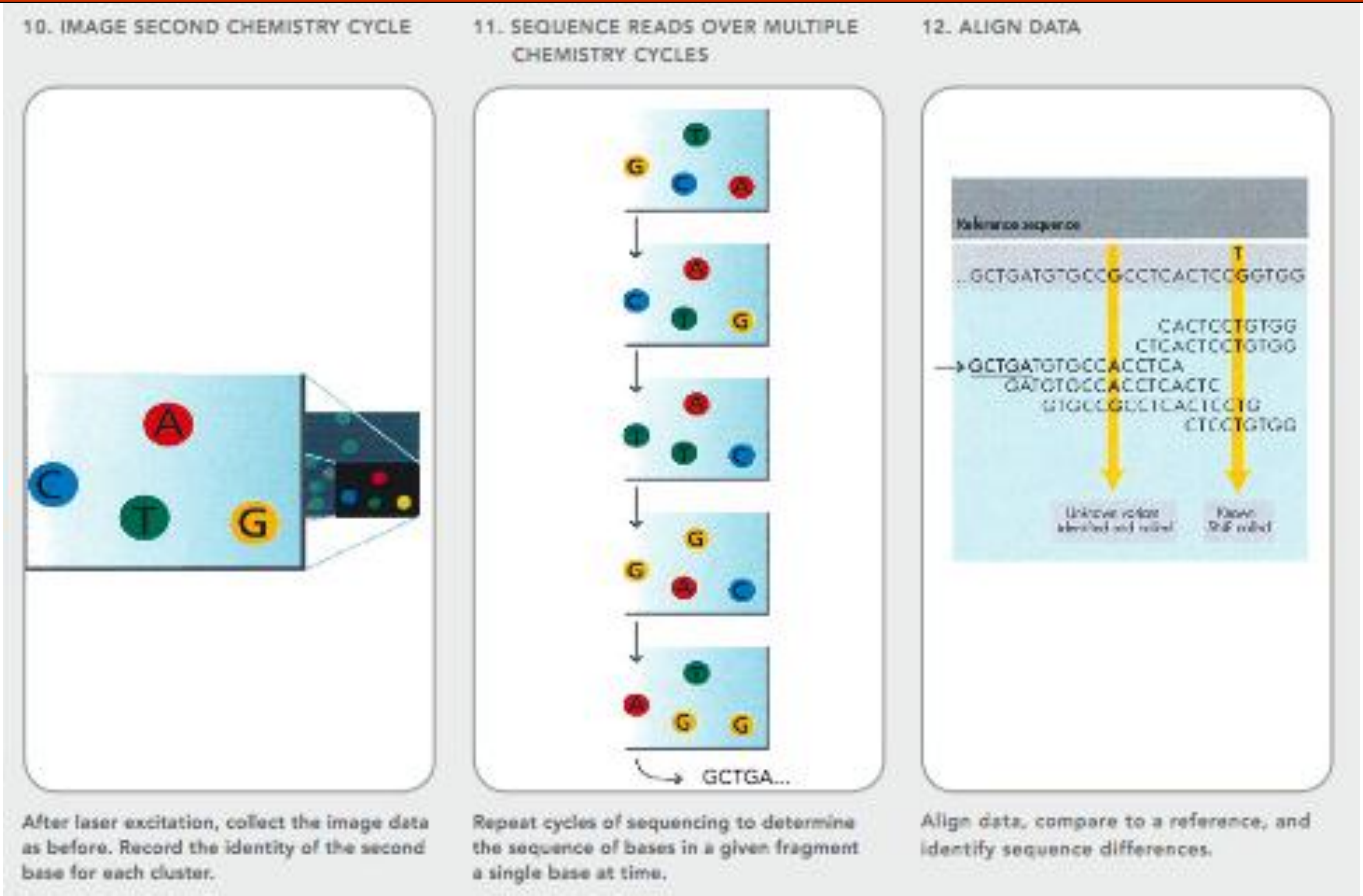
8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

Solexa Sequencing



10. IMAGE SECOND CHEMISTRY CYCLE

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA

Align data, compare to a reference, and identify sequence differences.

Assemblers

- TIGR Assembler (TIGR)
- Phrap (U Washington)
- Celera Assembler (Celera Genomics)
- Arachne (Broad Institute of MIT & Harvard)
- Phusion (Sanger Center)
- Atlas (Baylor College of Medicine)

Applications of Sequencing

- ❑ Sequencing
- ❑ Resequencing
- ❑ SNP detection
- ❑ RNA-Seq
- ❑ CHiP-Seq
- ❑ Metagenomics

Basic Assembler

- **Read**: sequenced fragment; **Contig**: contiguous segment. How to assemble a contig?

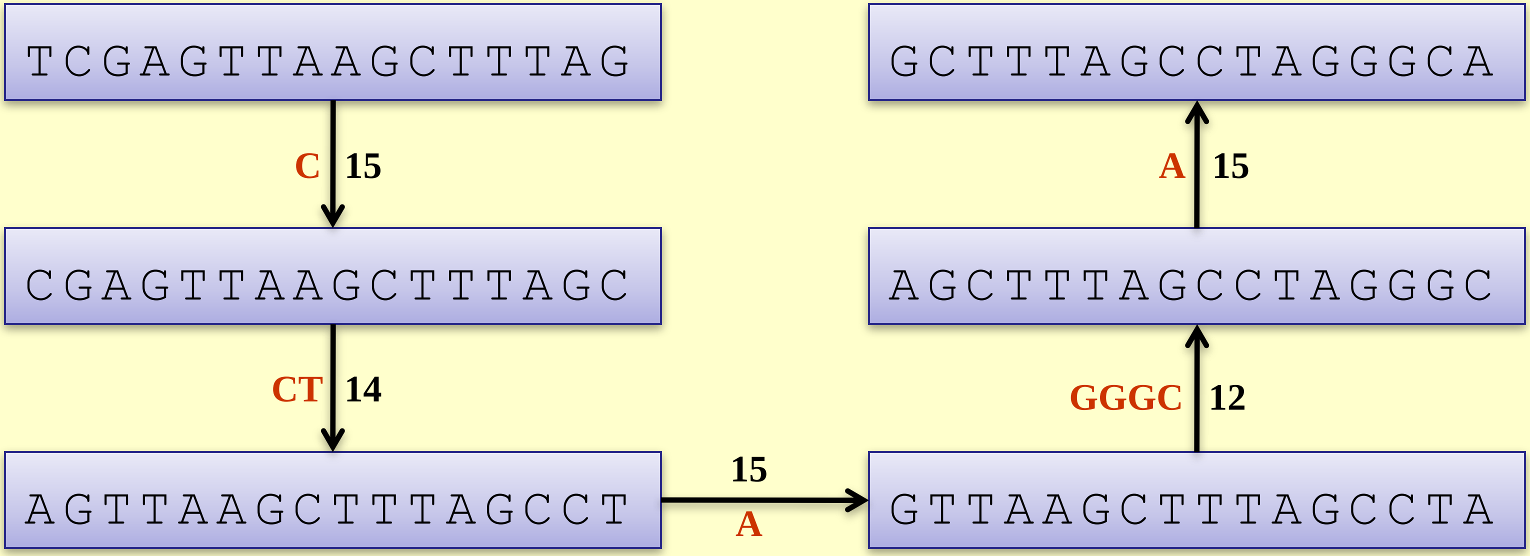
```
TCGAGTTAAGCTTTAG
 CGAGTTAAGCTTTAGC
   AGTTAAGCTTTAGCCT
    GTTAAGCTTTAGCCTA
       AGCTTTAGCCTAGGGC
         GCTTTAGCCTAGGCAG
            ...
```

```
AGCTTTAGCCTAGGGC
AGTTAAGCTTTAGCCT
CGAGTTAAGCTTTAGC
GCTTTAGCCTAGGCAG
GTTAAGCTTTAGCCTA
TAAGCTTTAGCCTAGG
TCGAGTTAAGCTTTAG
```

Problem: Need to try every pair of reads!

Reduce to Graph Problem

- How to assemble a contig?
  - Node ⟷ Read
  - Edge between Nodes ⟷ Overlapping Reads
  - **Problem**: Find a path through each node in graph.



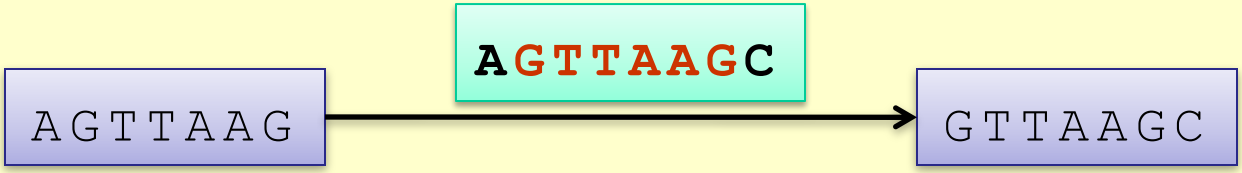
Issues: Problem is NP-Complete
# nodes = # reads
# of edges ≤ k(# nodes)

A better solution

- Take each read and chop it into k-mers.
- Represent k-mers by nodes in a graph and edges between k-mers that overlap in k-1 bases.
- **Consequence**:
  - Number of nodes = 4^k ;
  - Number of edges = $k4^k$;
- **Issues**:
  - Problem (i.e., find path through all vertices) remains NP-Complete

A more efficient solution

- Represent every possible (k-1)-mer by a node.
- Edges connect 2 nodes if they share k-2 bases.
- Label each edge by k-mer.



- Problem:
  - Find a path through each edge in the graph
- The Eulerian path problem is **NOT** NP-Complete. It can be solved in linear time!

Sources of Assembly Errors

- Errors in reads – caused by technology
  - Error in base calls, color calls (SOLID Technology), or repeated base calls (454 Technology)
- Missing reads – sequencing bias
- Read orientation error
  - One or both orientations may occur
  - Not told which ones are present
- Sequence Variations – mixed sample study
  - SNP, cancer, metagenomics studies
- **REPEATS**
- Combinations of the above

How to deal with REPEAT Regions

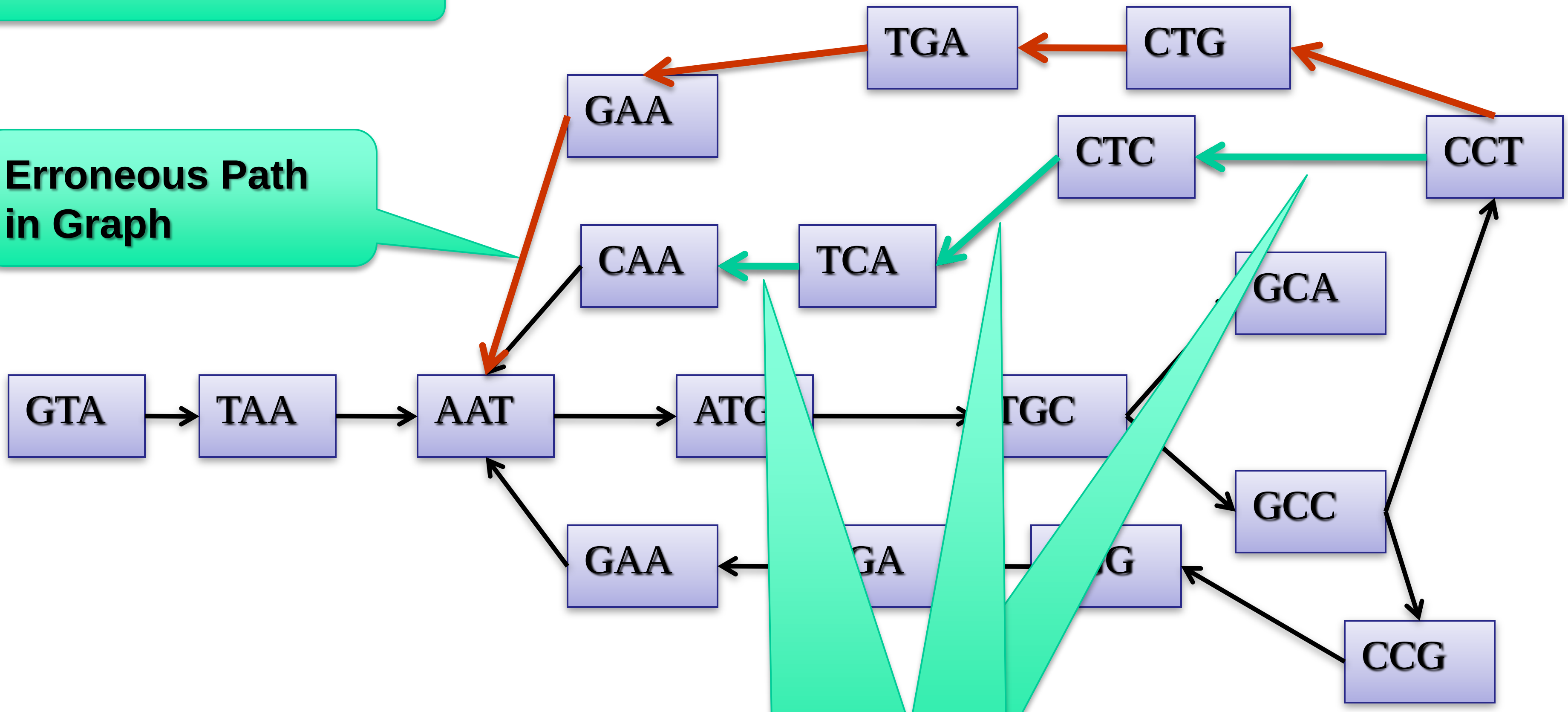
- If no errors or repeat regions, then the graph has a unique path through all the edges.
- **Problem**: REPEAT regions cause branching in graph. If no errors in reads, then the graph has a unique path through all edges, but with some edges traversed more than once.
- How to identify REPEAT regions:
  - Higher coverage of repeat regions
  - Branching of nodes

Handling Read Error

GTAATGCCTCAATGCCGGAATGCA

CTGAA

Erroneous Base Call



Erroneous Path in Graph

Potential Missing Edges in Graph

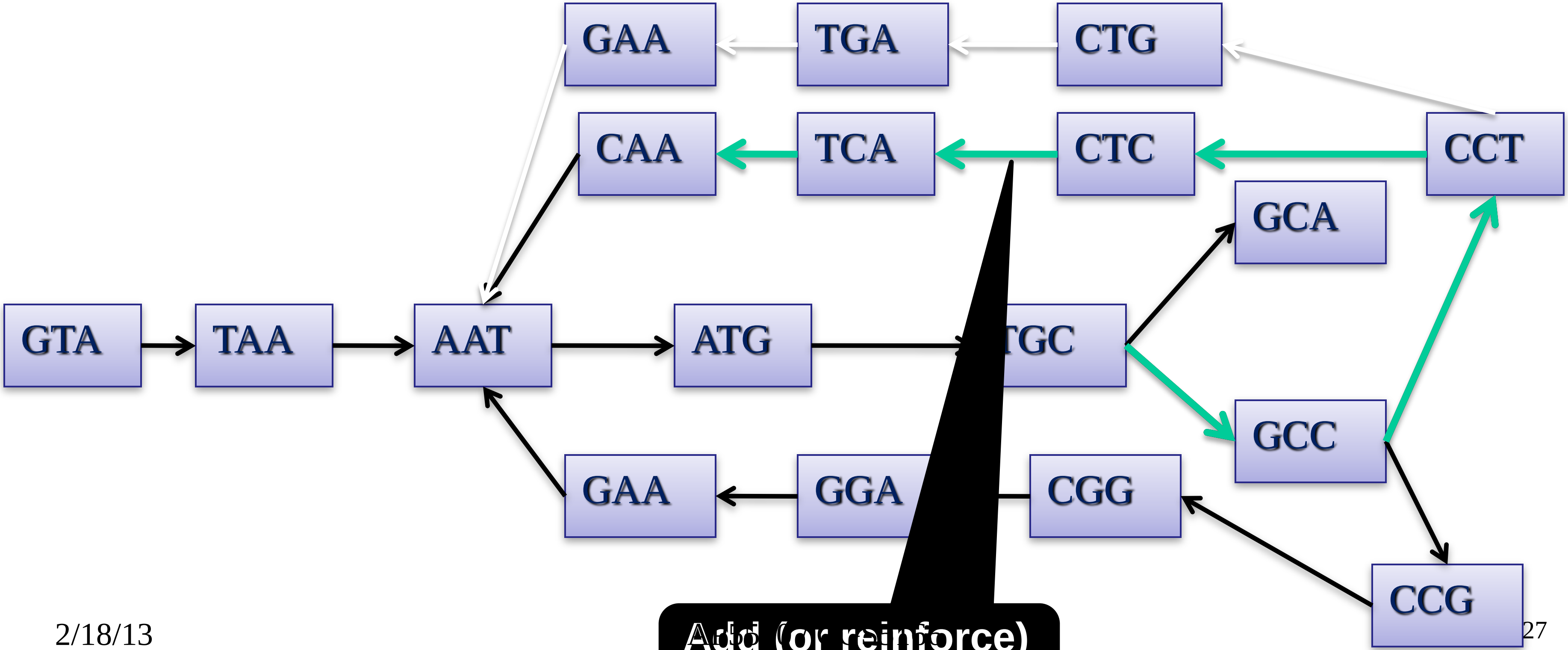
Well conserved regions in related genomes

TGCCTCAA

TGCCTCAA

GTAATGCCTCAATGCCGGAATGCA

CTGAA



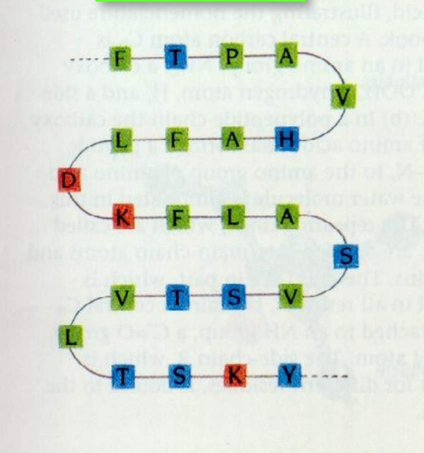
Add (or reinforce) path in graph

Protein Structures

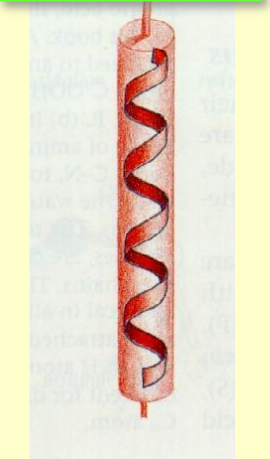
Protein Structures

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

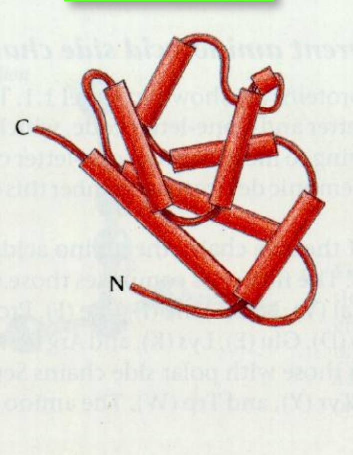
Primary



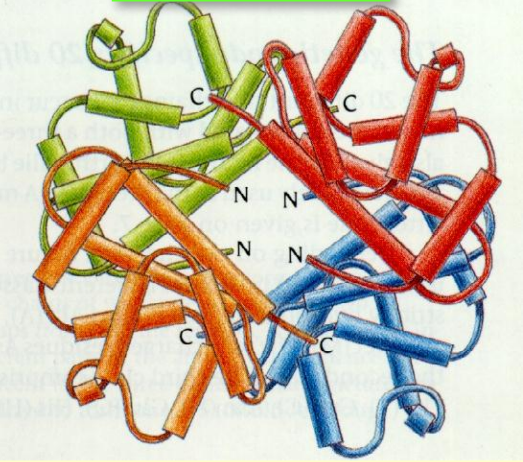
Secondary



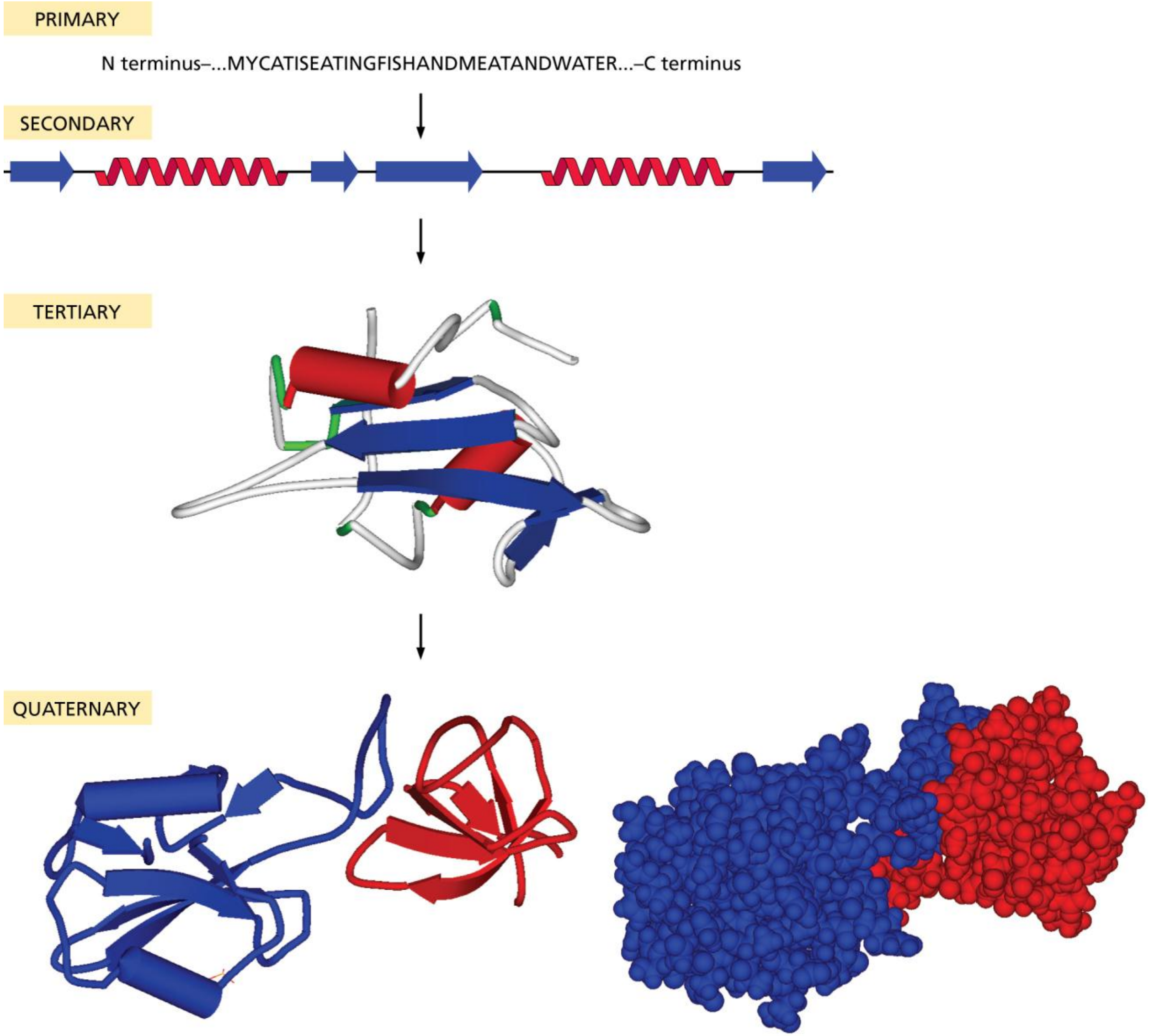
Tertiary



Quaternary



Proteins: Levels of Description

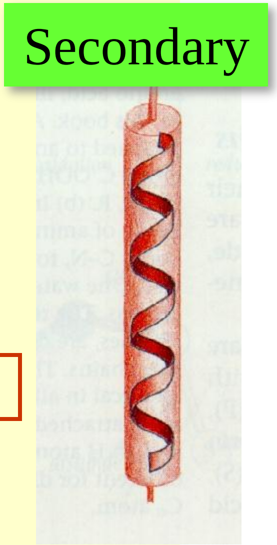

PRIMARY
N terminus–...MYCATISEATINGFISHANDMEATANDWATER...–C terminus
SECONDARY
TERTIARY
QUATERNARY

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…

Secondary

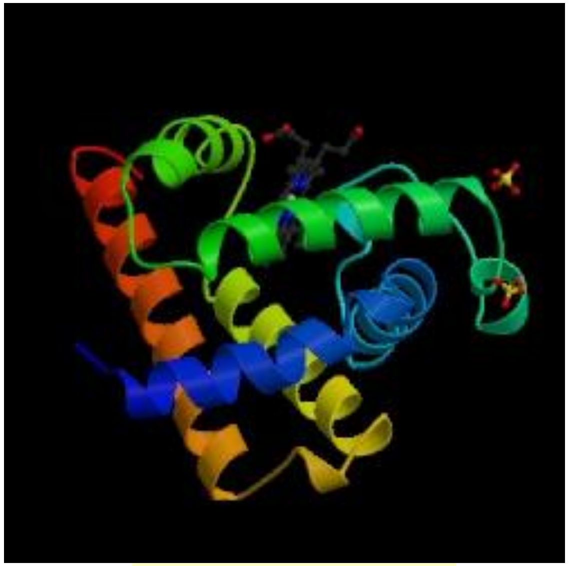


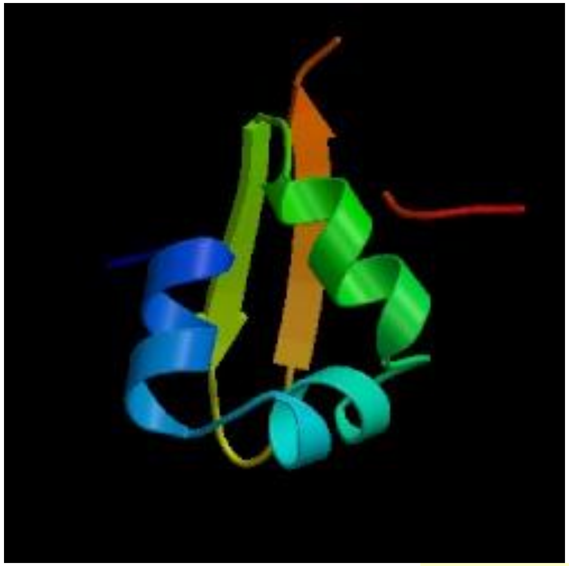
More on Secondary Structures

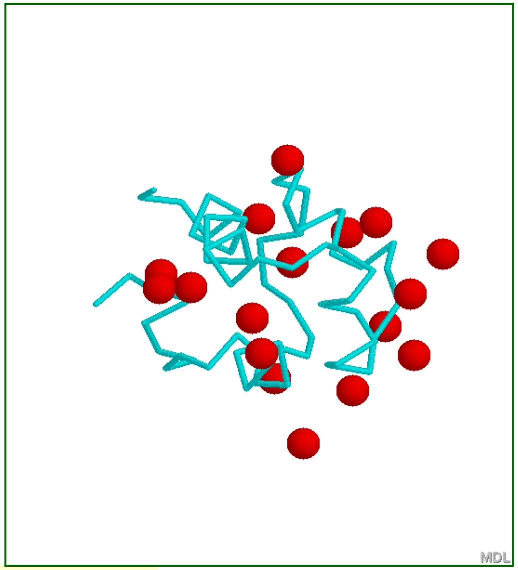
- **α-helix**
  - Main chain with peptide bonds
  - Side chains project outward from helix
  - Stability provided by H-bonds between CO and NH groups of residues 4 locations away.
- **β-strand**
  - Stability provided by H-bonds with one or more β-strands, forming β-sheets. Needs a β-turn.

Proteins

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





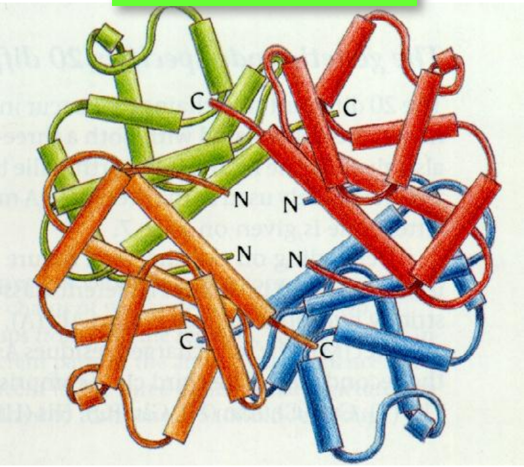


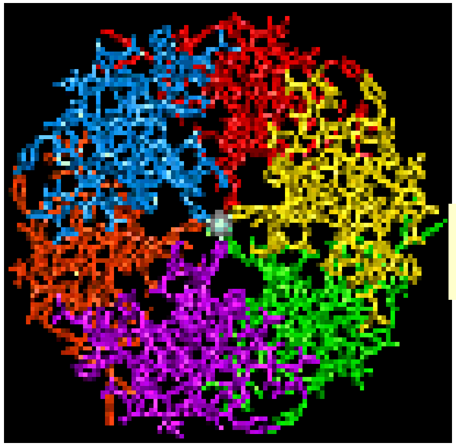
Myoglobin

Lambda Cro

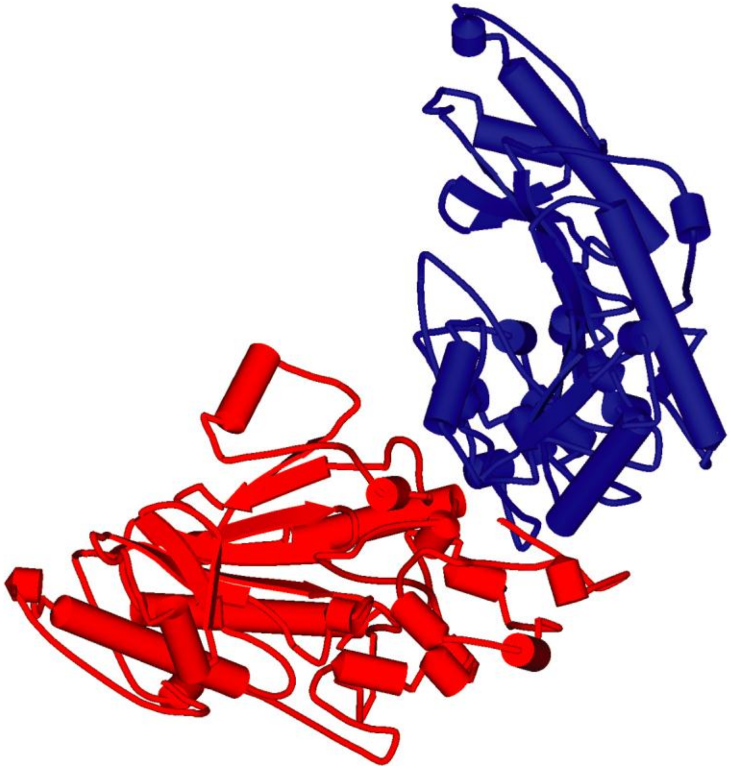
Quaternary Structures in Proteins

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


Quaternary

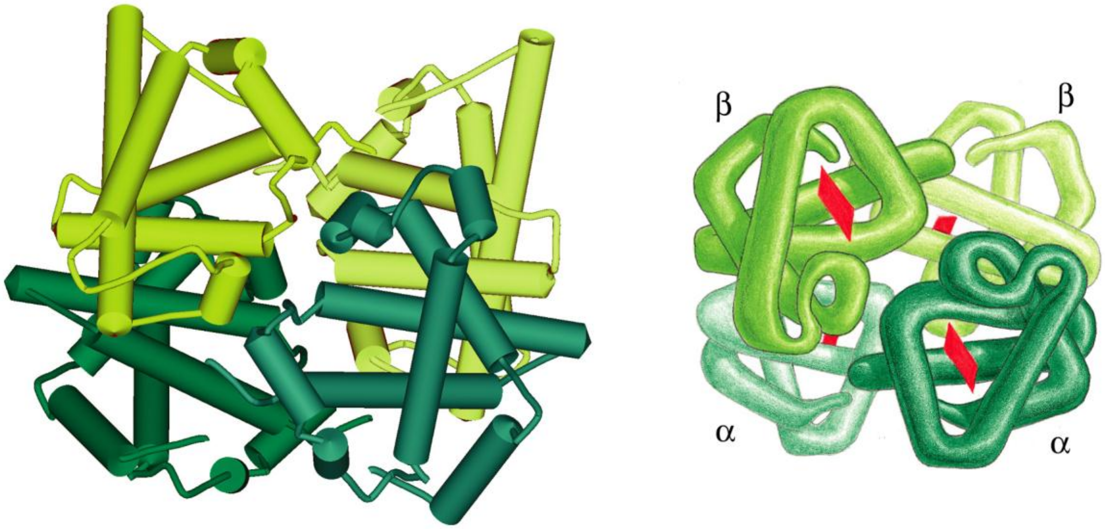

Insulin Hexamer

More quaternary structures



Muscle creatine kinase
(Homodimer)

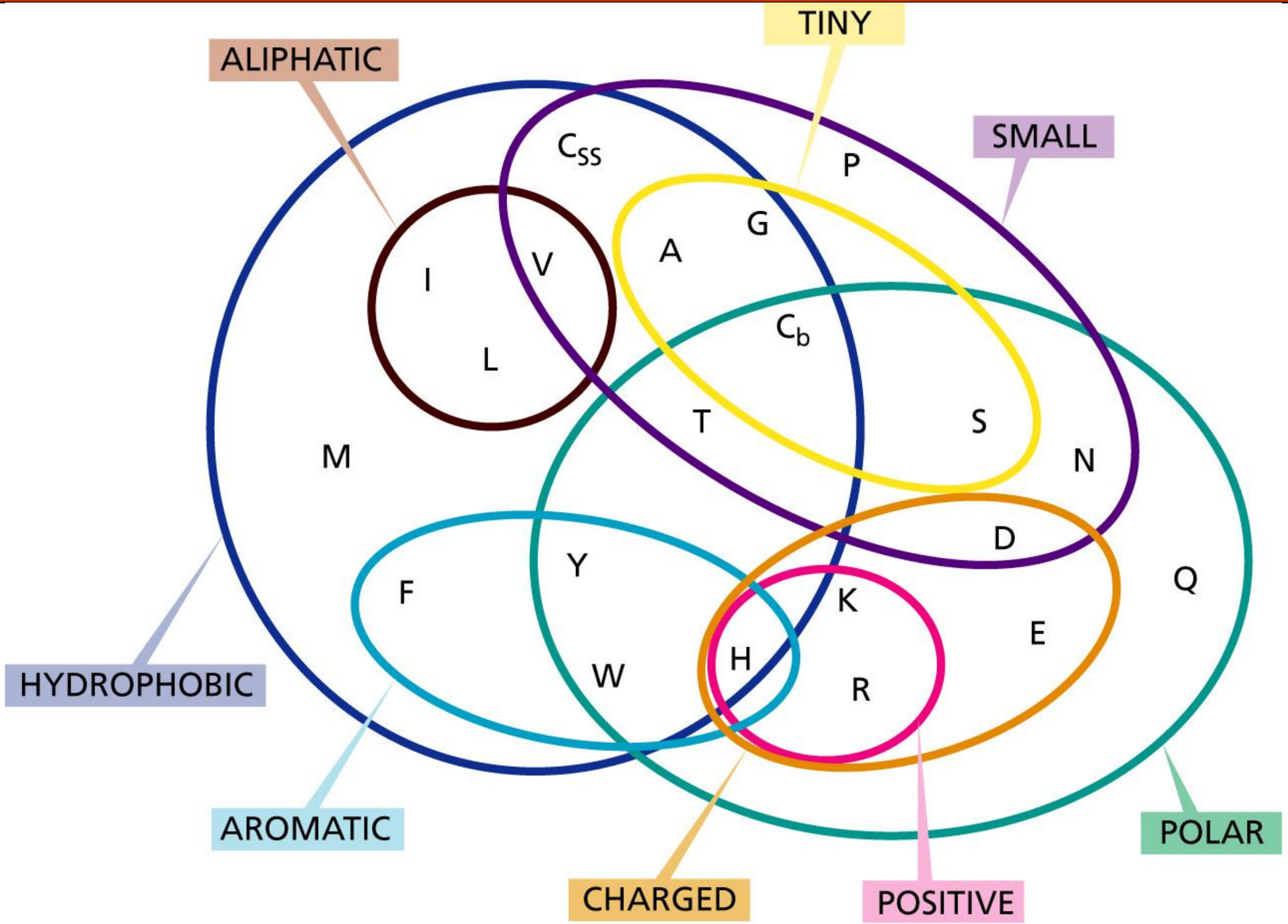
Bovine deoxyhemoglobin
(Heterotetramer)



Amino Acid Types

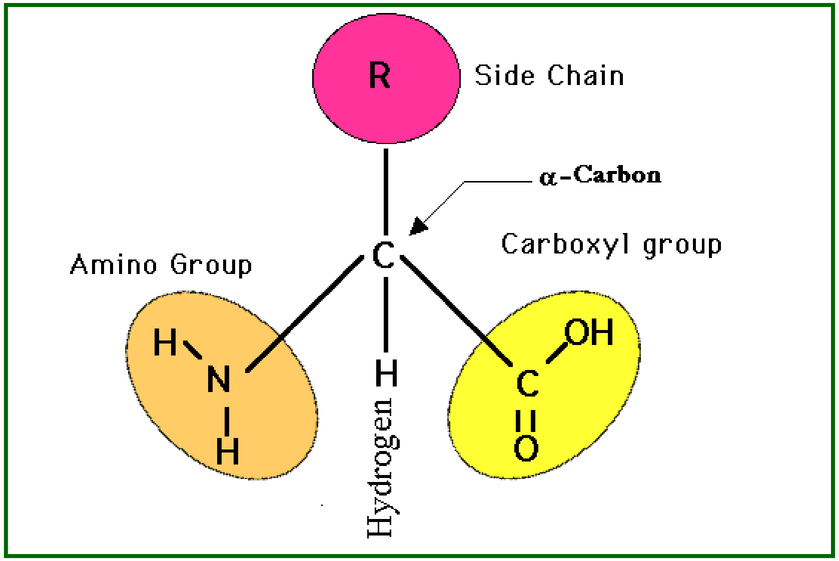
Hydrophobic	I,L,M,V,A,F,P
Charged	
● Basic	K,H,R
● Acidic	E,D
Polar	S,T,Y,H,C,N,Q,W
Small	A,S,T
Very Small	A,G
Aromatic	F,Y,W

Amino Acid Types



Structure of a single amino acid

All 3 figures are cartoons of an amino acid residue.



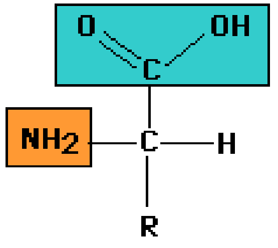
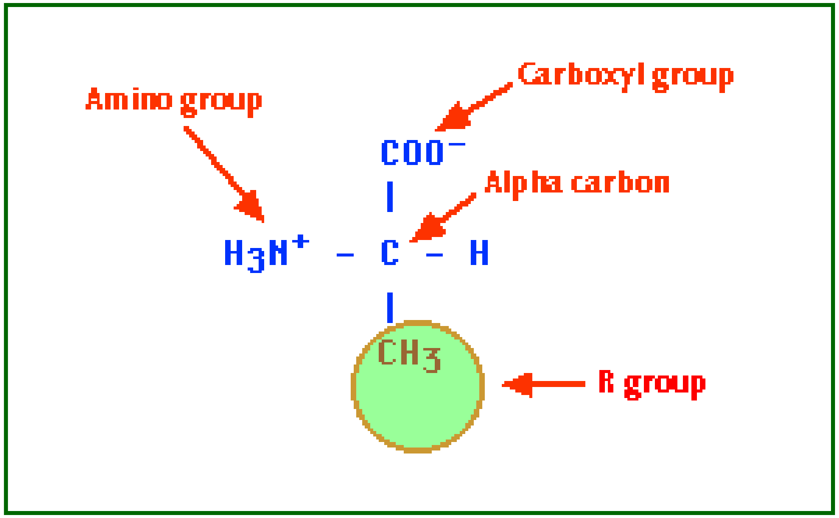
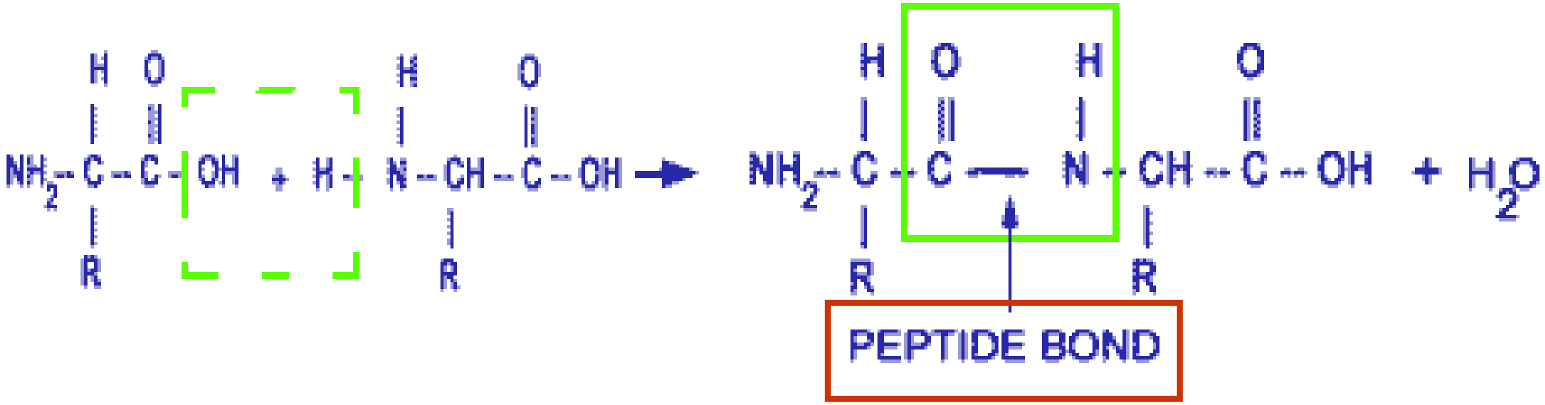


Fig. General formula for an amino acid molecule. "R" represents the variable groups that are attached to this basic molecule to make up the 20 common amino acids



Chains of amino acids



Amino acids **vs** **Amino acid residues**

Angles ϕ and ψ in the polypeptide chain

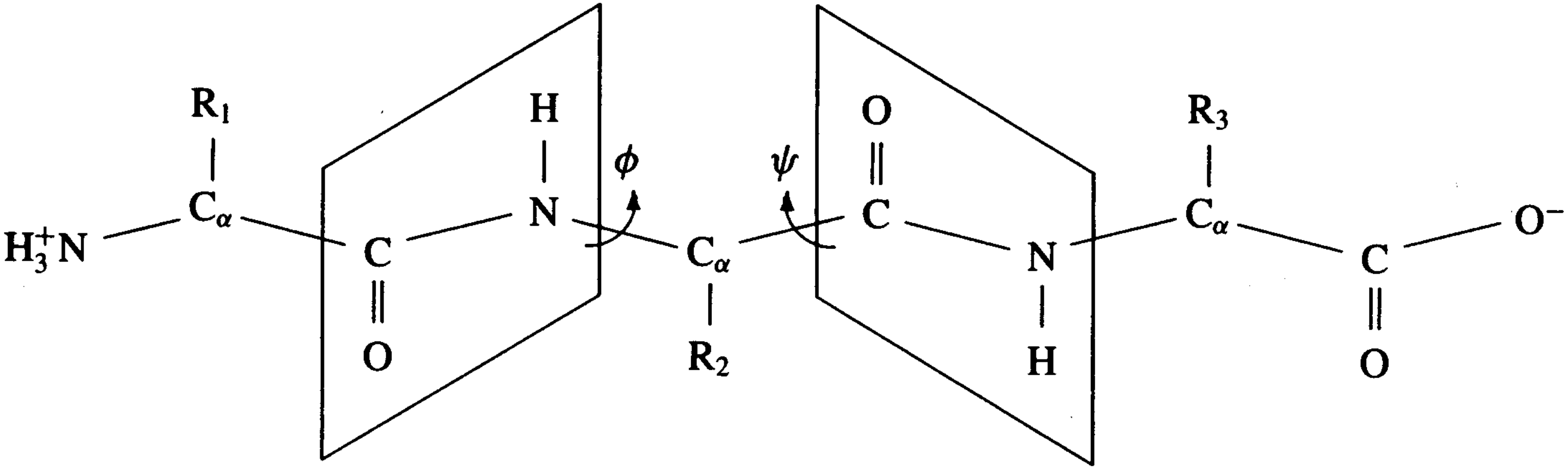


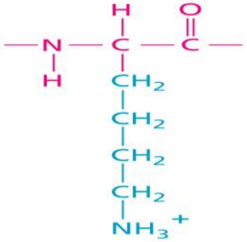
FIGURE 1.2

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.

BASIC SIDE CHAINS

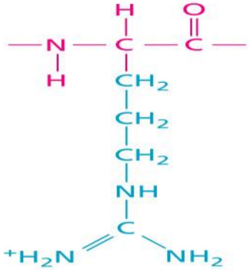
lysine

(Lys, or K)



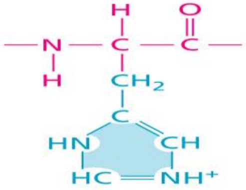
arginine

(Arg, or R)



histidine

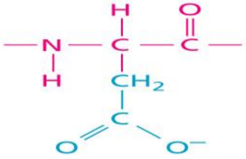
(His, or H)



ACIDIC SIDE CHAINS

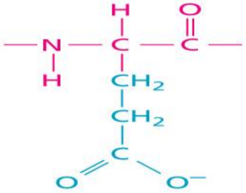
aspartic acid

(Asp, or D)



glutamic acid

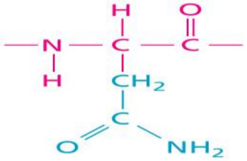
(Glu, or E)



UNCHARGED POLAR SIDE CHAINS

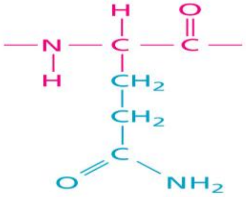
asparagine

(Asn, or N)



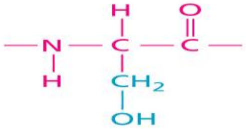
glutamine

(Gln, or Q)



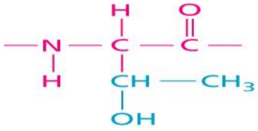
serine

(Ser, or S)



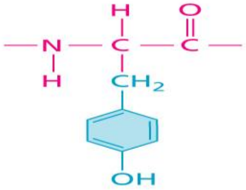
threonine

(Thr, or T)



tyrosine

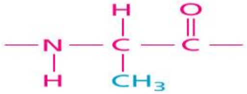
(Tyr, or Y)



NONPOLAR SIDE CHAINS

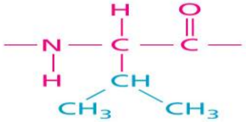
alanine

(Ala, or A)



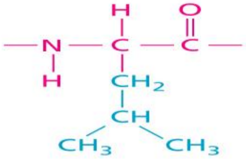
valine

(Val, or V)



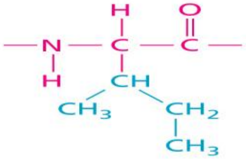
leucine

(Leu, or L)



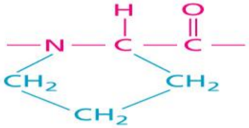
isoleucine

(Ile, or I)



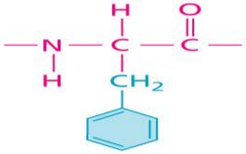
proline

(Pro, or P)



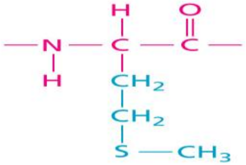
phenylalanine

(Phe, or F)



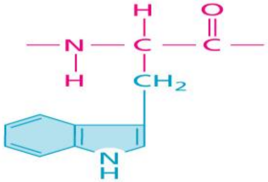
methionine

(Met, or M)



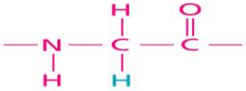
tryptophan

(Trp, or W)



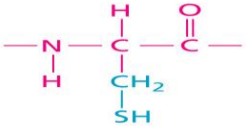
glycine

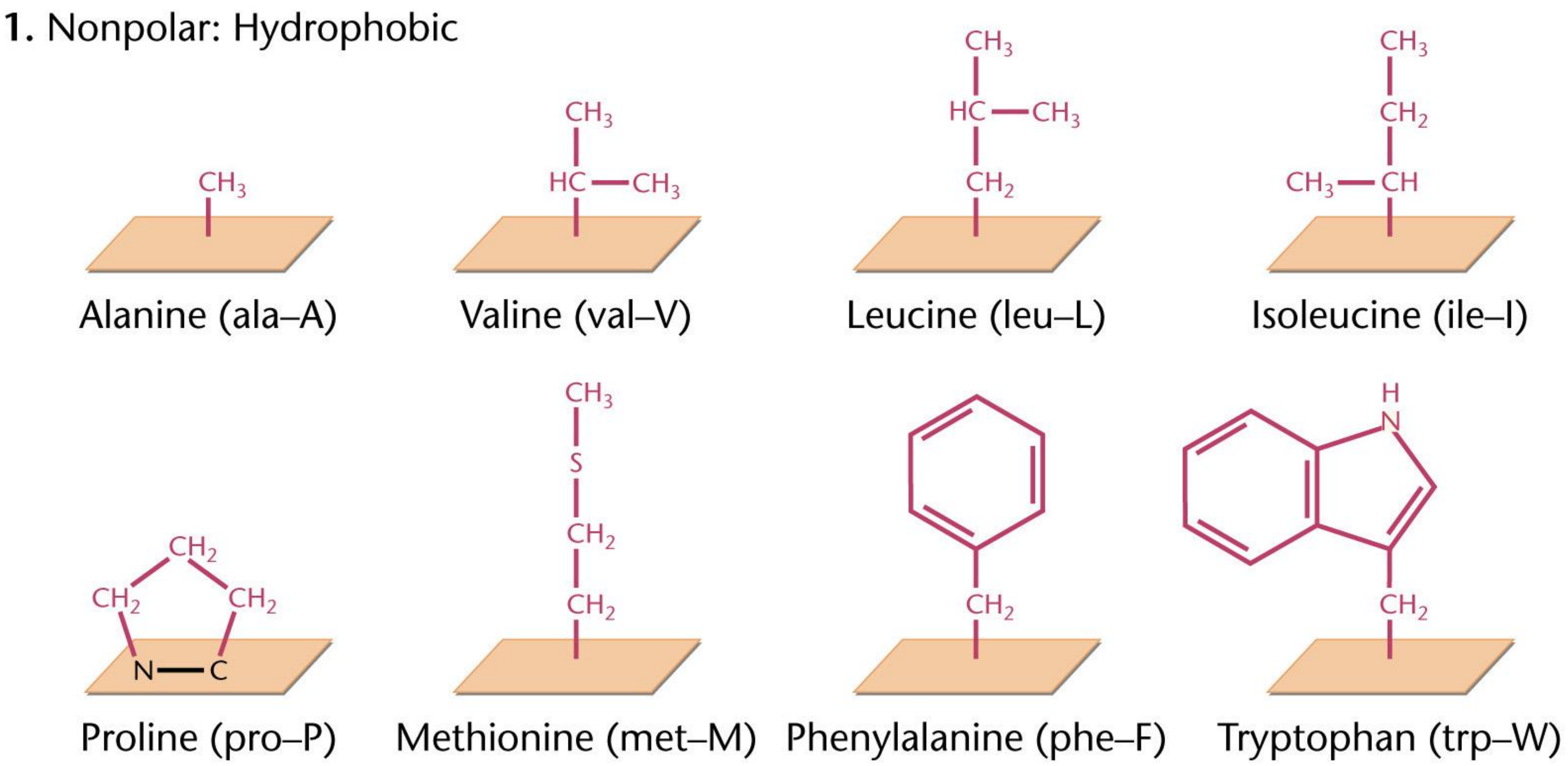
(Gly, or G)



cysteine

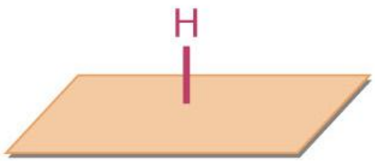
(Cys, or C)



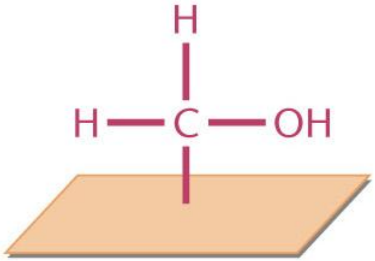


Amino Acid Structures from Klug & Cummings

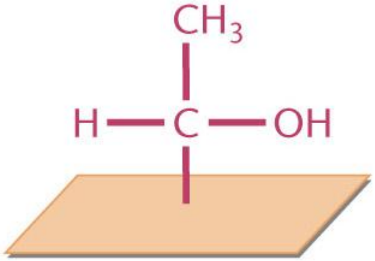
2. Polar: Hydrophilic



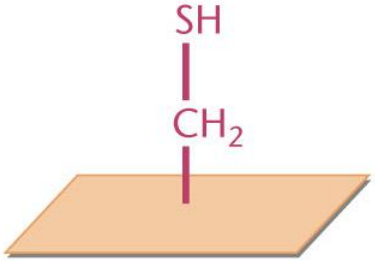
Glycine (gly–G)



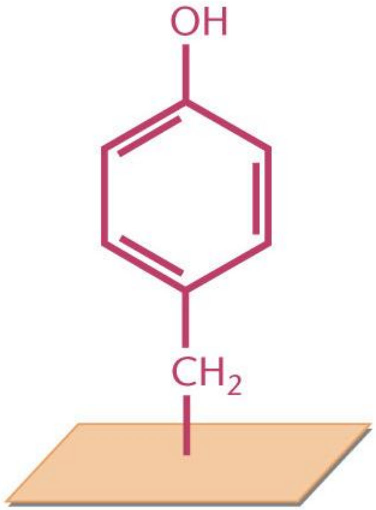
Serine (ser–S)



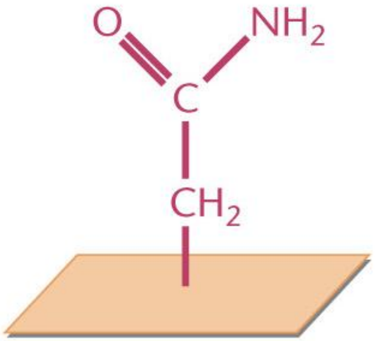
Threonine (thr–T)



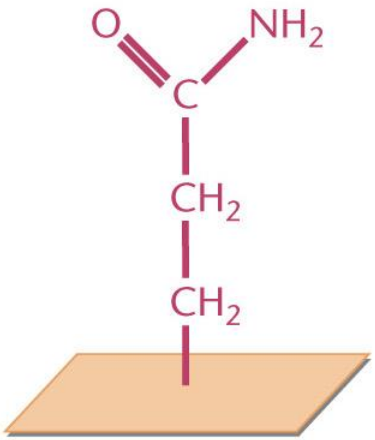
Cysteine (cys–C)



Tyrosine (tyr–Y)



Asparagine (asn–N)

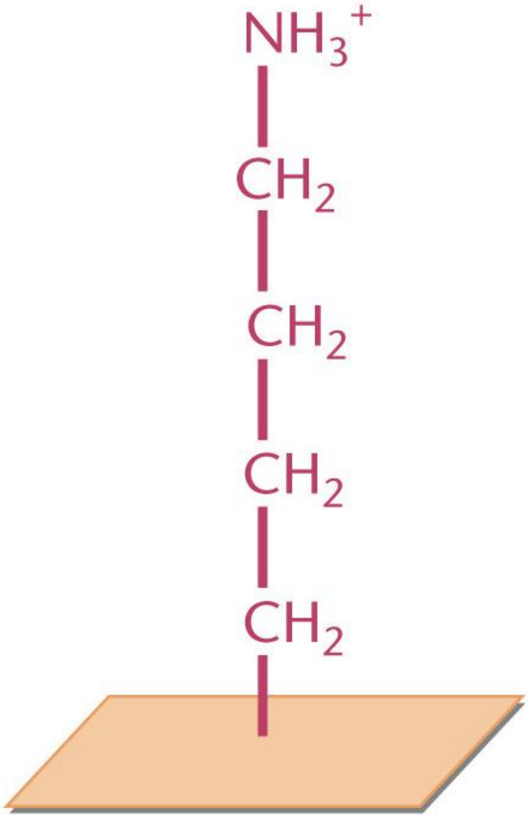


Glutamine (gln–Q)

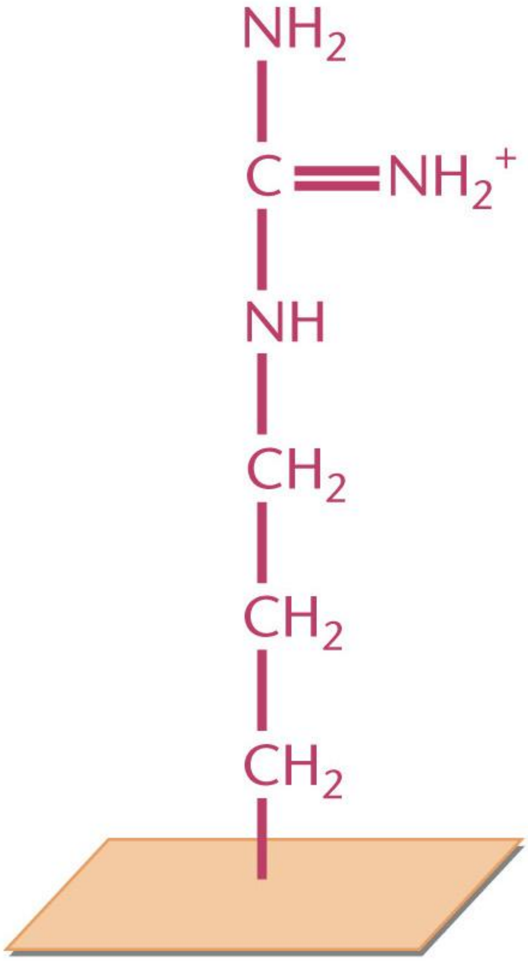
Amino Acid Structures from Klug & Cummings

3. Polar: positively charged (basic)

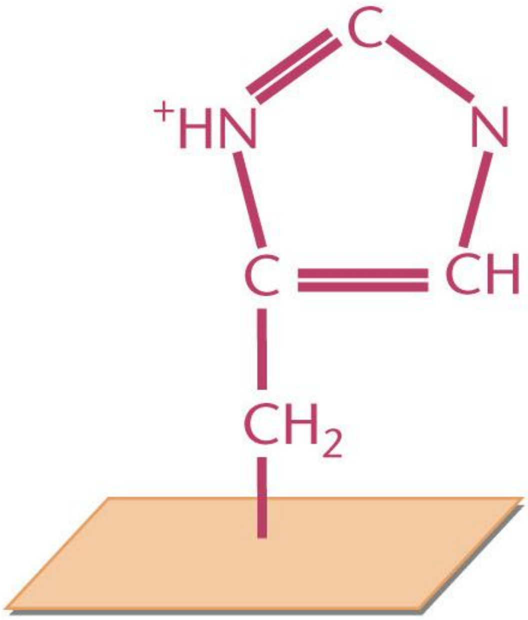
Amino Acid Structures from Klug & Cummings



Lysine (lys–K)

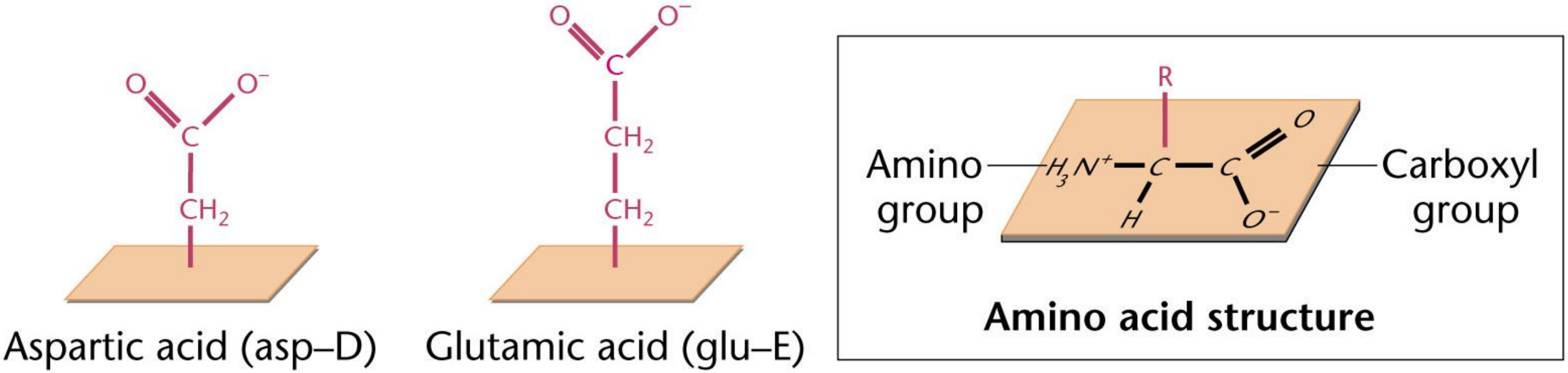


Arginine (arg–R)



Histidine (his–H)

4. Polar: negatively charged (acidic)

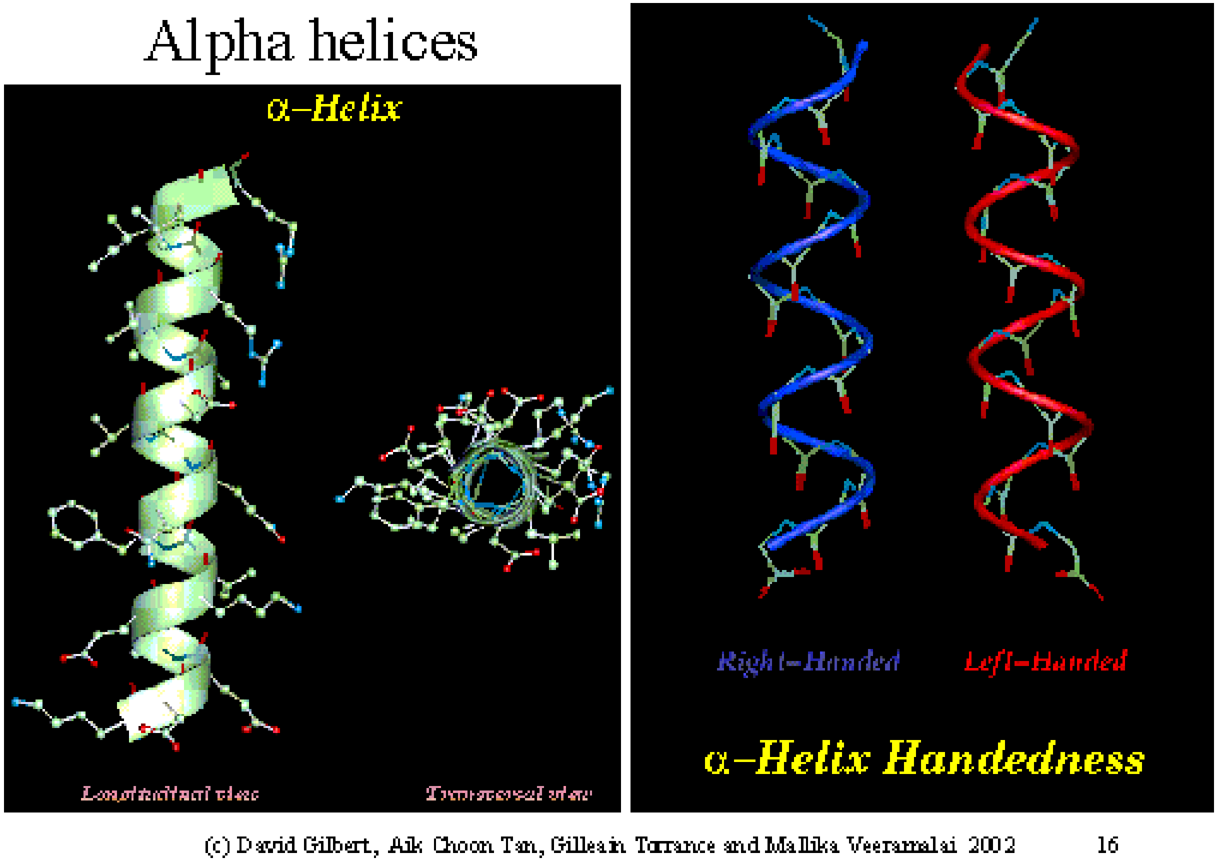


Aspartic acid (asp–D)

Glutamic acid (glu–E)

Amino acid structure

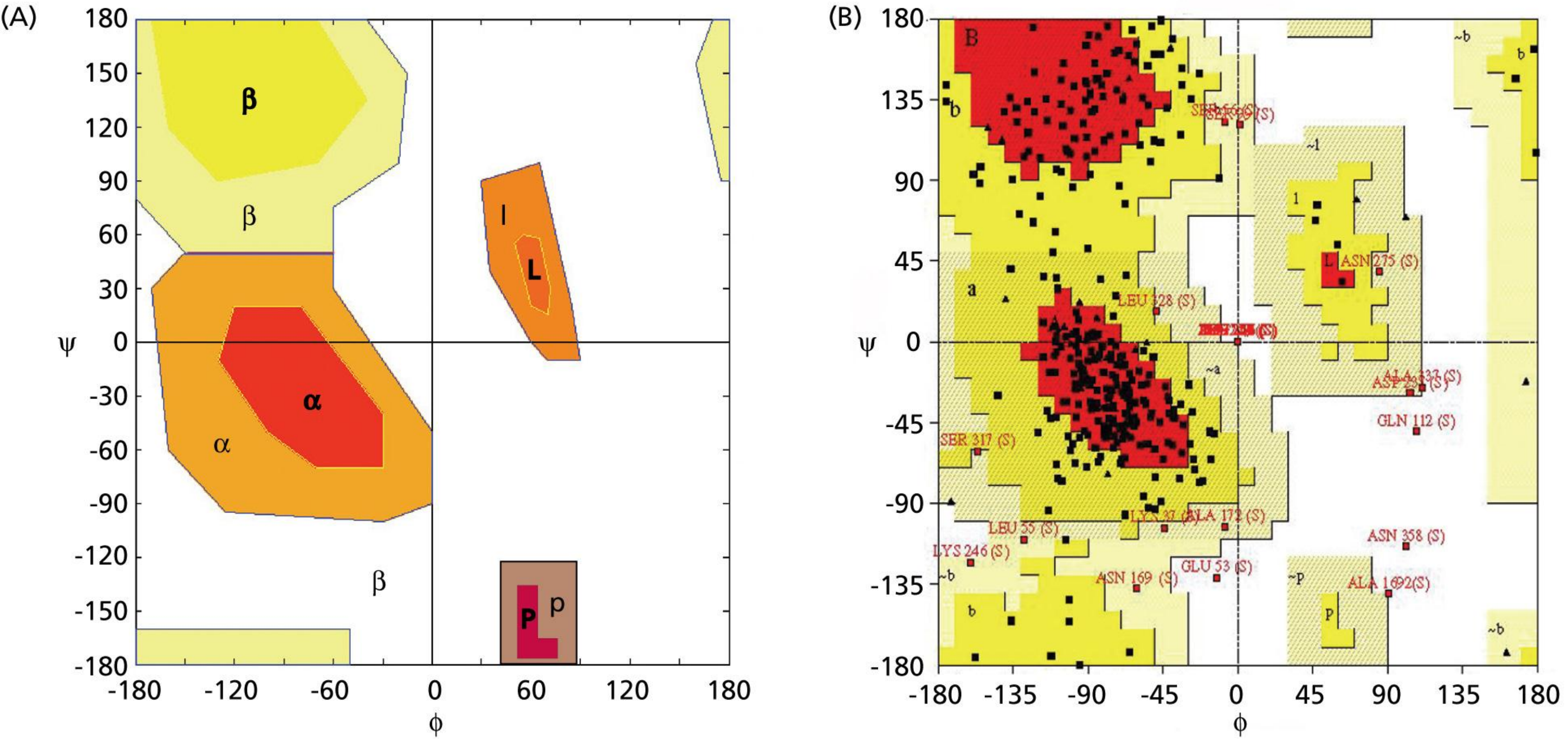
Amino Acid Structures from Klug & Cummings



Alpha helices

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

Ramachandran Plot



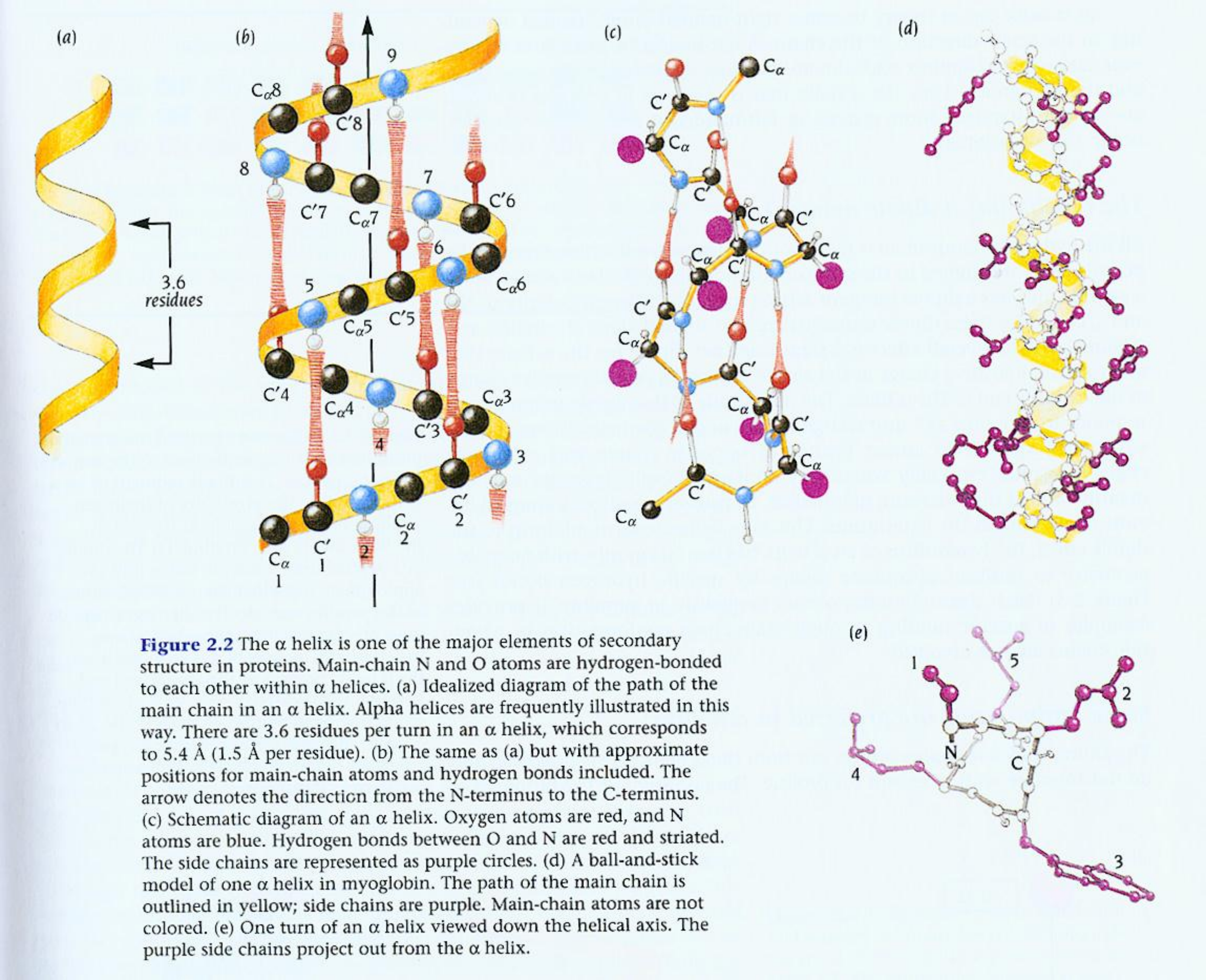
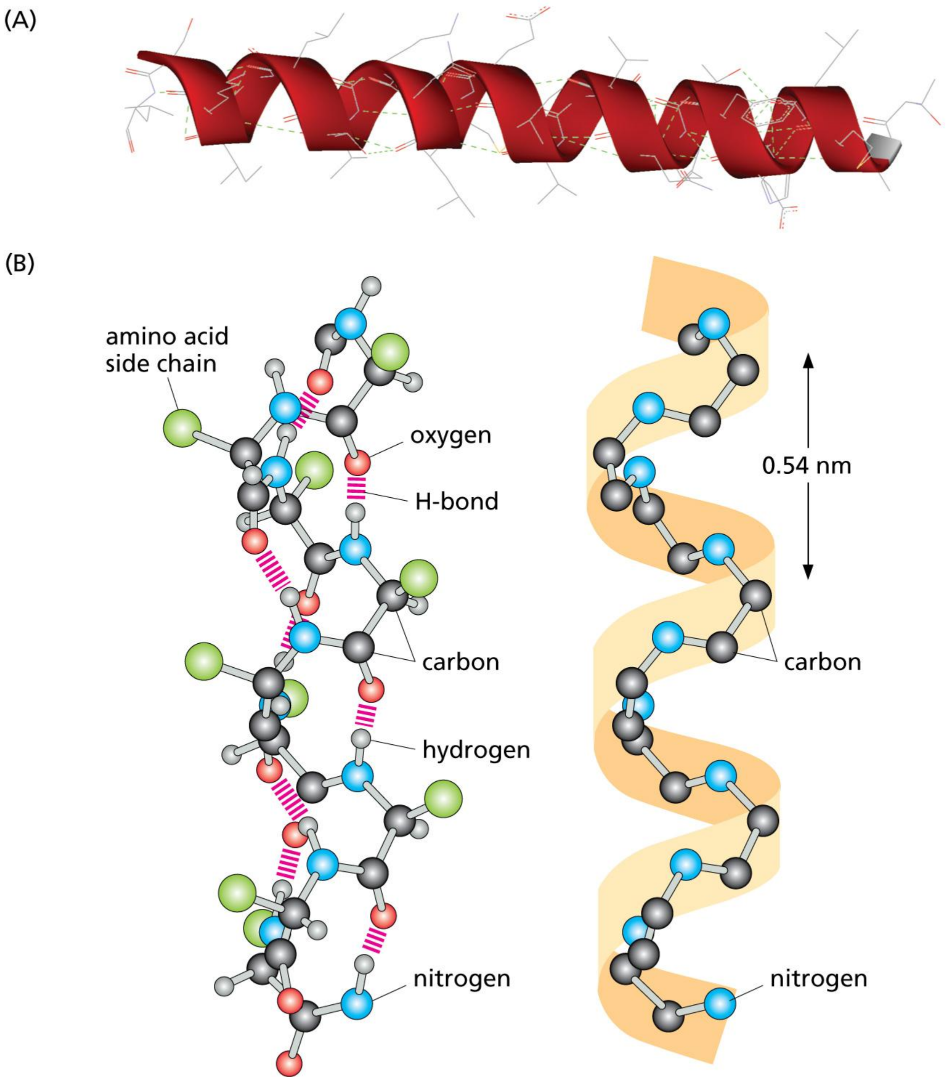
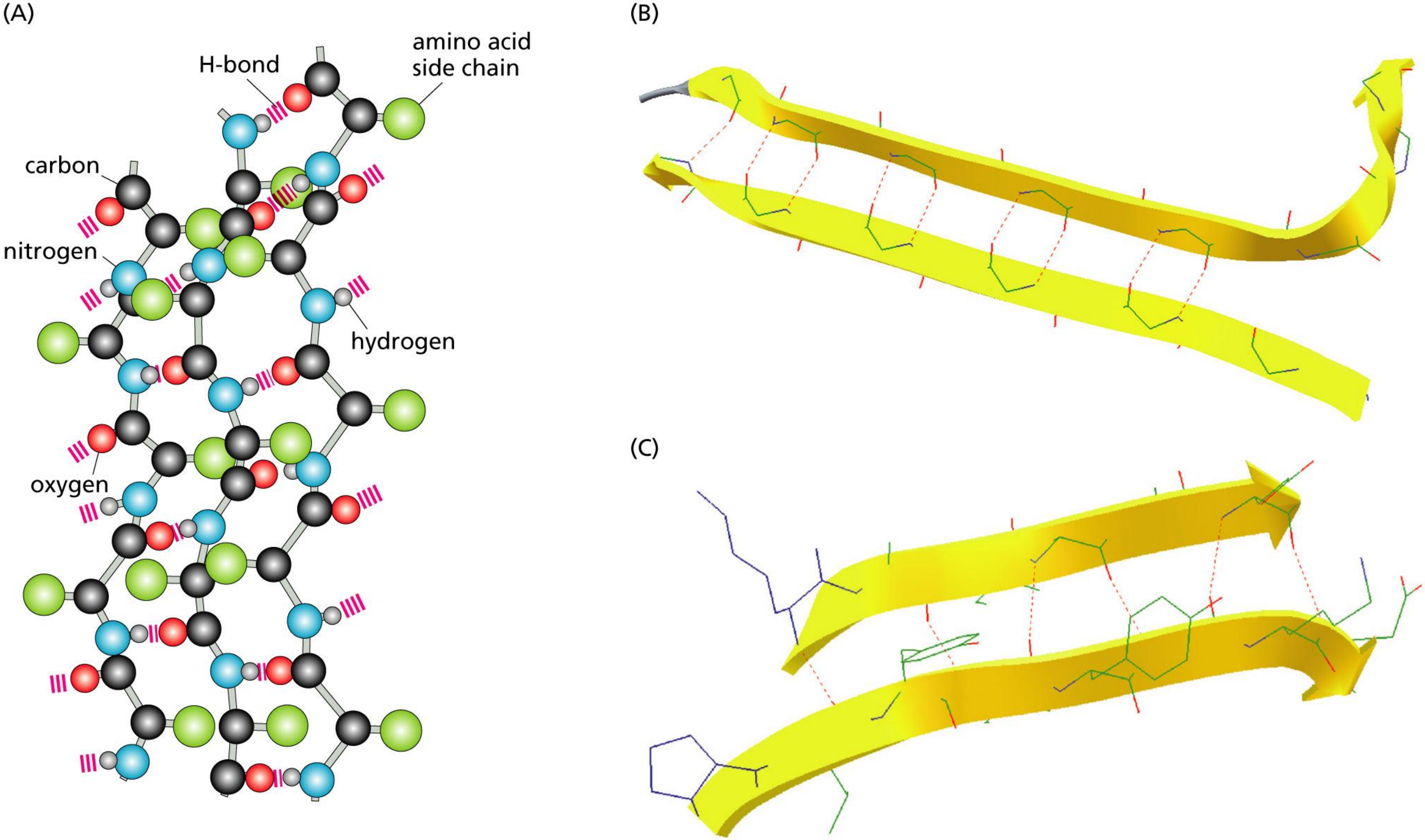


Figure 2.2 The α helix is one of the major elements of secondary structure in proteins. Main-chain N and O atoms are hydrogen-bonded to each other within α helices. (a) Idealized diagram of the path of the main chain in an α helix. Alpha helices are frequently illustrated in this way. There are 3.6 residues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

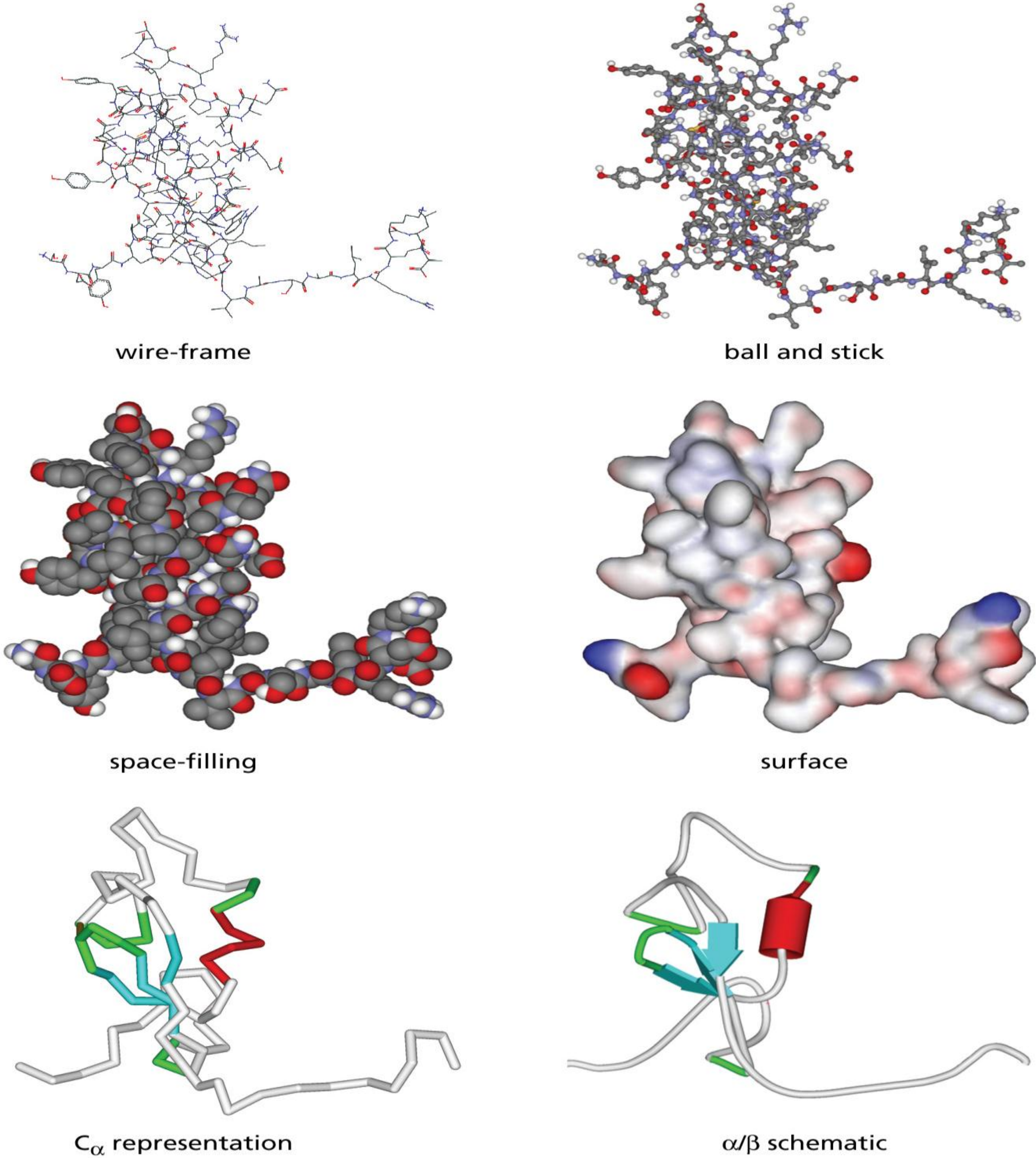
Alpha Helix


(A)
(B)
amino acid
side chain
oxygen
H-bond
carbon
hydrogen
nitrogen
0.54 nm
carbon
nitrogen

Beta Strands and Sheets



Molecular Representations



wire-frame

ball and stick

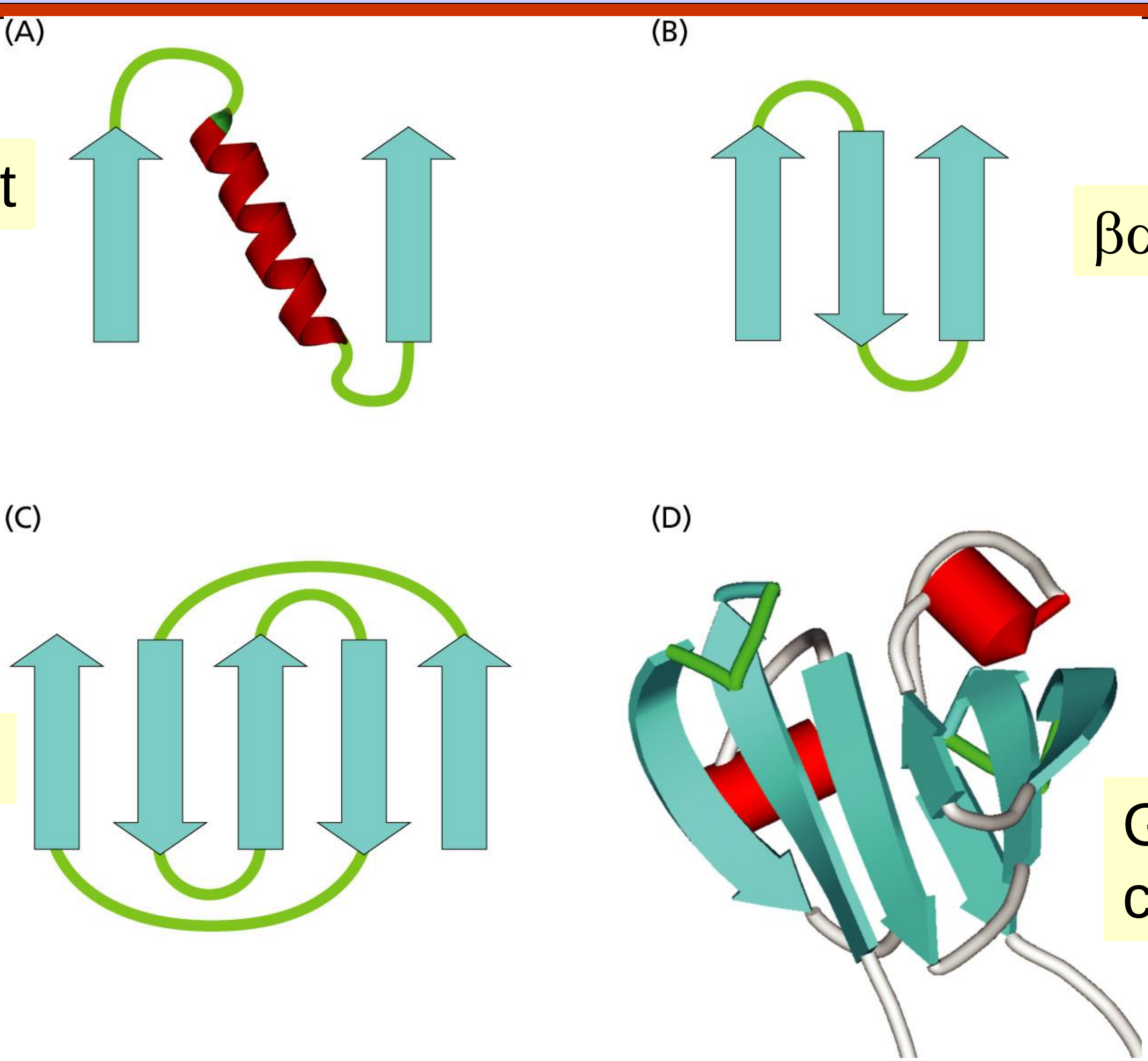
space-filling

surface

C_α representation

α/β schematic

Supersecondary structures



βαβ repeat

βαβ-meander

Greek Key

Gamma β crystallin

Secondary Structure Prediction Software

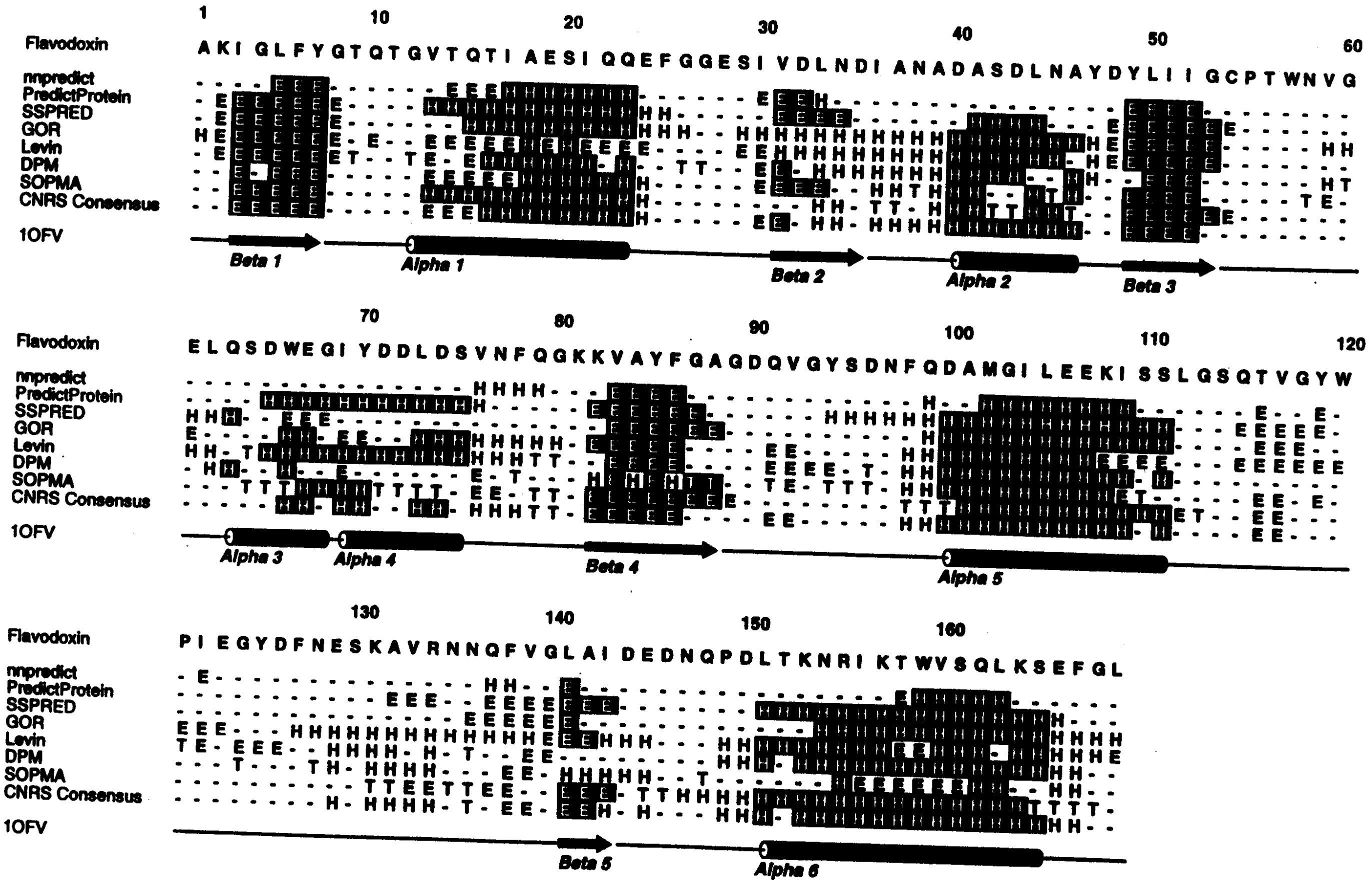


Figure 11.3 Comparison of secondary structure predictions by various methods. The sequence of flavodoxin, an α/β protein, was used as the query and is shown on the first line of the alignment. For each prediction, H denotes an α helix, E a β strand, T a β turn; all other positions are assumed to be random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

Recent Ones:

GOR V

PREDATOR

Zpred

PROF

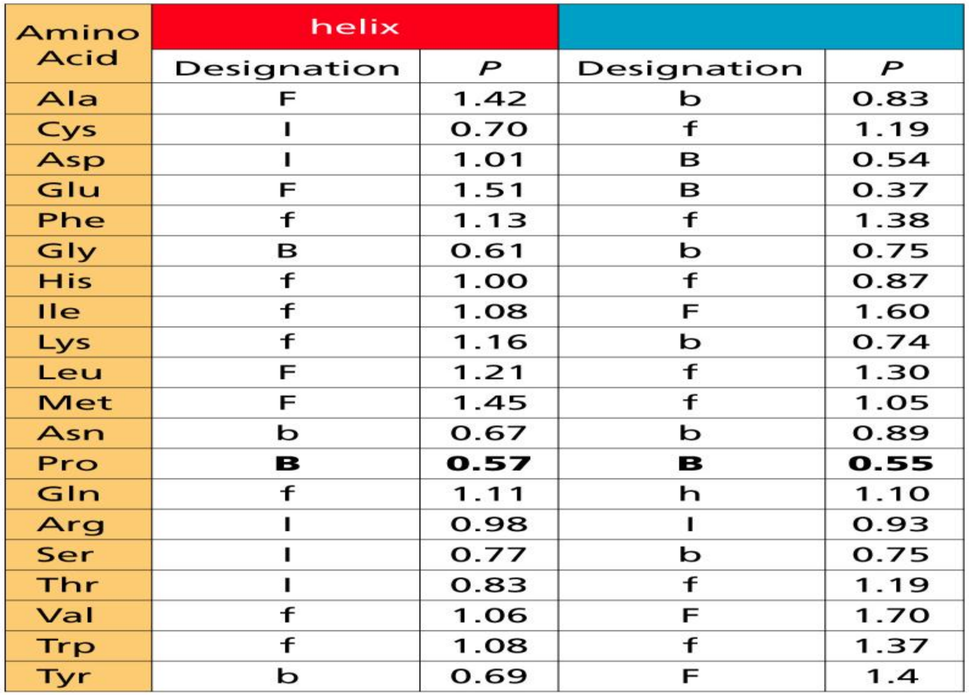
NNSSP

PHD

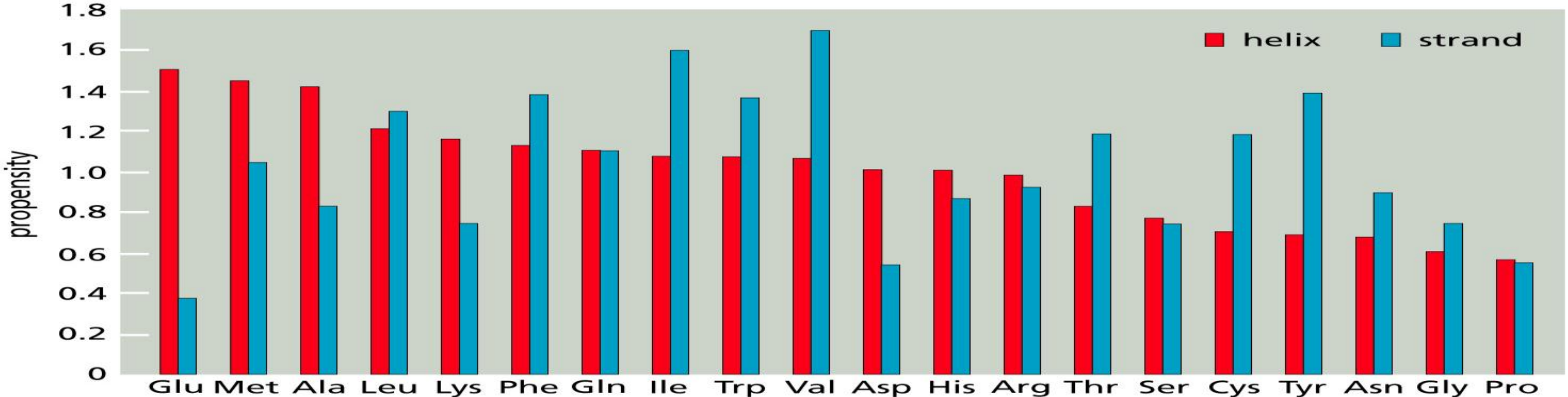
PSIPRED

Jnet

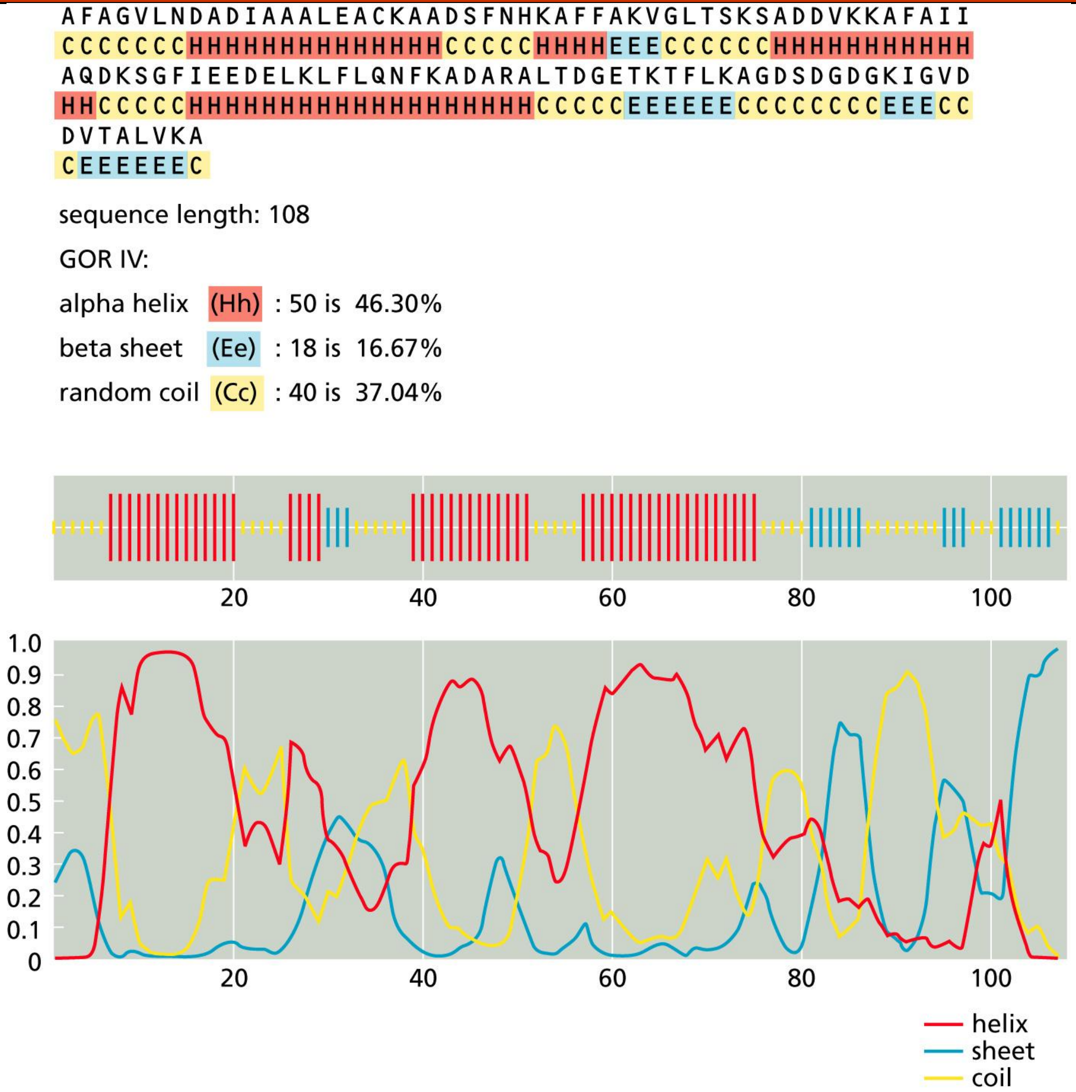
Chou & Fasman Propensities



Amino Acid	helix			
	Designation	*P*	Designation	*P*
Ala	F	1.42	b	0.83
Cys	I	0.70	f	1.19
Asp	I	1.01	B	0.54
Glu	F	1.51	B	0.37
Phe	f	1.13	f	1.38
Gly	B	0.61	b	0.75
His	f	1.00	f	0.87
Ile	f	1.08	F	1.60
Lys	f	1.16	b	0.74
Leu	F	1.21	f	1.30
Met	F	1.45	f	1.05
Asn	b	0.67	b	0.89
Pro	**B**	**0.57**	**B**	**0.55**
Gln	f	1.11	h	1.10
Arg	I	0.98	I	0.93
Ser	I	0.77	b	0.75
Thr	I	0.83	f	1.19
Val	f	1.06	F	1.70
Trp	f	1.08	f	1.37
Tyr	b	0.69	F	1.4



GOR IV prediction for 1bbc



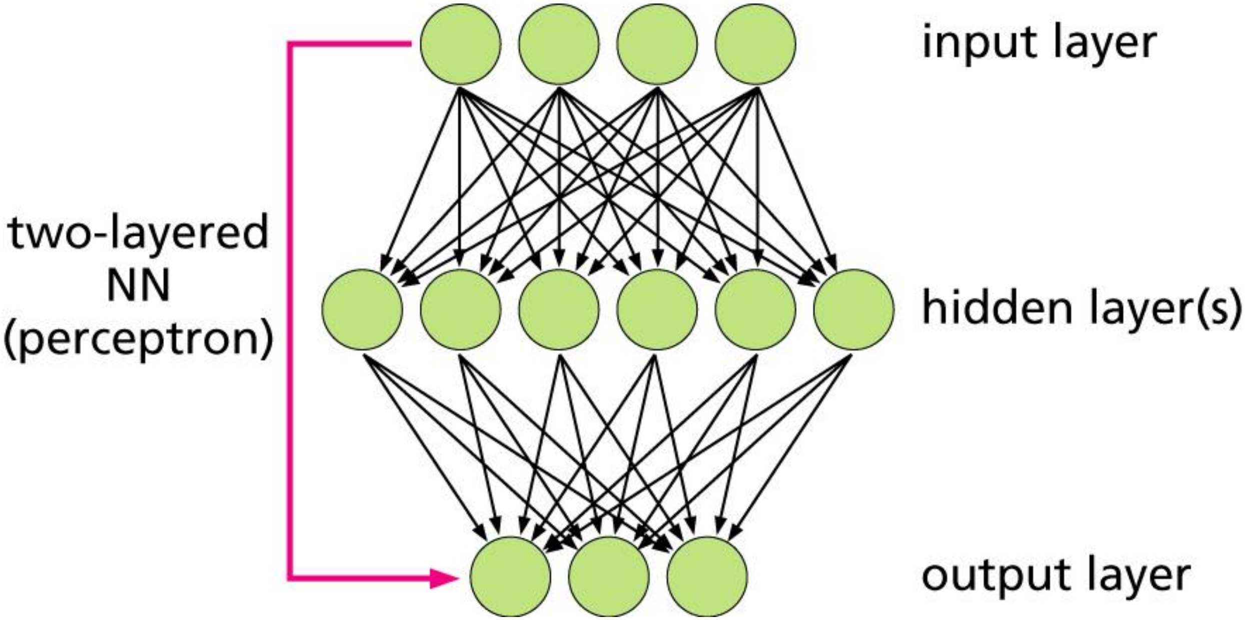
```
AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFAII
CCCCCCCHHHHHHHHHHHHHHCCCCCHHHHEEECCCCCCHHHHHHHHHHH
AQDKSGFIEEDELKLFLQNFKADARALTDGETKTFLKAGDSDGDGKIGVD
HHCCCCCHHHHHHHHHHHHHHHHHHHCCCCCEEEEEECCCCCCCCEEECC
DVTALVKA
CEEEEEEC
```

sequence length: 108

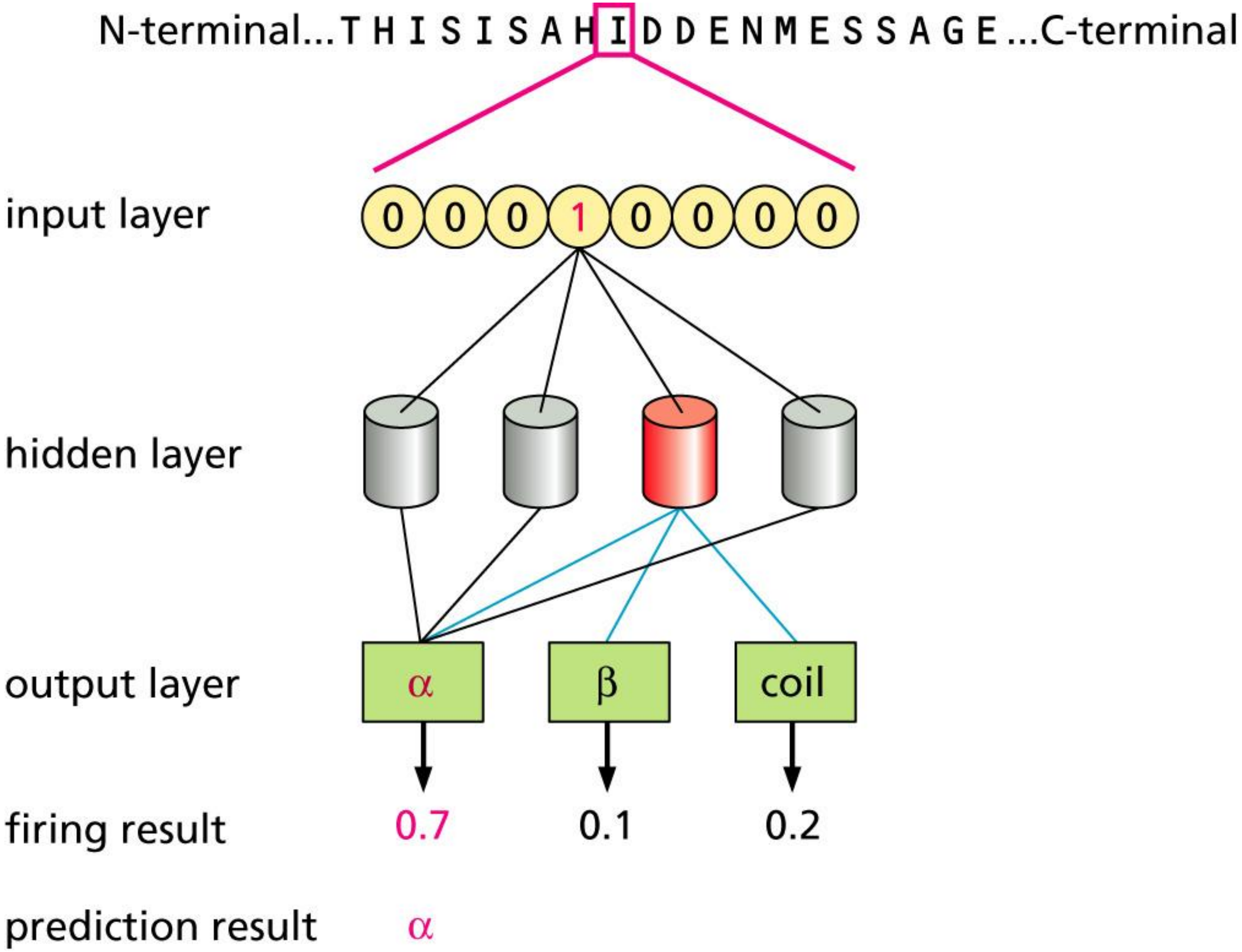
GOR IV:

alpha helix	(Hh)	: 50 is 46.30%
beta sheet	(Ee)	: 18 is 16.67%
random coil	(Cc)	: 40 is 37.04%

Neural Networks



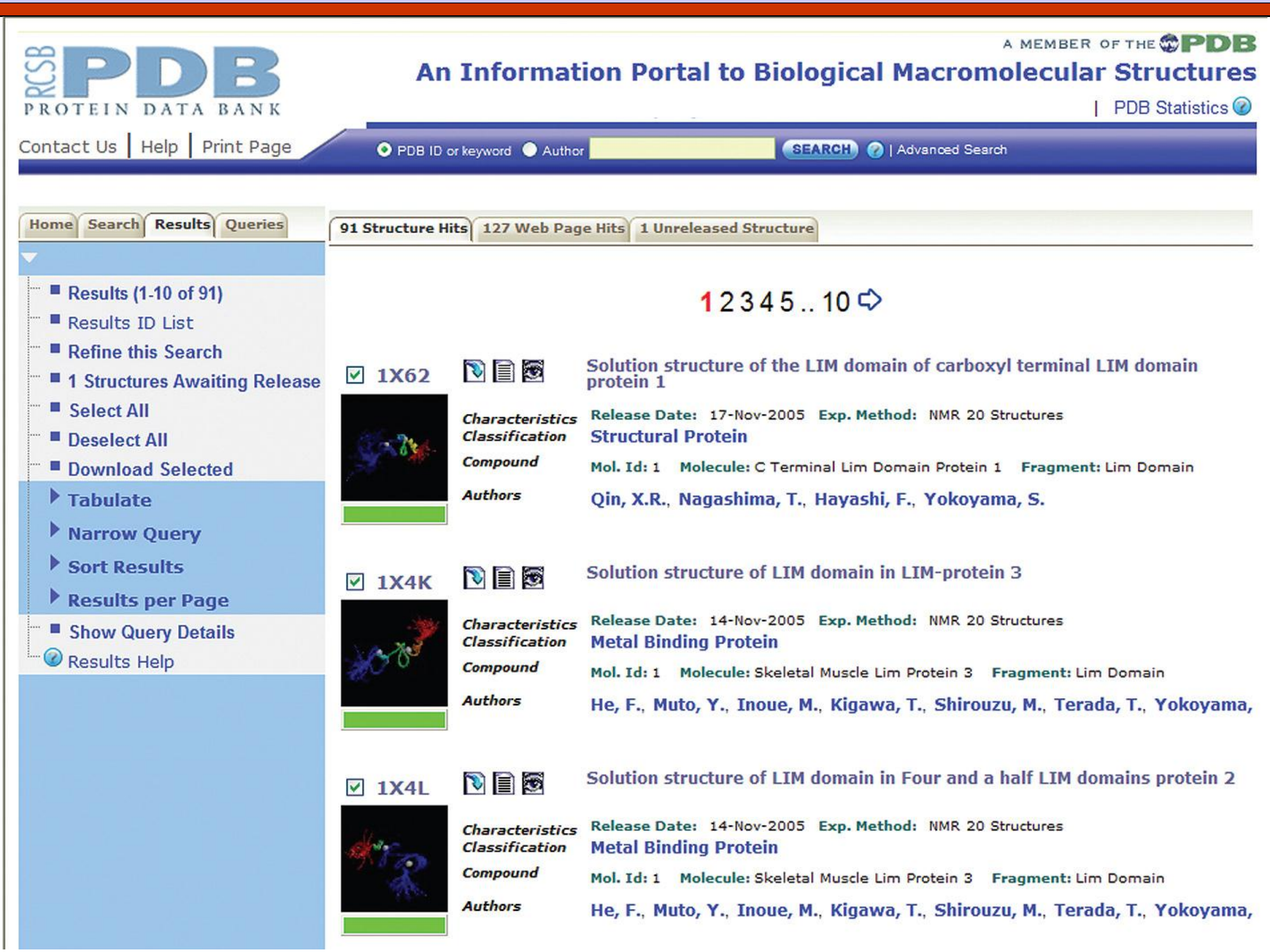
Neural Network Prediction of SS



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 [▪]

PDB Search Results



A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

| PDB Statistics

91 Structure Hits | 127 Web Page Hits | 1 Unreleased Structure

1 2 3 4 5 .. 10

☑ 1X62

Solution structure of the LIM domain of carboxyl terminal LIM domain protein 1

Characteristics Release Date: 17-Nov-2005 Exp. Method: NMR 20 Structures

Classification Structural Protein

Compound Mol. Id: 1 Molecule: C Terminal Lim Domain Protein 1 Fragment: Lim Domain

Authors Qin, X.R., Nagashima, T., Hayashi, F., Yokoyama, S.

☑ 1X4K

Solution structure of LIM domain in LIM-protein 3

Characteristics Release Date: 14-Nov-2005 Exp. Method: NMR 20 Structures

Classification Metal Binding Protein

Compound Mol. Id: 1 Molecule: Skeletal Muscle Lim Protein 3 Fragment: Lim Domain

Authors He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama,

☑ 1X4L

Solution structure of LIM domain in Four and a half LIM domains protein 2

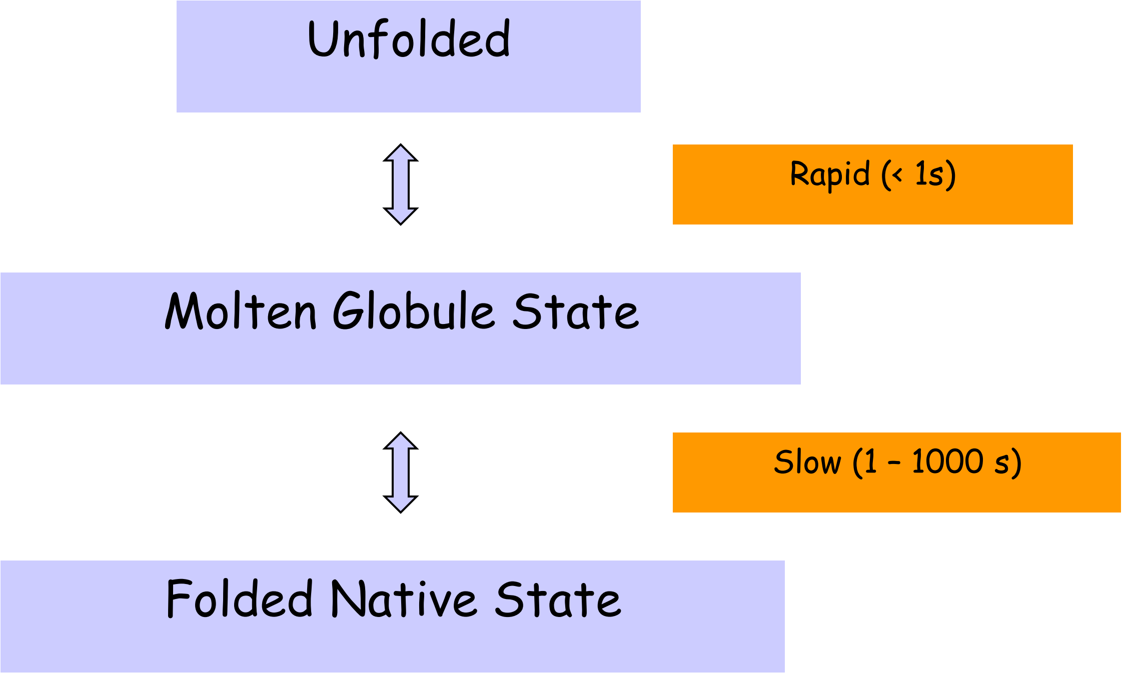
Characteristics Release Date: 14-Nov-2005 Exp. Method: NMR 20 Structures

Classification Metal Binding Protein

Compound Mol. Id: 1 Molecule: Skeletal Muscle Lim Protein 3 Fragment: Lim Domain

Authors He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama,

Protein Folding



Unfolded

Rapid (< 1s)

Molten Globule State

Slow (1 – 1000 s)

Folded Native State

- How to find minimum energy configuration?