Topic 5: the folding of biopolymers – RNA and Protein

Overview:

The main functional biomolecules in cells are polymers – DNA, RNA and proteins

For RNA and Proteins, the specific sequence of the polymer dictates its final structure

Can we predict the final structure of RNA or protein given just the sequence information?

Can we design any biomolecular structure that we want?

The world of RNA

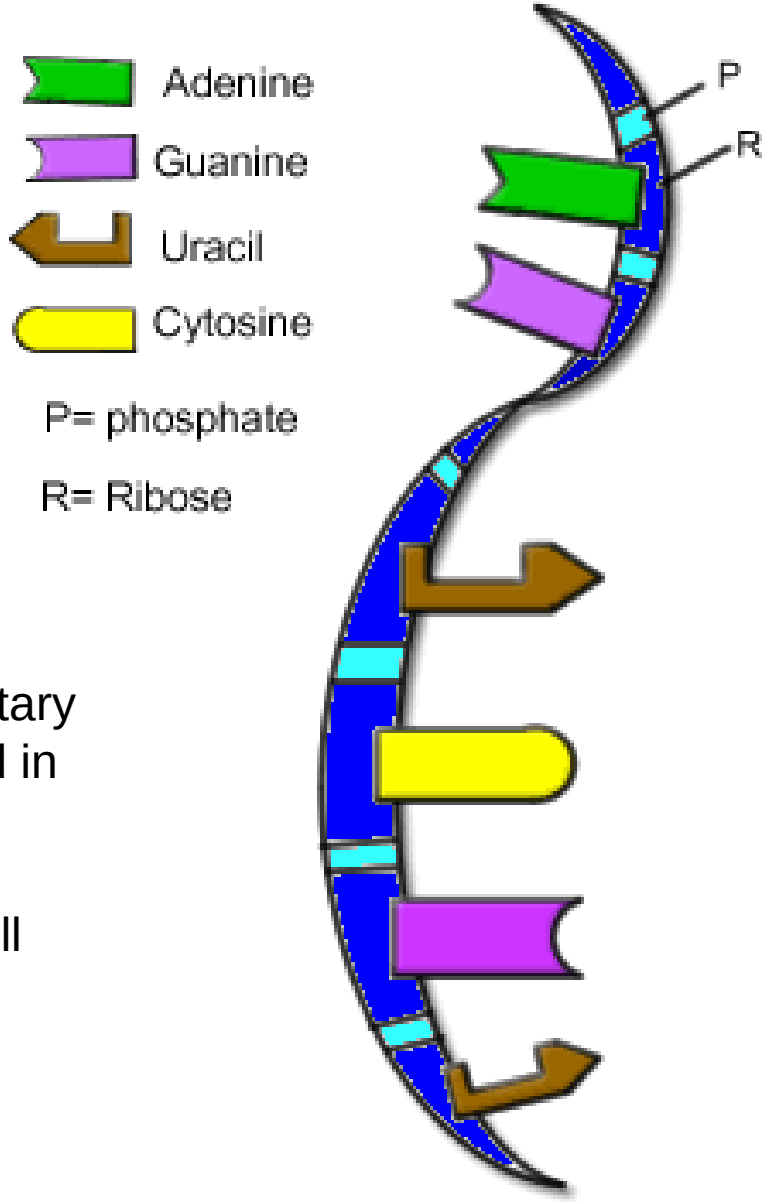
RNA is a linear polymer built from 4 possible monomers: A, G, U, C

These monomers can form complimentary interactions to form base-pairings:

A with U and C with G

Most often RNA is found as a single-stranded polymer that via base-pairing with complementary regions within its own sequence, is able to fold in on itself

RNA has a wide variety of functions that we will now explore



RNA function in cell

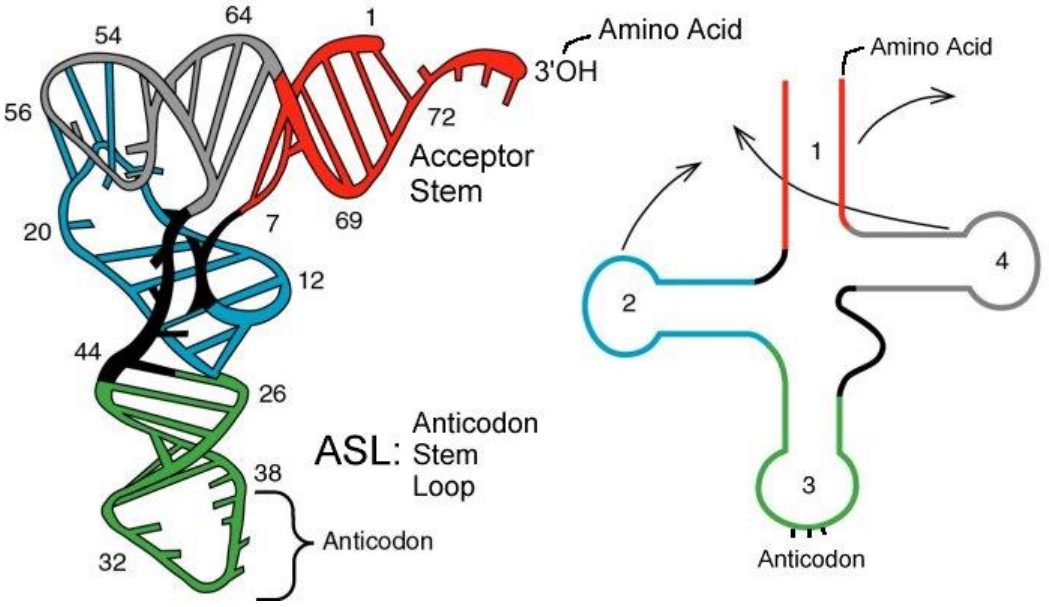
mRNA = messenger RNA

RNA's main function in the cell is to act as a messenger molecule in the process of making a protein

DNA (gene) → mRNA → Protein

tRNA = transfer RNA

these are used in translation to recognize the 64 codons. There is one tRNA for each codon, each representing one of the 20 amino acids



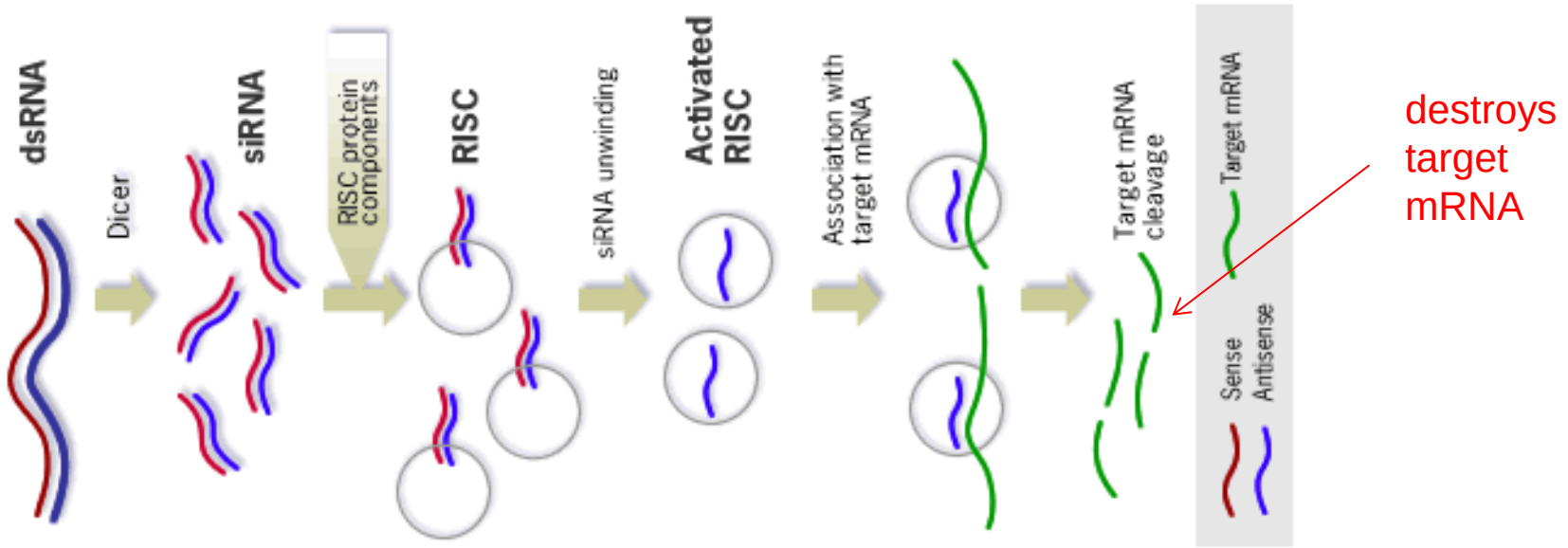
RNA function in cell

rRNA = ribosomal RNA

these are RNAs that get incorporated into the ribosome to give it part of its function.

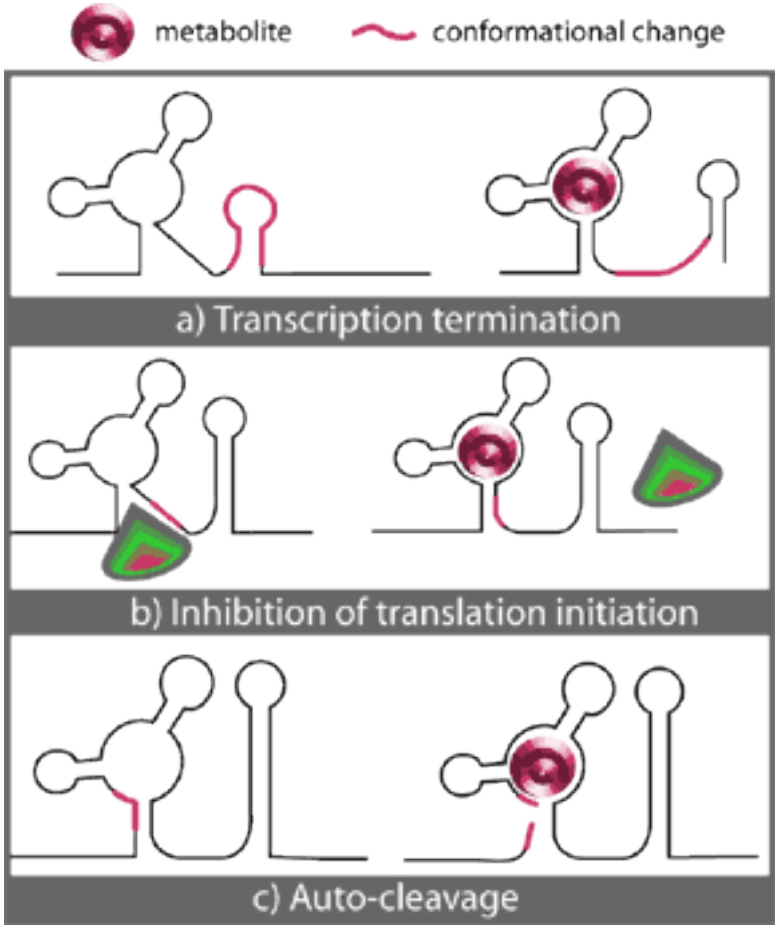
microRNA or siRNA

these are relatively recent discovered form of RNA that is used to regulate gene expression – so called RNA interference (won Nobel prize)



many developmental genes are regulated by miRNAs in your genome

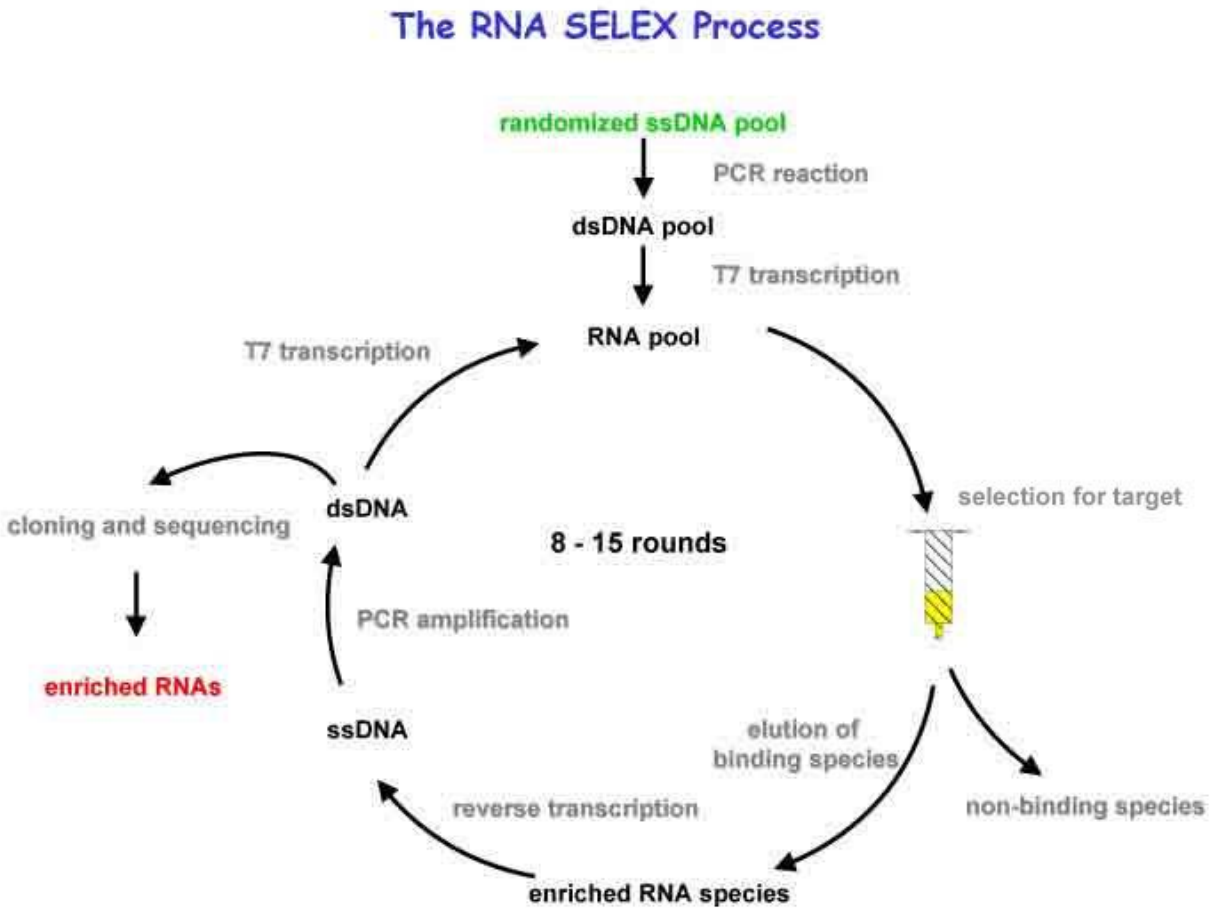
Riboswitches – many mRNA molecules can detect metabolites by binding them and changing the structure of the mRNA = regulation



RNA function in cell

Ribozymes – like enzymes (catalytic proteins) except made from RNA. Able to catalyze reactions.

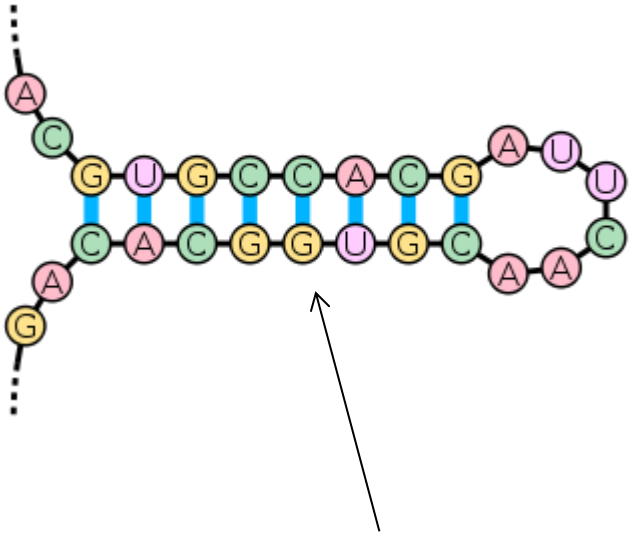
in-vitro selection experiments – can select RNA molecules out of a random library of sequences to catalyze a specific chemical reaction



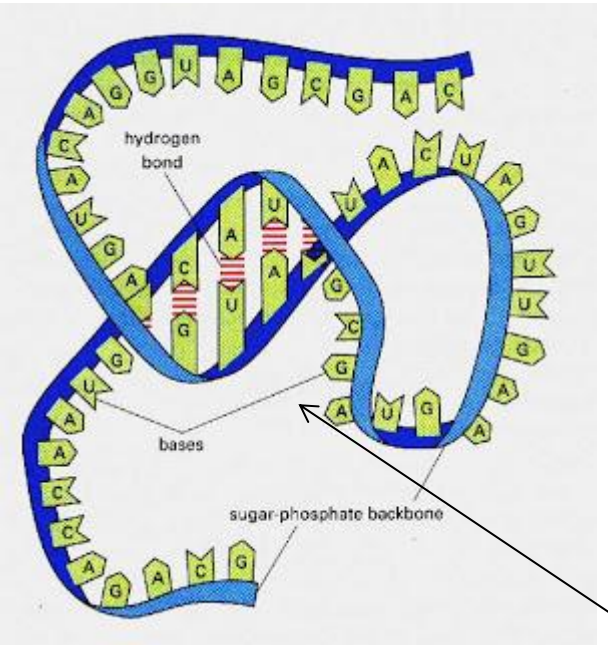
RNA structure:

A polymer of RNA first folds by forming complimentary base pairings: G = C and A = U

The simplest form of RNA structure is a hairpin loop that in 3D looks like a double helix



2ndary structure

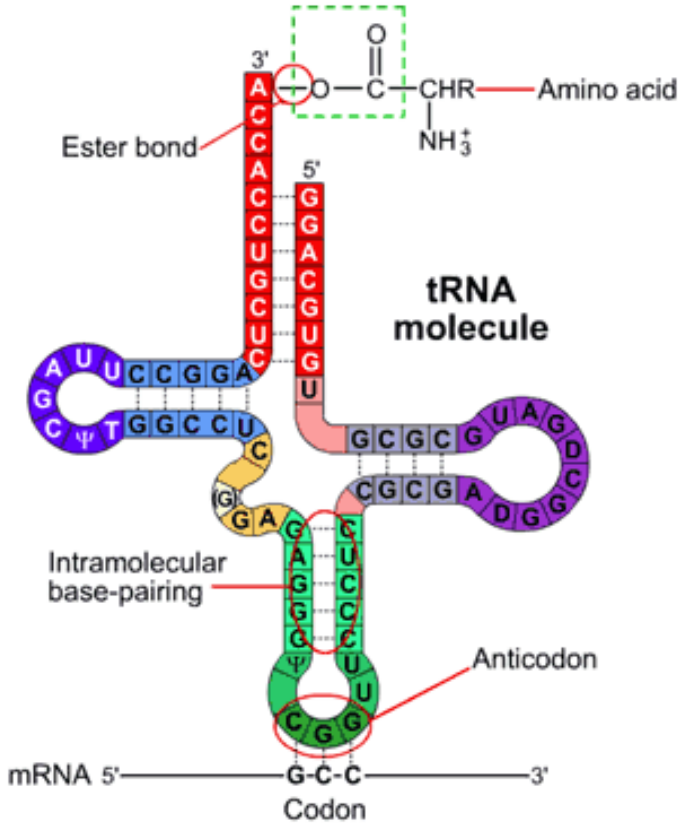


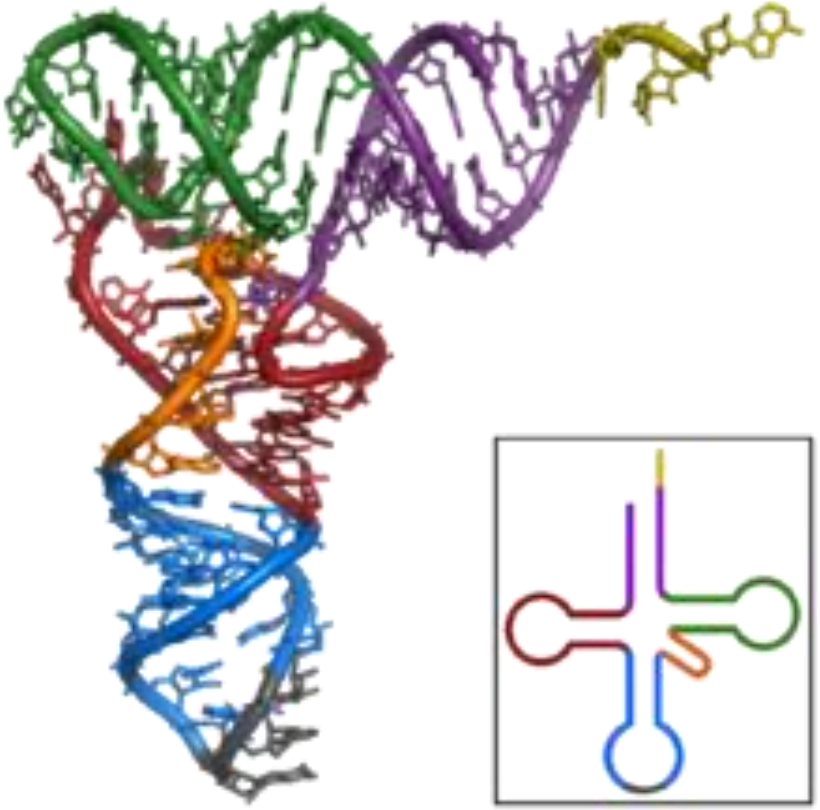
double helix

RNA structure representation

The 2ndary structure of RNA is a particular pairing of its complimentary bases

RNA tertiary structure is the final 3D fold of the polymer





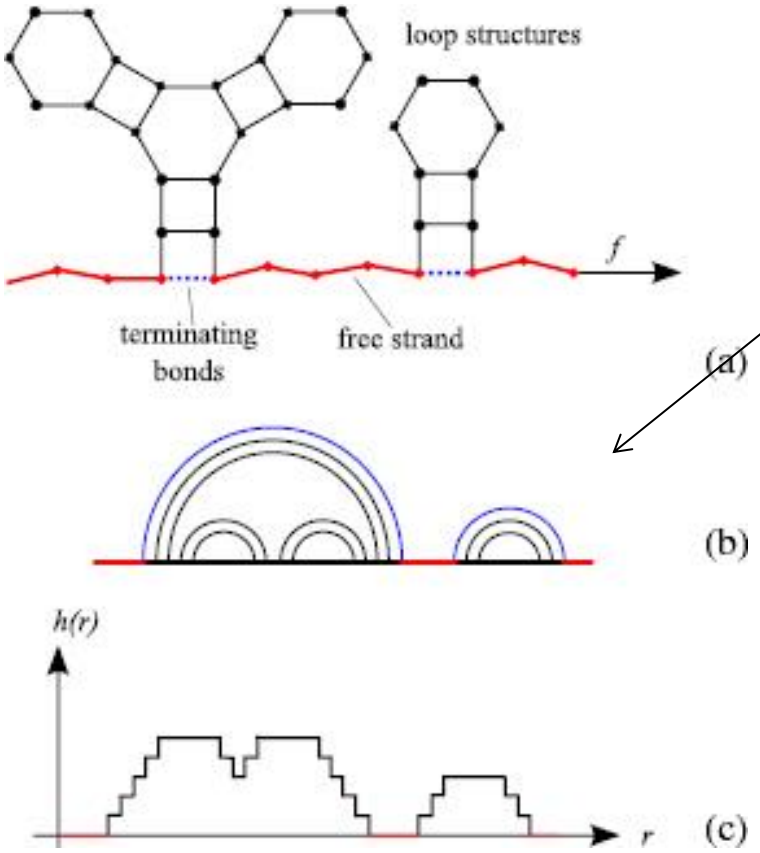
2ndary structure possesses 3 hairpin loops

Tertiary structure

RNA structure representation

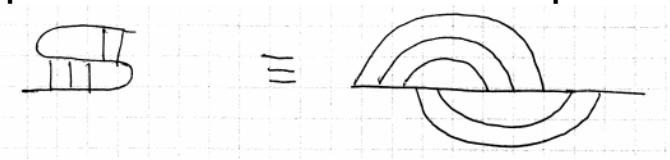
We will be interested in studying the formation of RNA secondary structure – tertiary is too hard of a problem

We need ways to represent the structure in diagrams or strings for use in calculations



Rainbow diagram – shows pairings as loops

no loops are allowed to cross = pseudo-knot



these occur but are computationally hard to deal with

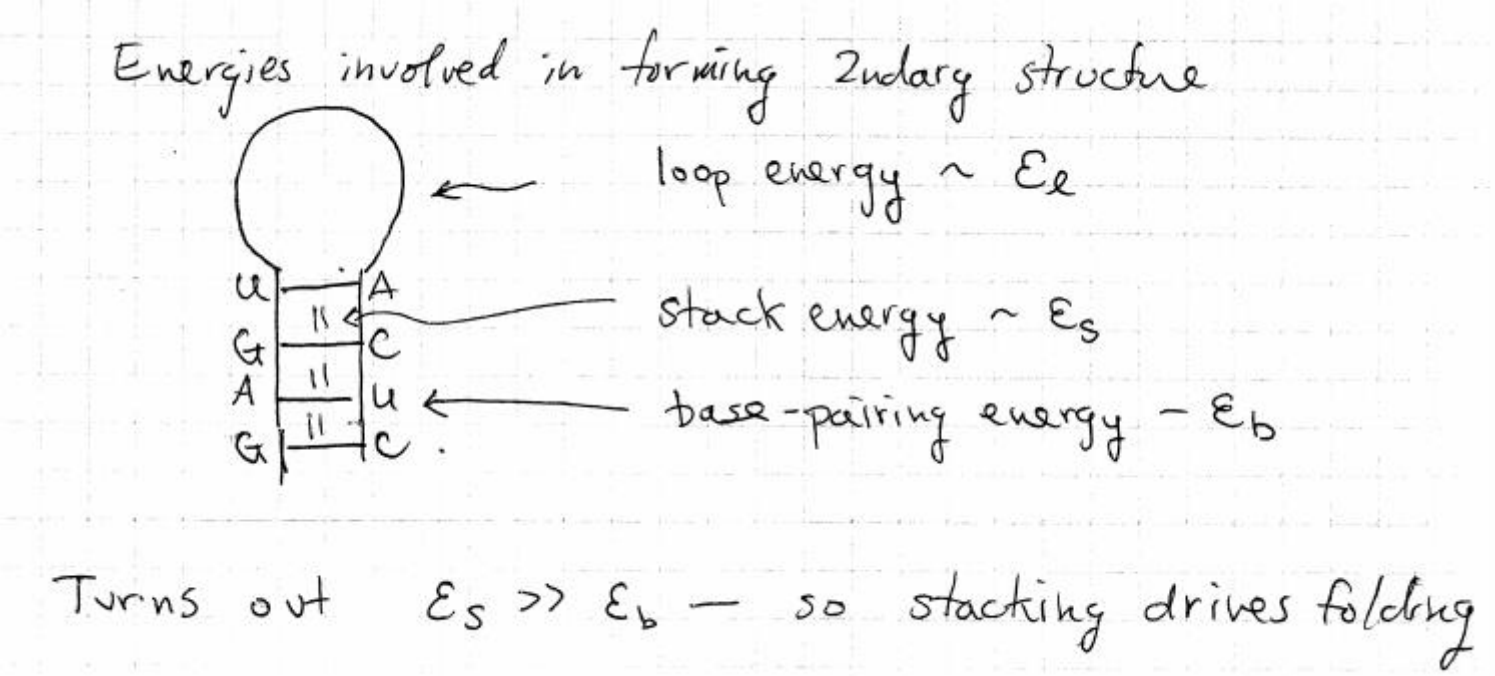
String representation:
(((((..))((..)))))…((…))…

1,1,1,1,0,0,-1,-1,1,1,0,0,-1,-1, …

nice property: sum of the string = 0

A biomolecules function depends on it's structure. Can we predict the most probable 2ndary structure of an RNA molecule by just knowing it's sequence?

Ans: yes! enumerate all the possible structures (states), each has an energy, then use Boltzmann distribution to determine the probabilities



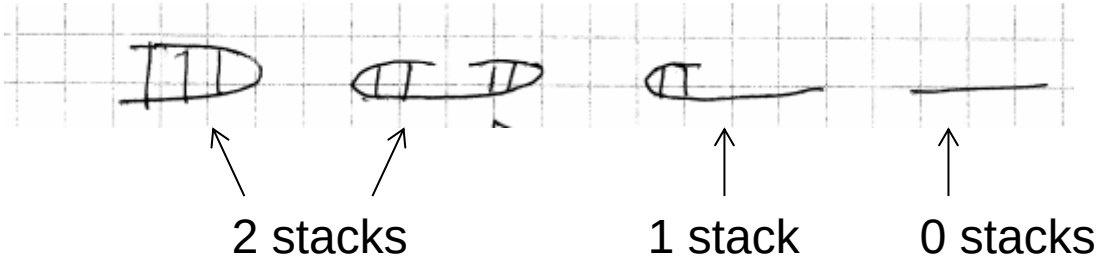
Consecutive base-pairs is a stack – a lone base-pair is not

Simple model for RNA folding

Since stacking energy dominates, ignore the contribution of lone base-pairs and only consider the energy that comes from forming stacks

Each stack lowers the energy of the structure by, $-|\varepsilon_S|$

A structure that has n stacks has an energy of $E = -n\,|\varepsilon_S|$



For small RNA sequences, we can enumerate all possible structures that possess stacks (do not draw structures that have lone base-pairs that are not part of a stack)

Calculate their energy (and possible entropy for the unpaired bases).

Ground state is the lowest (free) energy structure

For probabilities use Boltzman: P(structure) = $\exp(-E_{structure}/kT)/Z$

RNA structure prediction in the real world

In reality, we can not draw all these structures by hand. Use a computer to enumerate possible structure sequences and calculate the energy of the sequence on each structure

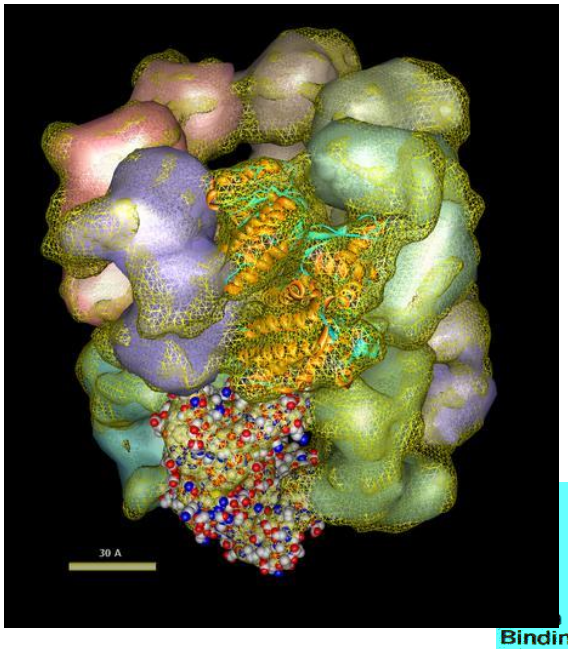
http://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html

Real-world RNA secondary structure prediction uses energies for base-pairing, stacking, looping and forming pseudo-knots

Wide range of applications from predicting mRNA secondary structure, the locations of miRNAs in a genome, designing PCR templates

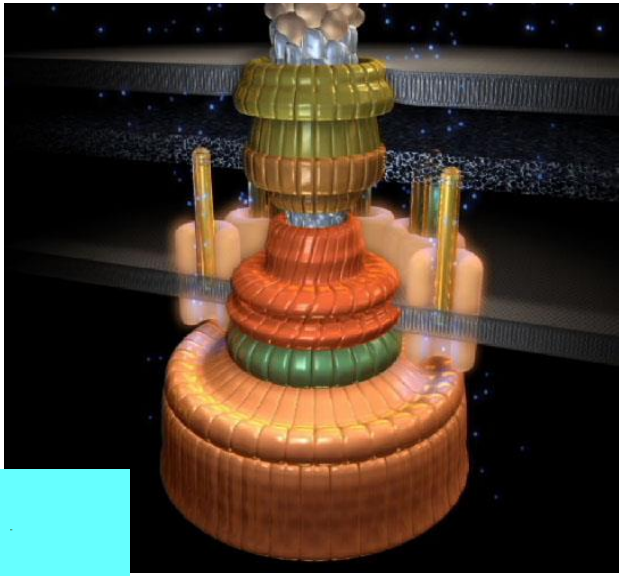
Proteins:

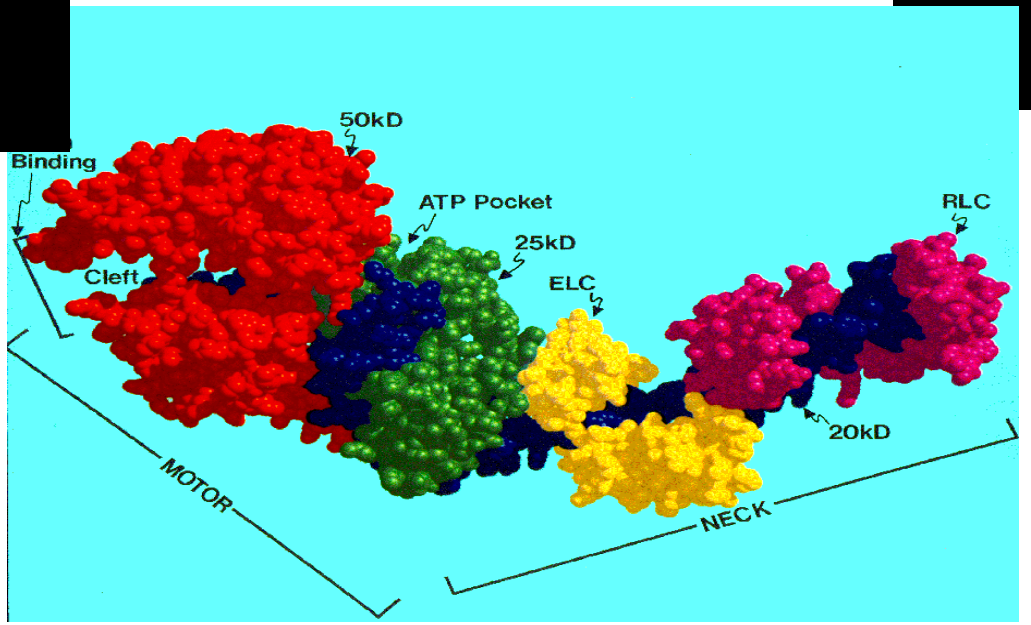
- Proteins are biopolymers that form most of the cellular machinery
- The function of a protein depends on its 'fold' – its 3D structure



Chaperone

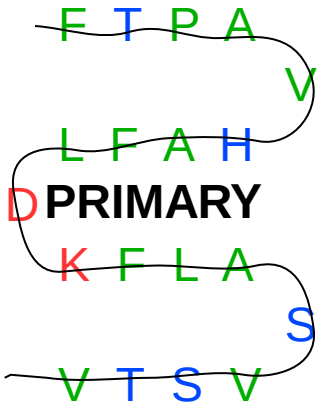
Motor



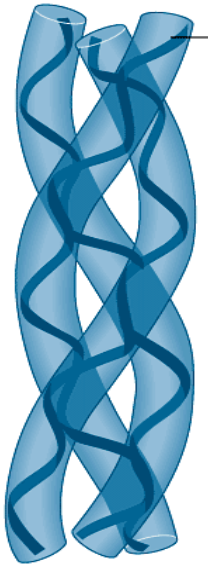


Walker

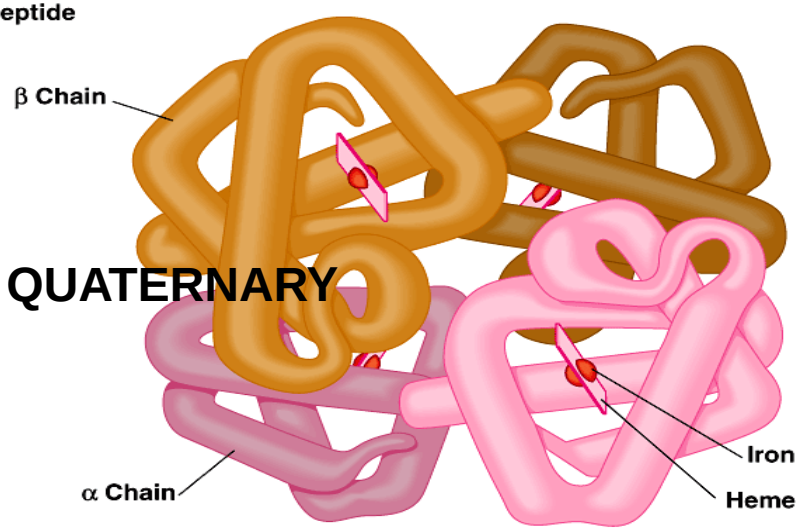
Levels of Folding:



PRIMARY

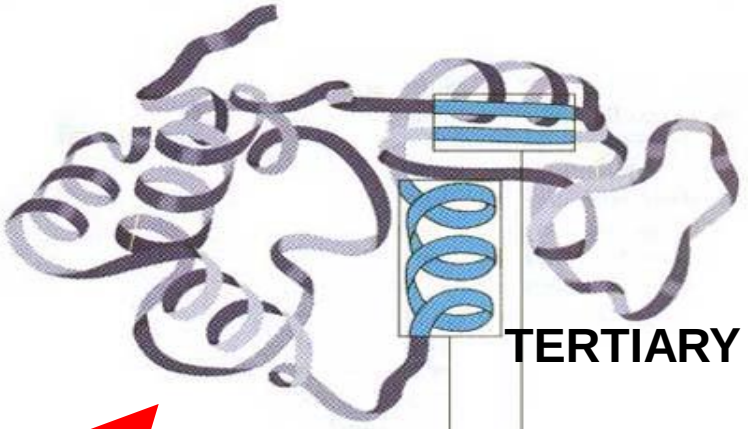


(a) Collagen

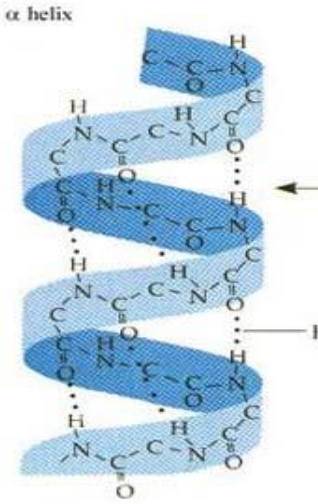


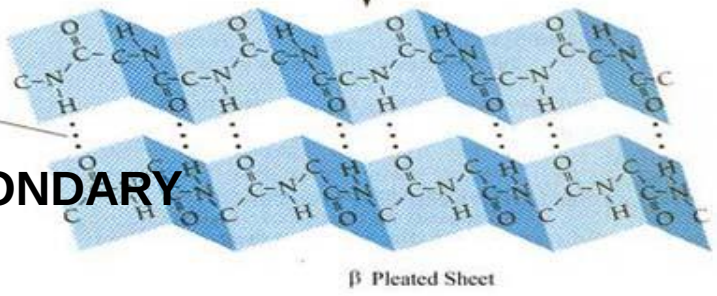
QUATERNARY

(b) Hemoglobin



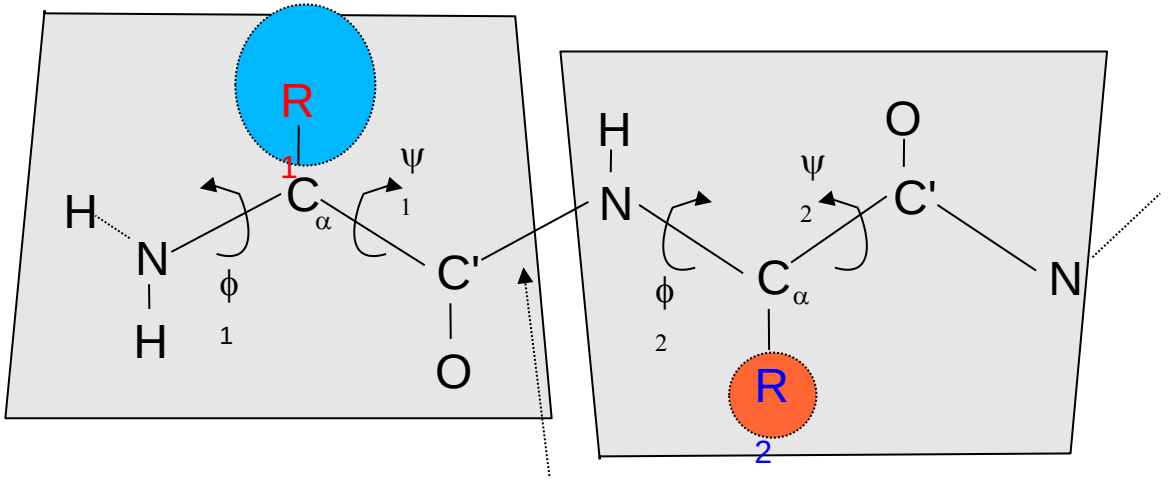
TERTIARY





SECONDARY

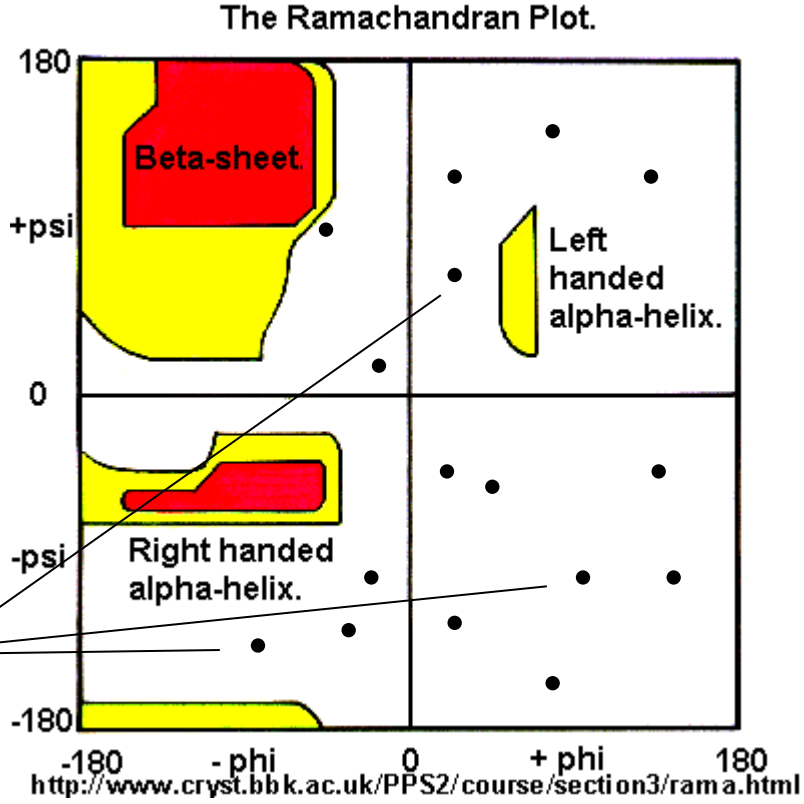
The Backbone



Peptide Bond

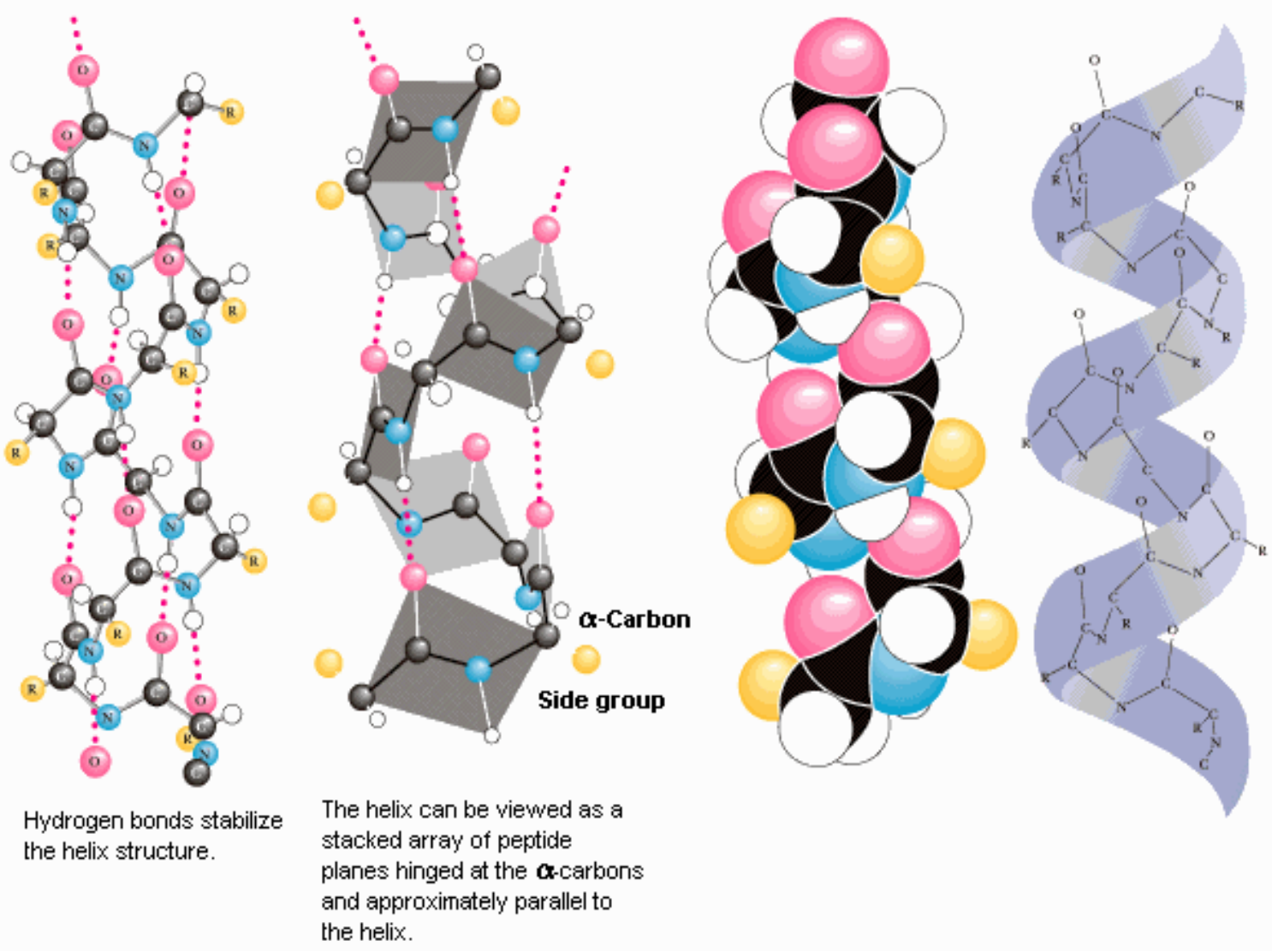
- Amino acids linked together by peptide bonds

Steric constraints lead
only to a subset of possible angles
--> Ramachandran plot



Glycine residues
can adopt many
angles

α Helices

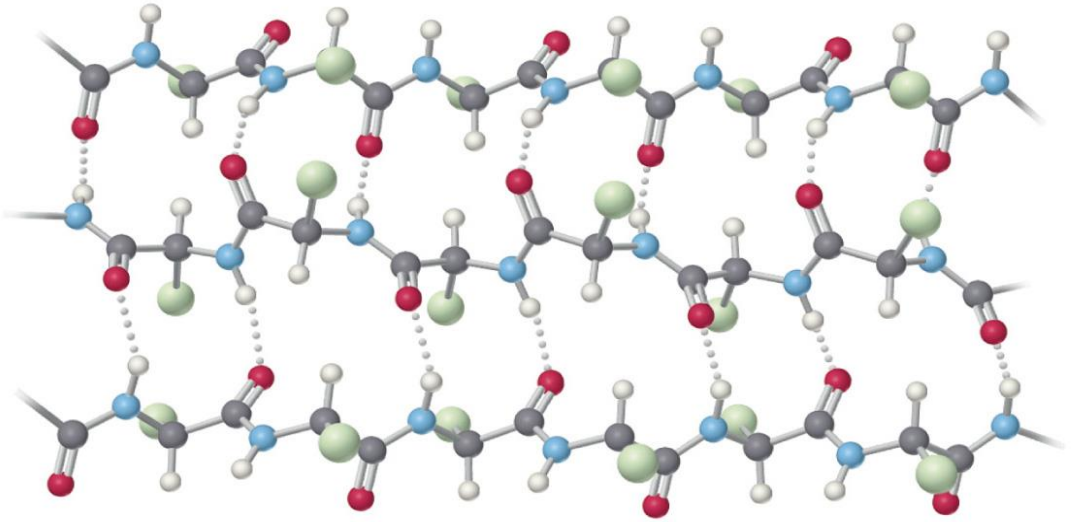


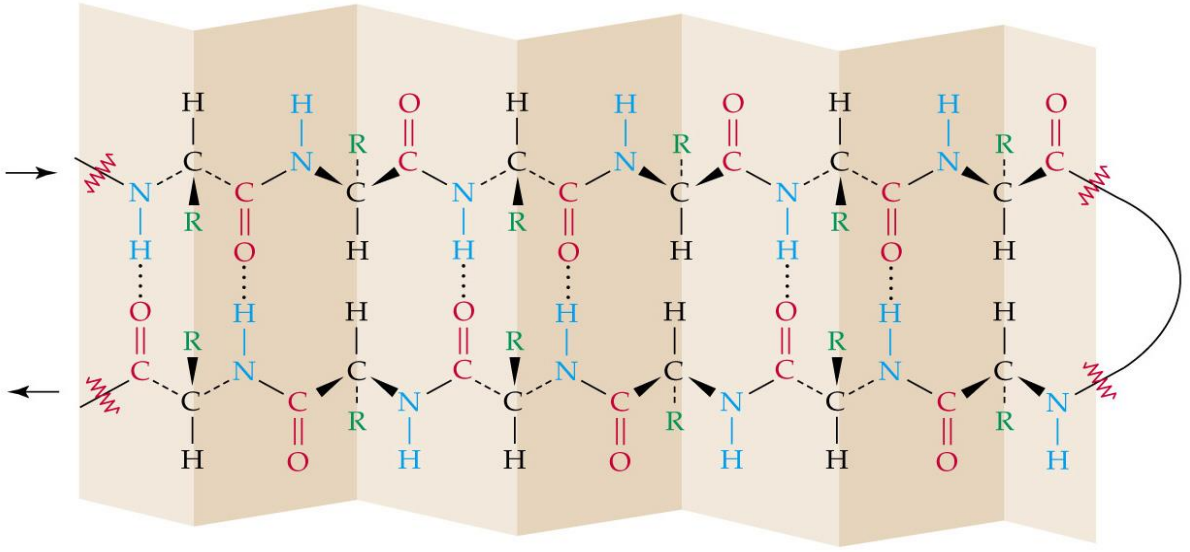
Hydrogen bonds stabilize the helix structure.

The helix can be viewed as a stacked array of peptide planes hinged at the α-carbons and approximately parallel to the helix.

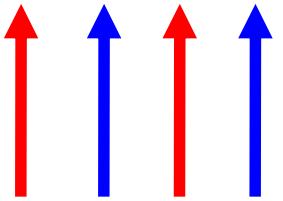
3.6 residues/turn

β Sheets

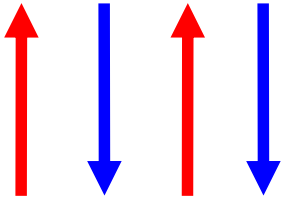




parallel sheet



anti-parallel sheet



other topologies possible
but much more rare

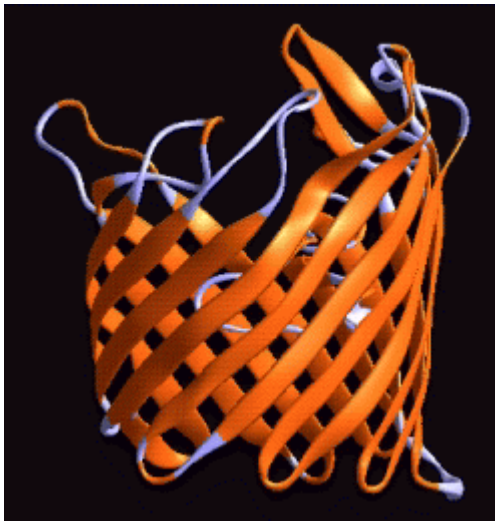
Classes of Folds:

- There are three broad classes of folds: α, β and α+β
- as of today, 25973 known structures --> 945 folds (SCOP 1.65)

alpha class

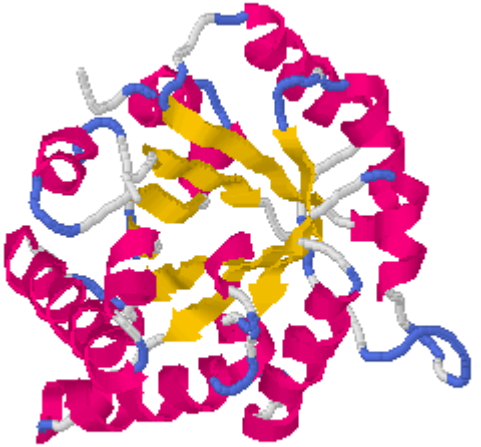
myoglobin – stores oxygen in muscle tissue

beta class



streptavadin – used a lot in biotech, binds biotin

alpha+beta class



TIM barrel – 10% of enzymes adopt this fold, a great template for function

Databases:

SWISSPROT:

contains sequence data of proteins – 100,000s of sequences

Protein Data Bank (PDB):

contains 3D structural data for proteins – 20,000 structures, x-ray & NMR

SCOP:

classifies all known structures into fold classes ~ 800 folds

Protein Folding:

amino acid sequence $\underset{\text{DESIGN}}{\overset{\text{FOLDING}}{\rightleftharpoons}}$ structure

- naturally occurring sequences seem to have a unique 3D structure

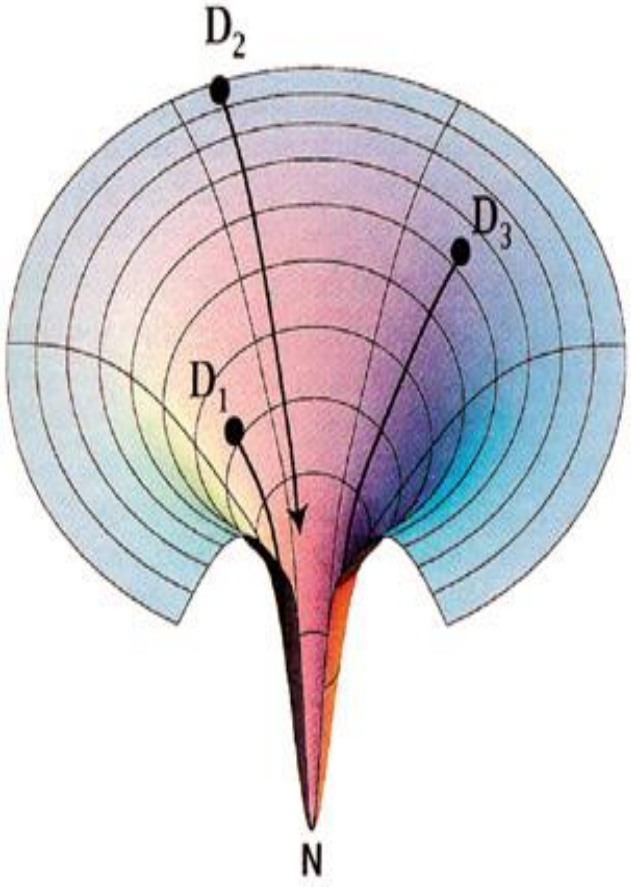
Levinthal paradox: if the polymer doesn't search all of conformation space, how on earth does it find its ground state, and in a reasonable time?

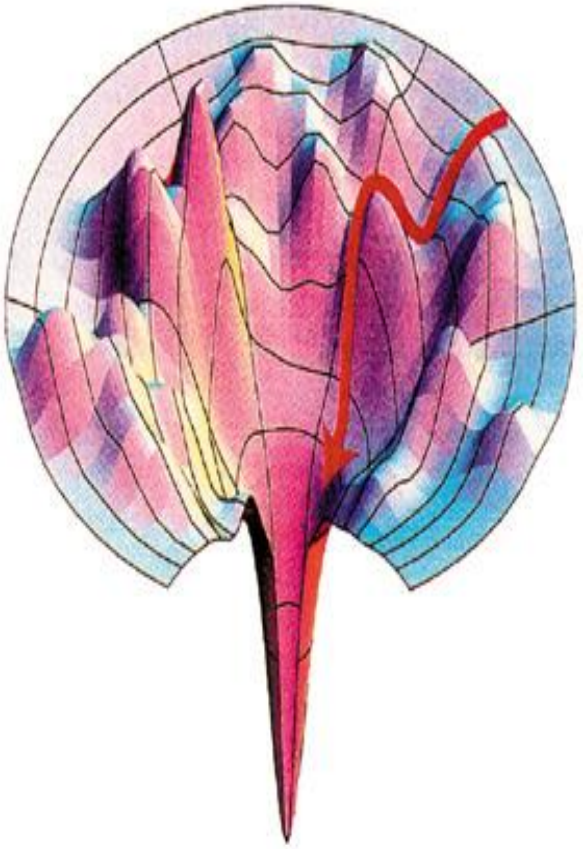
if 2 conformation/residue & dt ~ 10^{-12} -> t=10^{25} years for a protein of L = 150!!!

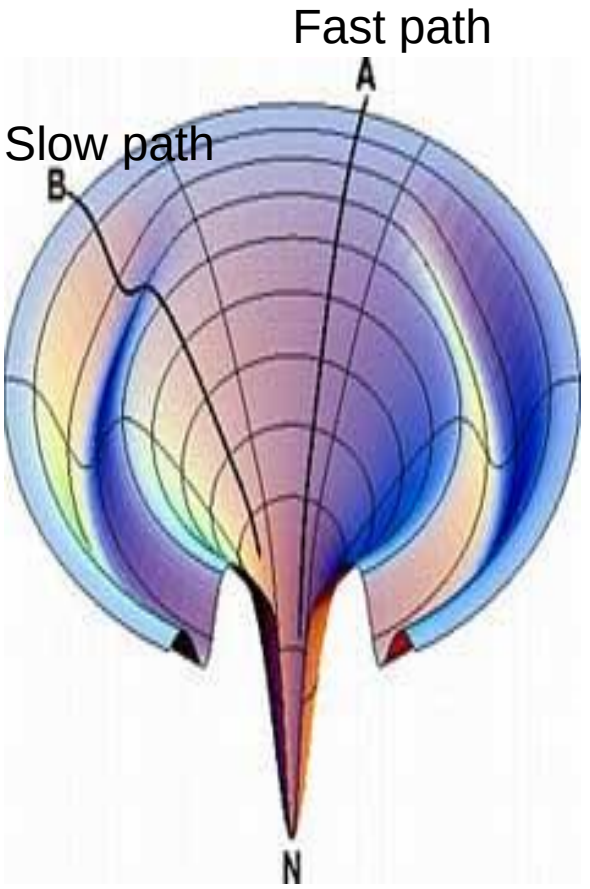
Reality: t = .1 to 1000 s

How do we resolve the paradox?

Paradox Resolved: Funnels







- there are multiple folding pathways on the energy landscape – slow & fast
- If a protein gets stuck (misfolded) there are chaperones to help finish the fold

Factors Influencing folding:

Hydrogen bonding:

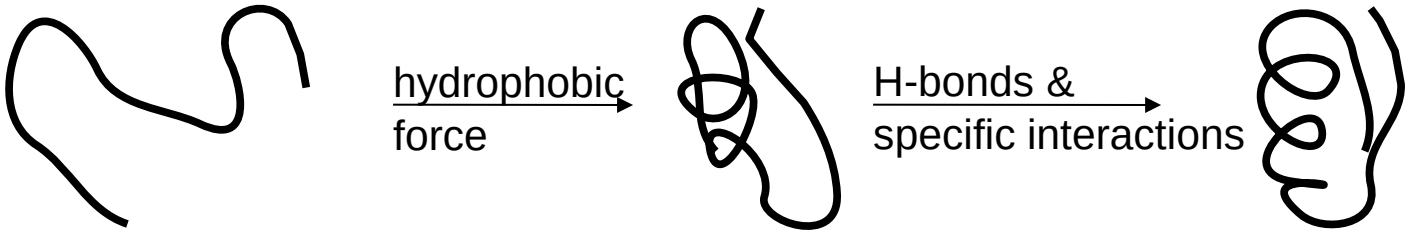
doesn't drive folding since unfolded structure can form H-bonds with H20
drives 2ndary structure formation after compaction

Hydrophobicity:

main driving force
significant energy gain from burying hydrophobic side-chains
leads to much smaller space to search

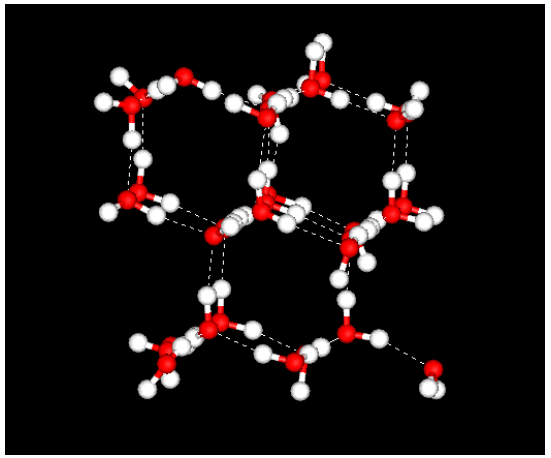
Other interactions:

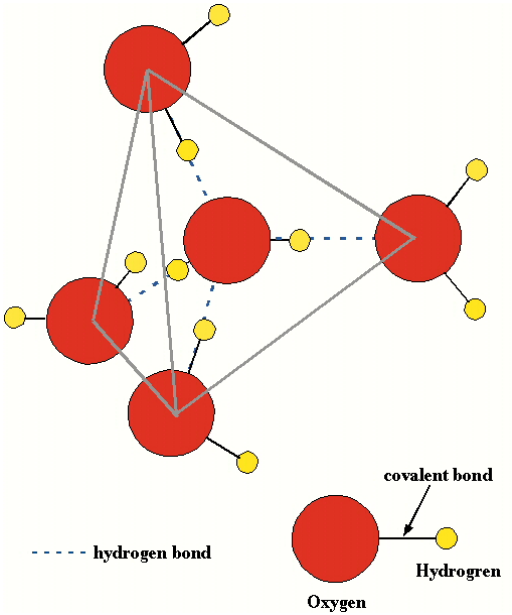
give specificity and ultimately favour final unique state
disulfide bridges = formed between contacting Cystine residues
salt-bridges = formed between contacting -ve and +ve charged residues
secondary structure preferences = from entropy



More on Hydrophobicity:

- Hydrophobicity is an entropic force – water loses entropy due to the presence of non-polar solvent



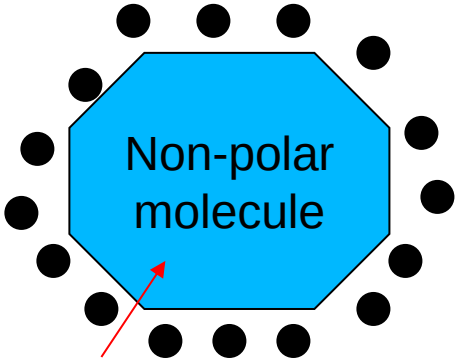


H20 molecules form a tetrahedral structure, and there are 6 hydrogen-bonding Orientations/H20

When a non-polar molecule occupies a vertex → reduces to only 3 orientations

$dS = k \ln 3 - k \ln 6 = - k \ln 2$ → $dG = + kT \ln 2$ costs energy to dissolve

Hydrophobicity and Packing:



Area = A

A non-polar object with area A will disrupt
The local H20 environment

For 1 nm^2 of area ~ 10 H20 molecules are affected

So hydrophobic cost per unit area

$$\gamma = 10\ k\ T\ \ln 2/nm^2 = 7\ k\ T\ /\ nm^2$$

Hydrophobic energy cost = G = γ A

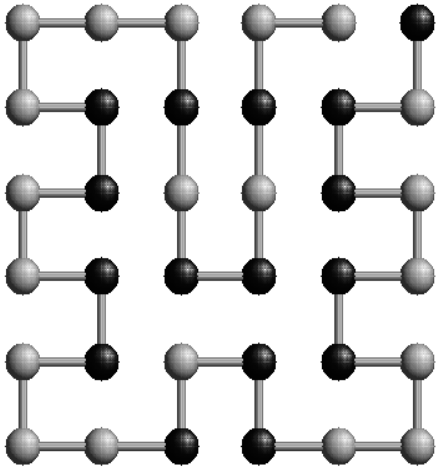
For an O2 molecule in H20, A = 0.2 nm2 so G ~ 1 kT. So O2 easily dissolves in H20

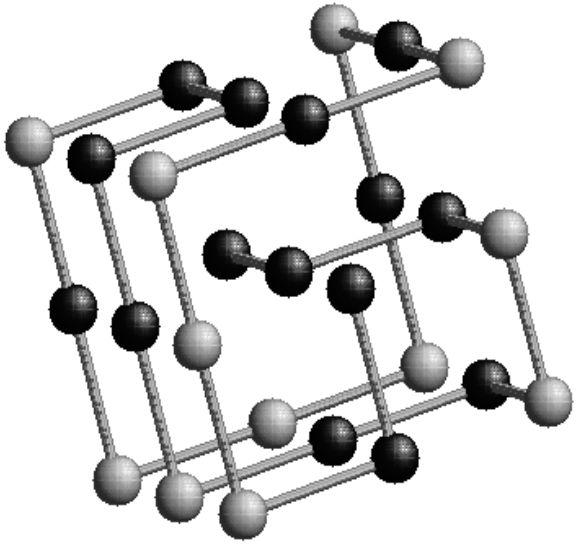
For an octane molecule, G ~ 15 kT, so octane will aggregate so as to minimize the combined exposed area

Simple Models of Folding: Getting at the big picture

- folding proteins in 3D with full atomic detail is HARD!!! essentially unsolved
  --> study tractable models that contain the essential elements

SIMPLE STRUCTURE MODEL = LATTICE MODELS:





- enumerate all compact structures that completely fill a 2D or 3D grid
- can also study non-compact structures by making larger grid

Simple Energy functions:

H-P Models:

- amino acids come in only two types, H = hydrophobic, P = polar
- interactions: H-H, H-P & P-P with $E_{PP} > E_{HP} > E_{HH}$

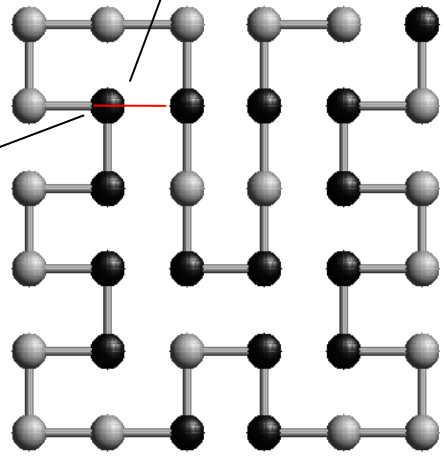
Energy = $\Sigma E_{ij} \Delta(r_i - r_j)$

- could use full blown 20 x 20 E_{ij} matrix = Miyazawa-Jernigan matrix

Solvation Models:

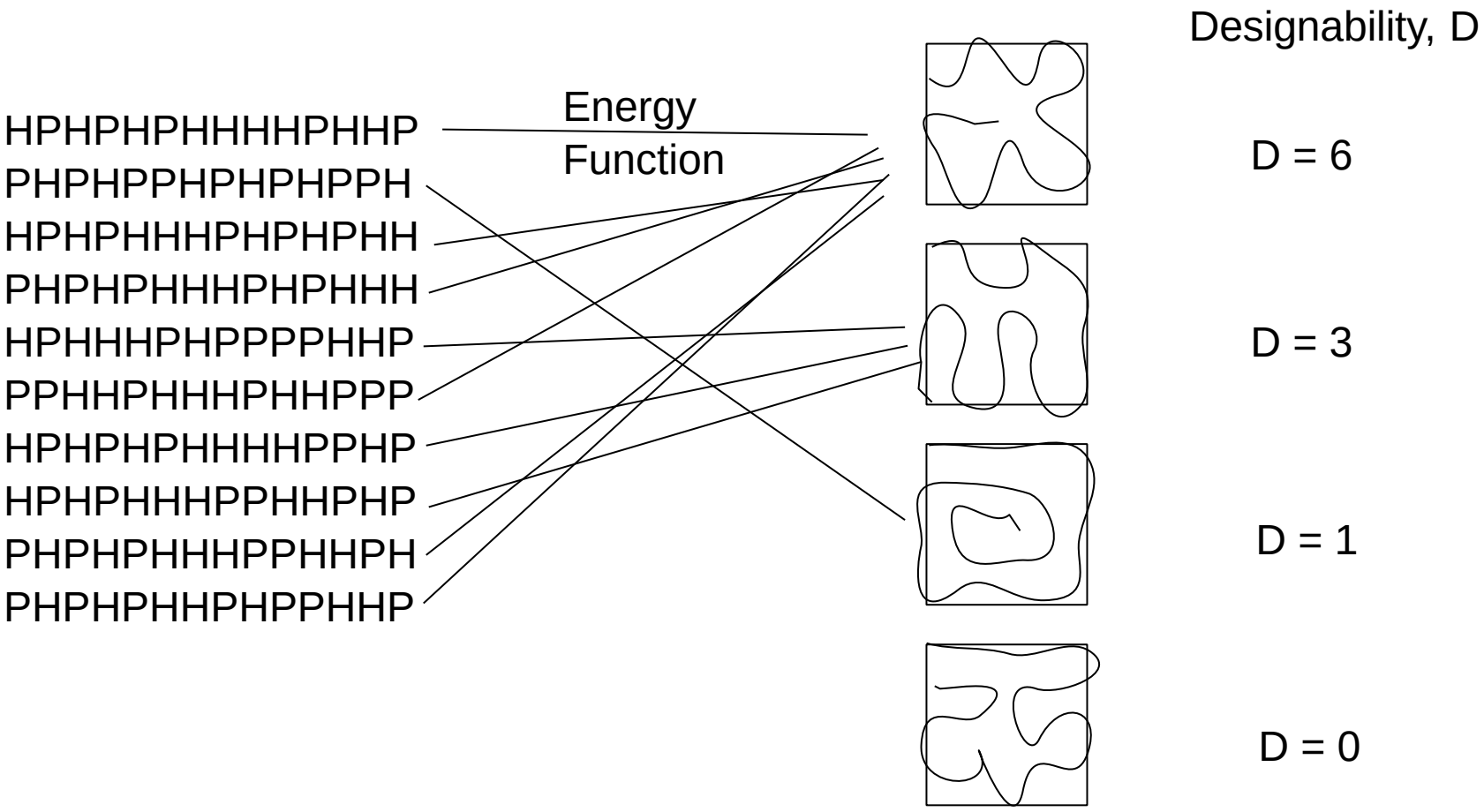
- energy is gained for burying hyrdophobic residues
- if residue is buried, surface exposure, s = 1
- if residue is exposed, surface exposure, s = 0
- hydrophobicity scale: H: h = -1, P: h= 1
- Energy = $\Sigma h_i s_i$

Ground state structure has the lowest energy for given sequence



Model Results: Designability Principle

- Fold random HP sequences, and determine the ground state for each
- Designability = # of sequences which fold into a given structure

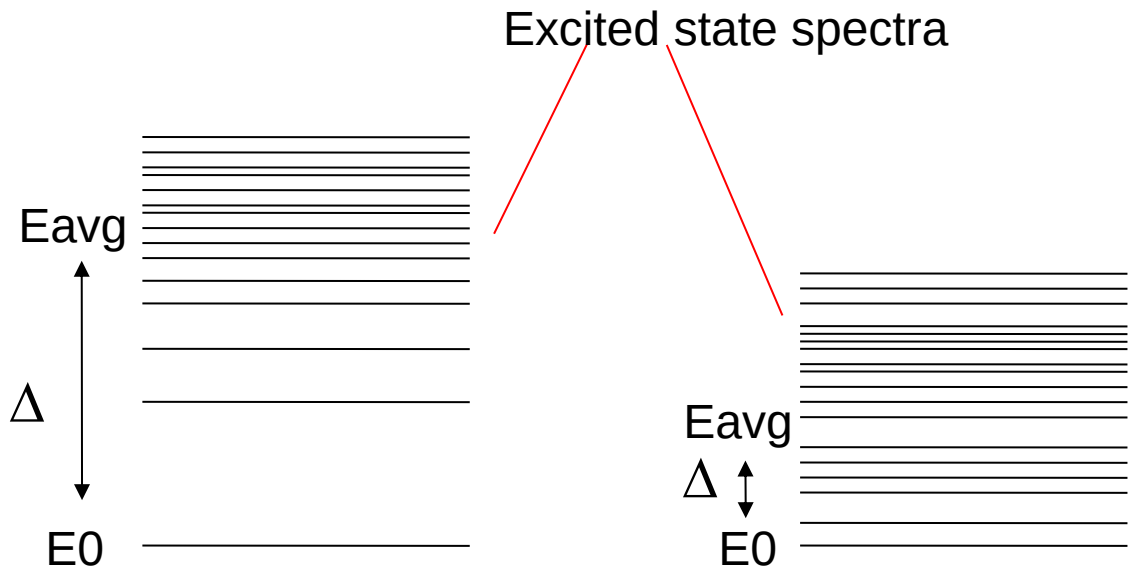


Designability Principle: there are only a few highly designable structure, most structures have very few sequences that fold into them

Thermodynamic Stability

- high designabilty implies mutational stability, does it imply thermodynamic stability?

YES



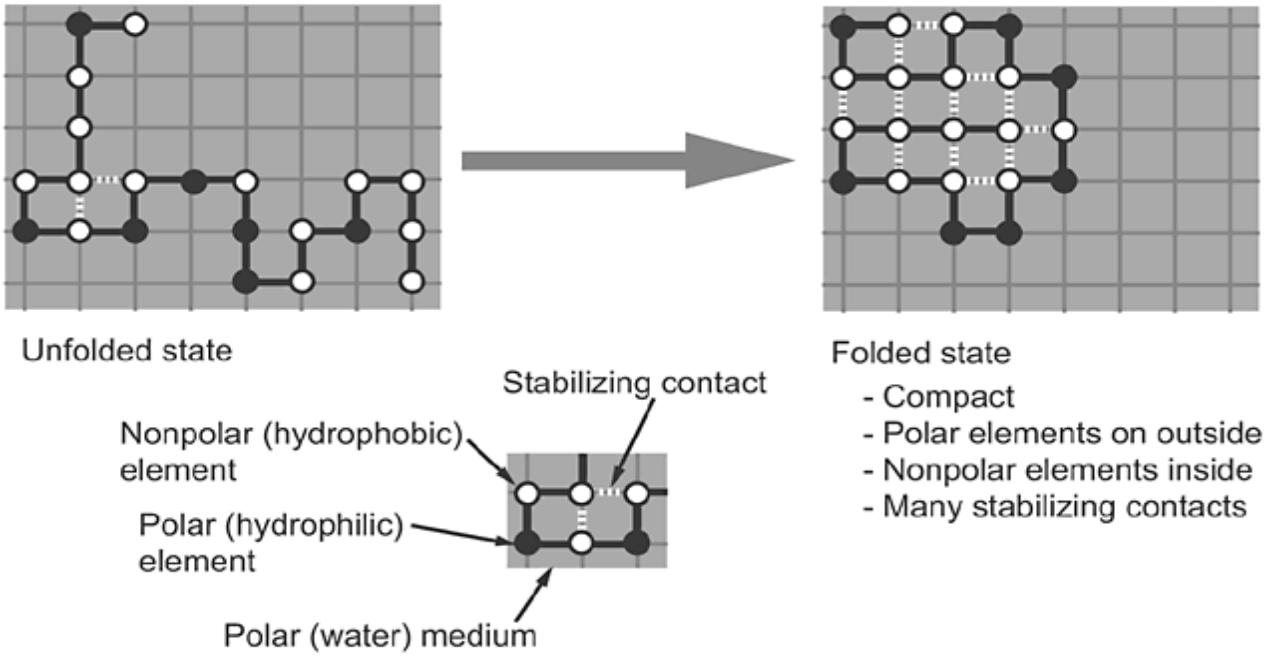
HIGH DESIGN STRUCTURE

LOW DESIGN STRUCTURE

- Highly designable structures are characterized by a large energy gap, Δ

Fast Folding

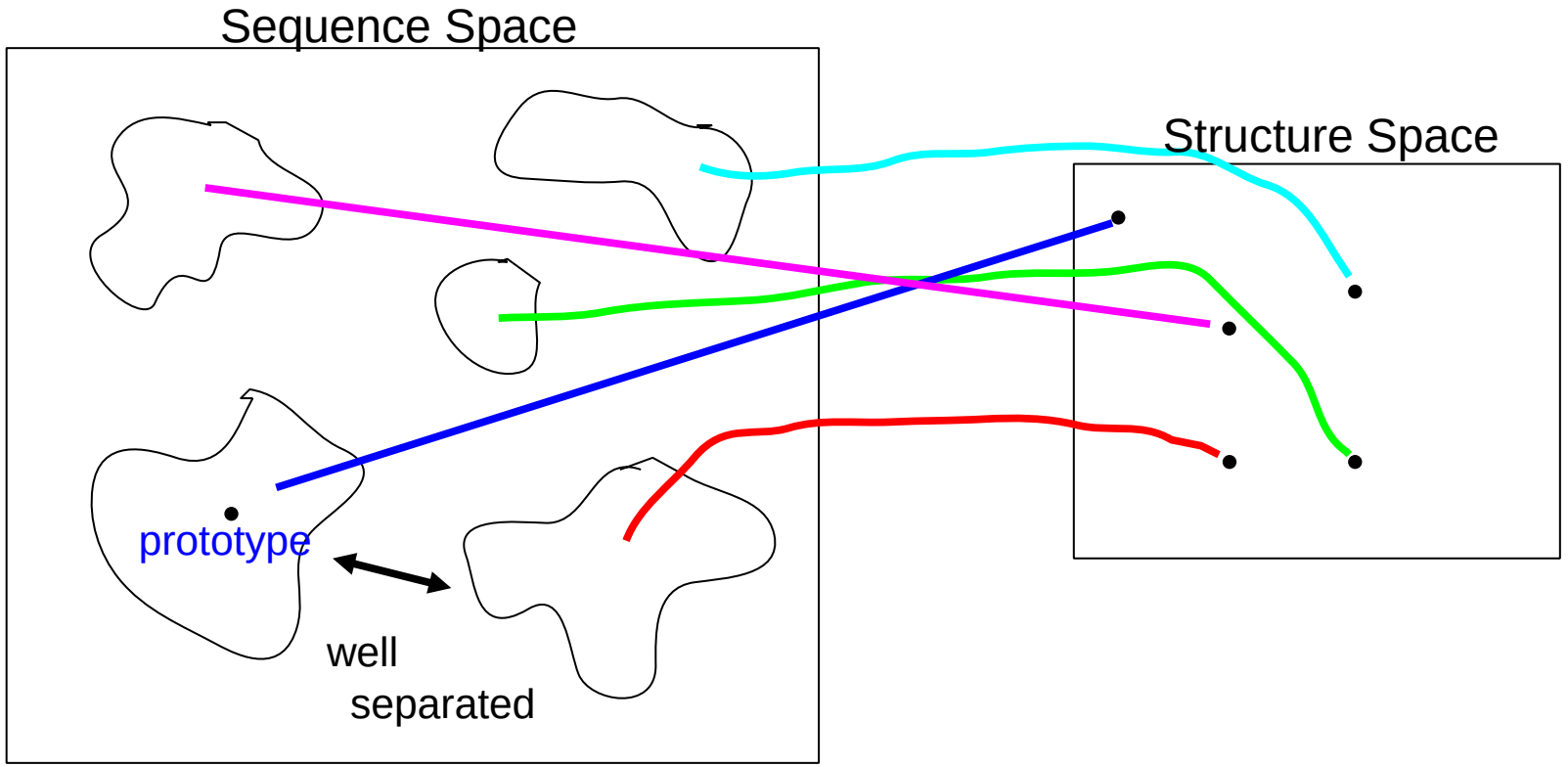
- High designability structures are fast folders, since there are few low lying energy structures to compete with – no kinetic traps
- Low designability structures are slow – have many competing low energy alternatives which act as kinetic traps



- Determine kinetics using Metropolis Monte-carlo

  t ~ # of monte-carlo steps needed to first achieve near native state (90%)

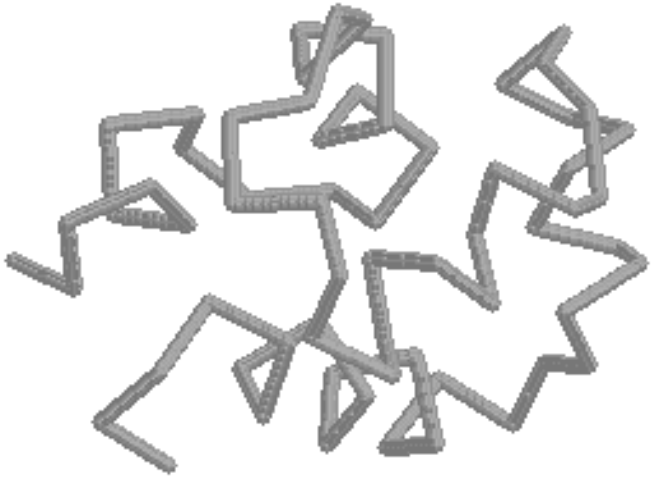
Neutral Networks in Protein Folding:



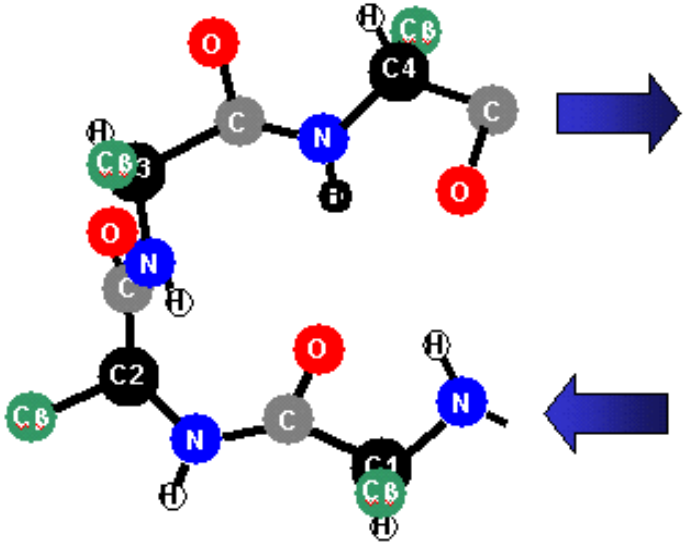
- Just like RNA, designable proteins have well connected neutral networks
- Unlike RNA, these neutral networks are well separated, so they are not space covering
- Prototype sequence tends to have best thermodynamic properties (cluster center)

Protein Folding in the Real World:

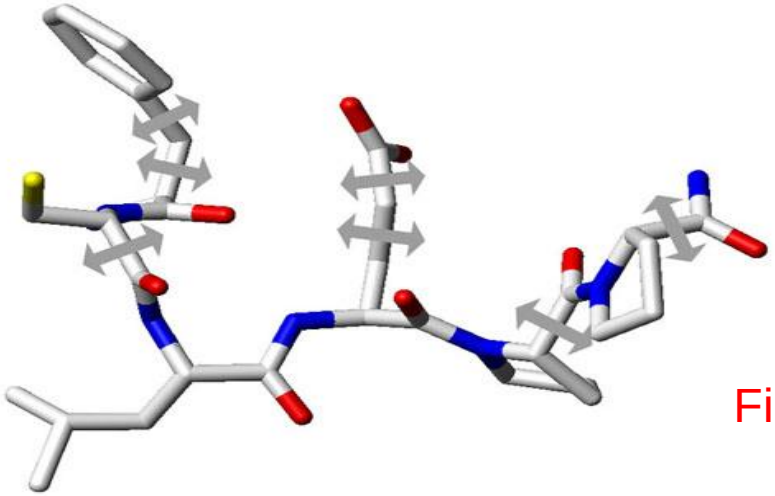
OFF-LATTICE MODELS:



Coarse: just C_α and C_β



Medium: all backbone and C_b



Fine: all atoms and use side chain rotamers

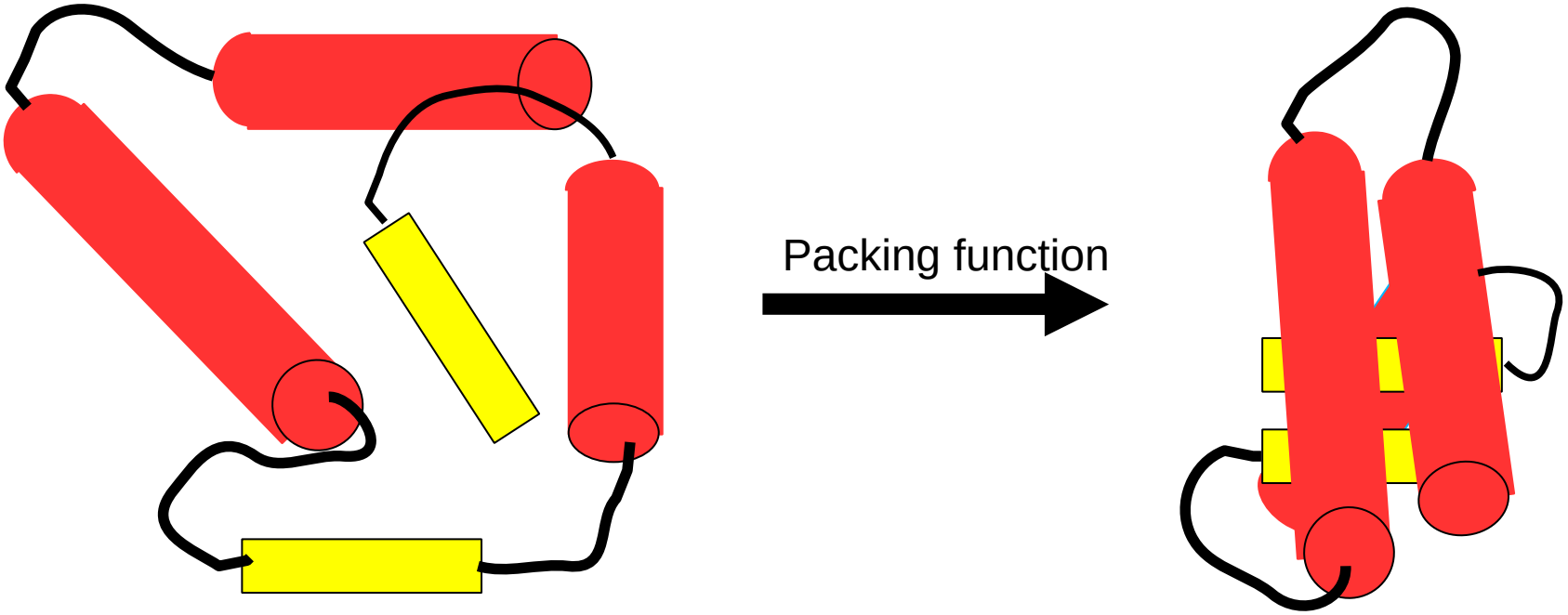
Structure Construction:

Enumerate structures:

- enumerate all structures that are possible using a finite # of (ϕ,ψ) angles
- e.g. 4 pairs, L = 20 --> 4^{20} = 1 x 10^{12} structures!!!

Packing of secondary elements:

- pack together in 3D a fixed set of secondary structural elements
- can go to much larger structures
- must sample the space



Protein Design:

1) Improve natural folds:

give natural proteins new function, stability, kinetics

2) The search for novel folds: for L = 100 --> 100^{20} sequences !!!

There may be sequences that fold into structures not seen in nature

Inverse folding problem: given a structure find a compatible sequence for which the structure is the ground state fold

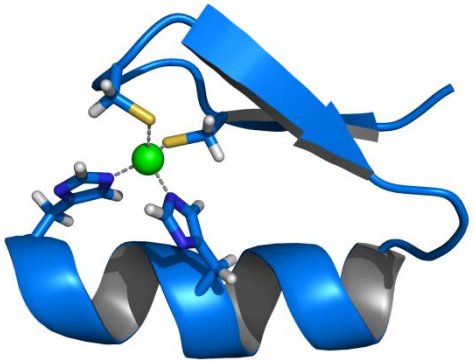
E(s) →

VLMQEGGFVLMS
MQEFTDGVMAA
AAVKRGTWWSR
EFVKLILAAIRST

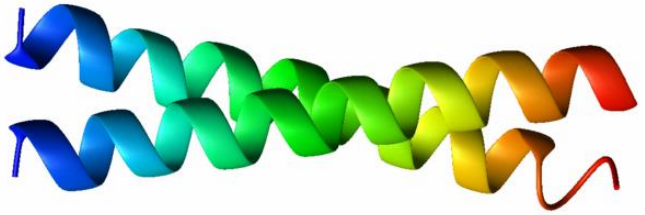
Can we design any structure we want? NO, designability principle.

Successful Designs

Redesigned Zinc Finger (Steve Mayo Lab)



Design of right-handed coiled coil (Harbury & Kim)



Binary patterning of helical bundle (Michael Hecht Lab)

Design of novel fold (David Baker Lab)