800 ANNI 1222·2022

UNIVERSITÀ DEGLI STUDI DI PADOVA

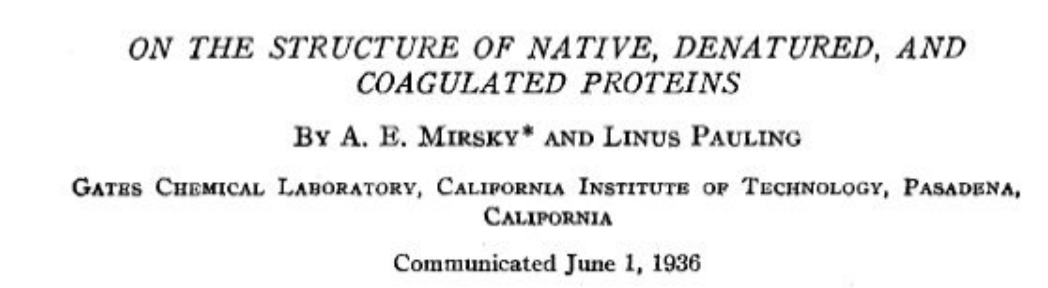
DIPARTIMENTO MATEMATICA

DATA SCIENCE
UNIVERSITY OF PADOVA

PROTEIN FOLDING PROBLEM

Master of Science in Data Science

Damiano Piovesan

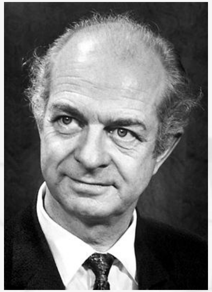


ON THE STRUCTURE OF NATIVE, DENATURED, AND COAGULATED PROTEINS

By A. E. Mirsky* and Linus Pauling

Gates Chemical Laboratory, California Institute of Technology, Pasadena, California

Communicated June 1, 1936

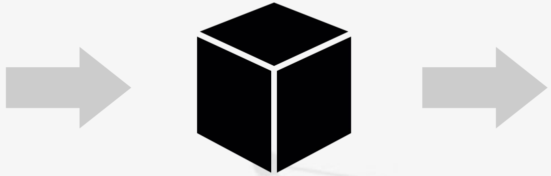


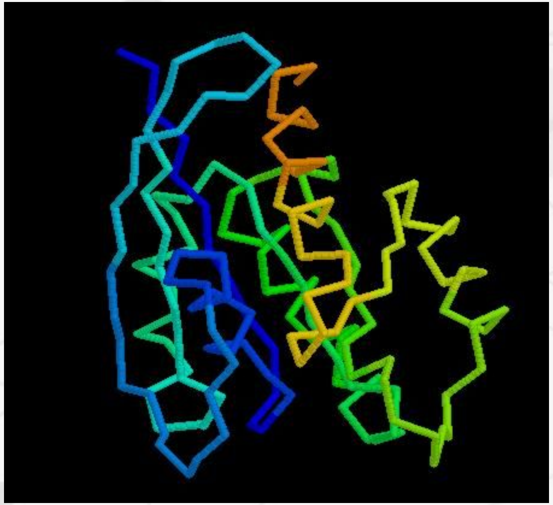
*“Our conception of a **native protein** molecule (showing specific properties) is the following. The molecule consists of one polypeptide chain which continues without interruption throughout the molecule (or, in certain cases, of two or more such chains), this chain is **folded into a uniquely defined configuration**”*

Linus Pauling, 1904 - 1994

Protein folding problem

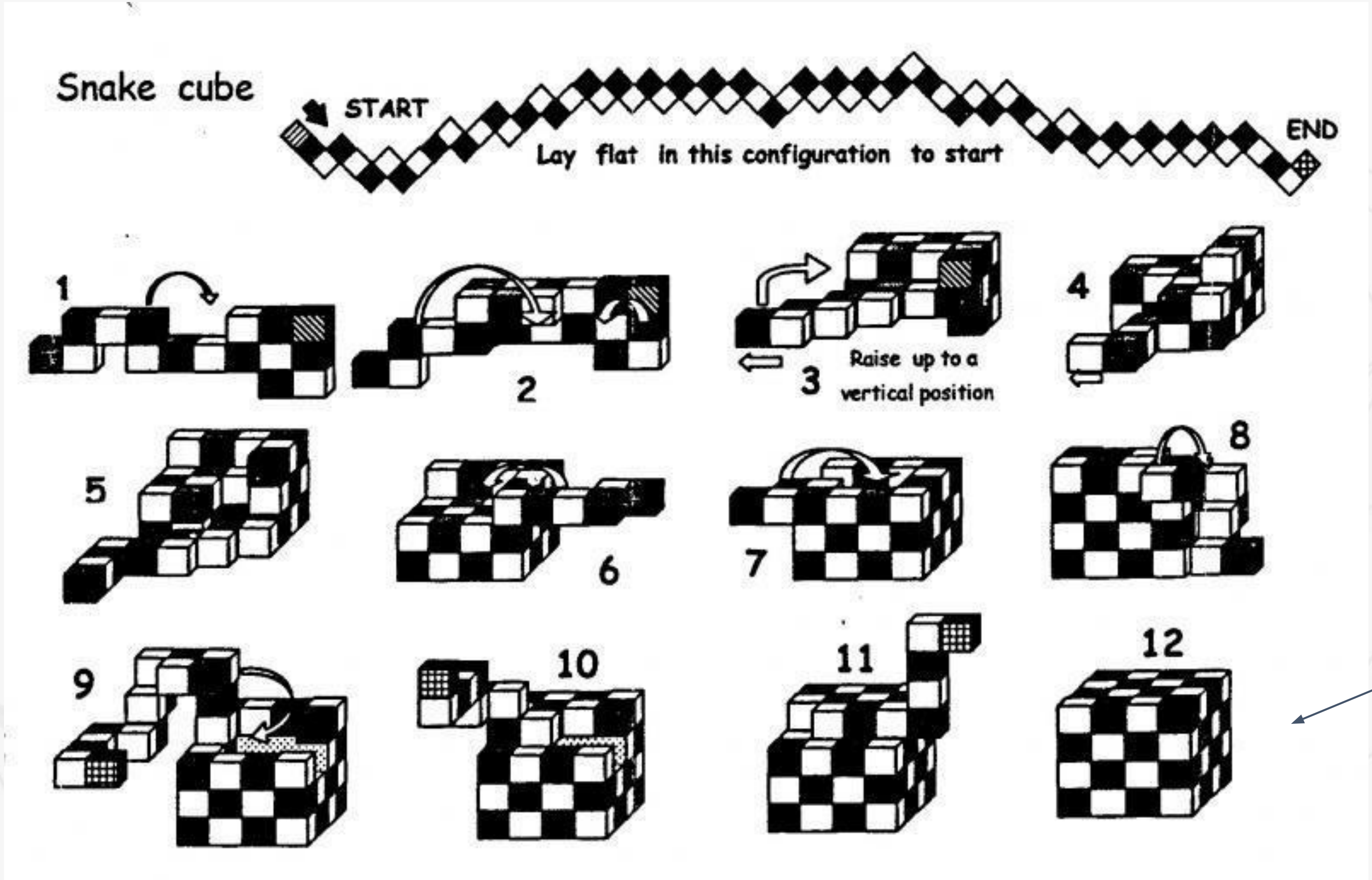
```
>P37840 Alpha-synuclein
MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKE
GVLYVGSKTKEGVVHGVATVAEKTKEQVTNVGGAV
VTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEE
GAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA
```





- Single polypeptide chains
- 20 L-amino acids, no modifications
- Water solvent, no reagents

Conformations / Configurations


Snake cube
START
Lay flat in this configuration to start
END
1
2
3
Raise up to a vertical position
4
5
6
7
8
9
10
11
12

Native conformation

Levinthal’s paradox

Assumptions (wrong)

- A protein sample all possible conformations (random walk)
- The conformation of a residue is independent of the rest

Statement

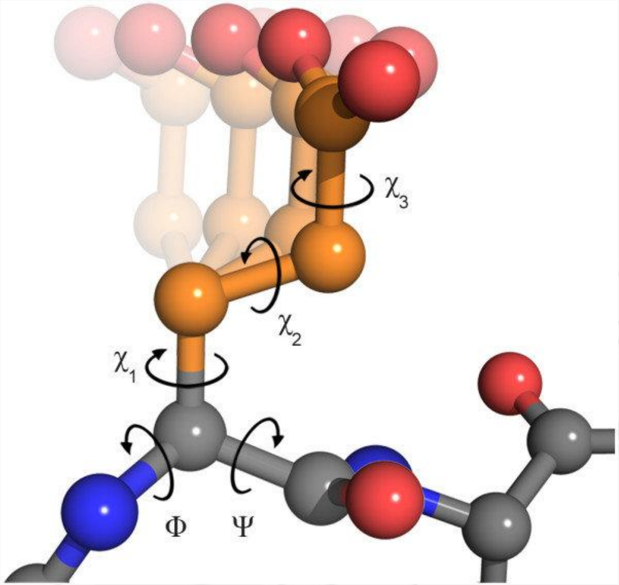
- The protein will never fold to its native structure

Example

- 6 possible conformations (type of secondary structure) x 100 residues
- $6^{100} \simeq 10^{78}$ conformations
- 10^{58} years to fold. 1 picosecond (10^{-12} seconds) for a single molecular vibration

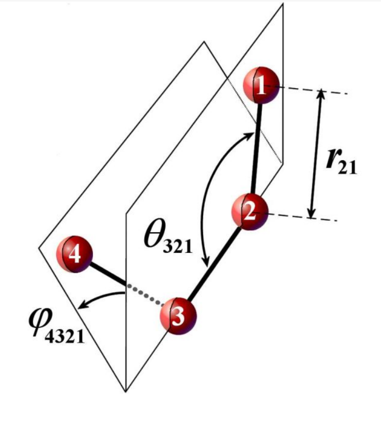
Local conformations

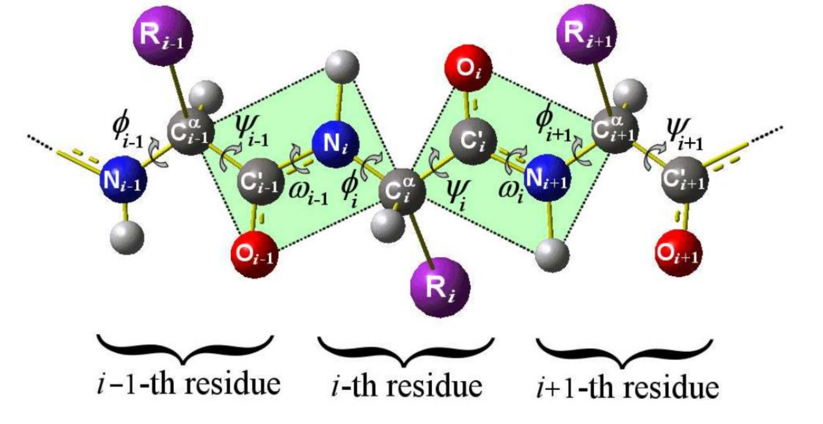
Amino acid rotamers (degree of freedom)

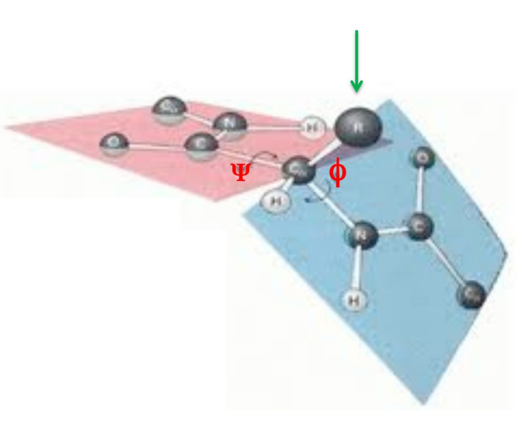


Beyond rotamers: A generative, probabilistic model of side chains in proteins.
Harder at al. 2010, BMC Bioinformatics, 11(1):306

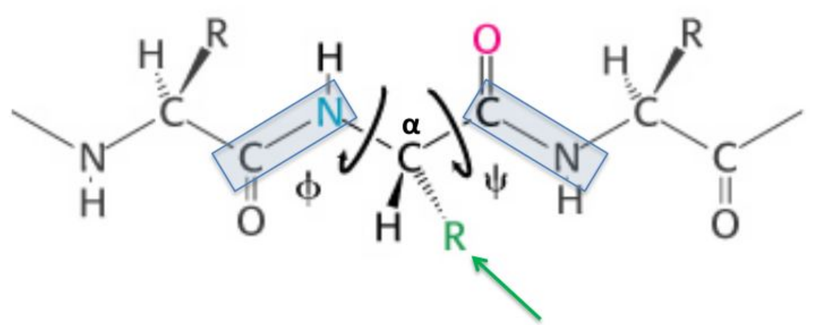
Dihedral angles (backbone)







- The **peptide bond** is **rigid** and **planar** bond because it has a **partial double bond** character
- It is **0.13 Angstrom shorter** than the C-N single bond yet not as short as a double bond



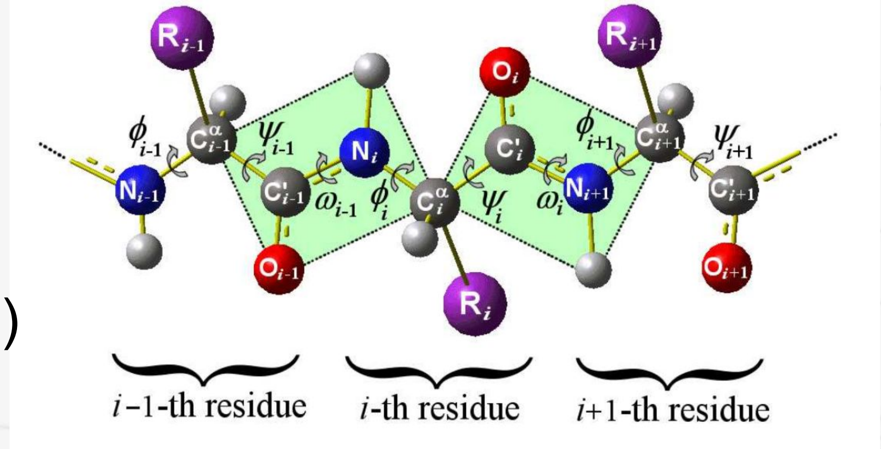
Degrees of freedom

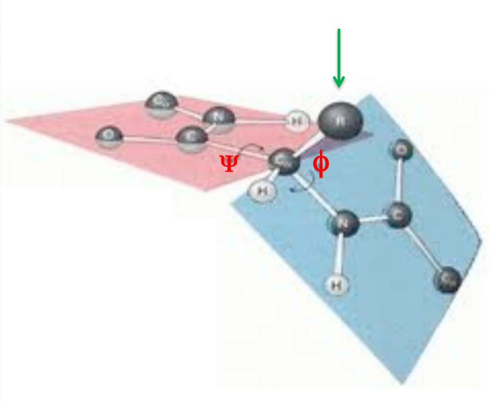
Hard (no freedom)

- Bond lengths
- Bond angles
- Dihedral angles (peptide bond)
  - main chain → ω

Soft

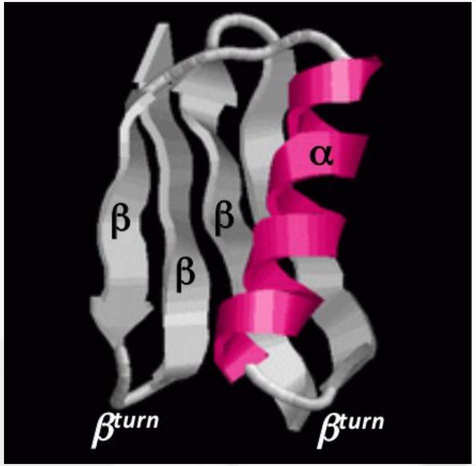
- Dihedral angles (single bond)
  - main chain → Φ, Ψ
  - sidechain → X

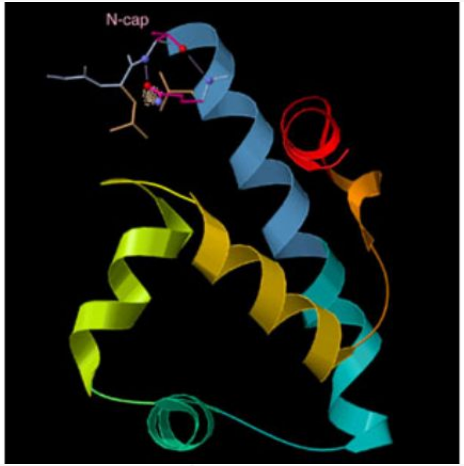


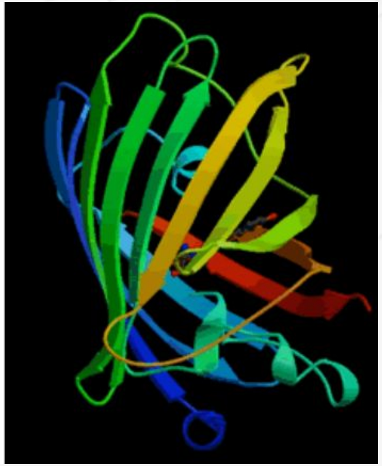


Secondary structures

- **α-helix** and **β-sheet** are regular structures, stable and frequent in proteins. They minimize steric repulsion and maximize H bonds
- **Random coil**, apparently not regular

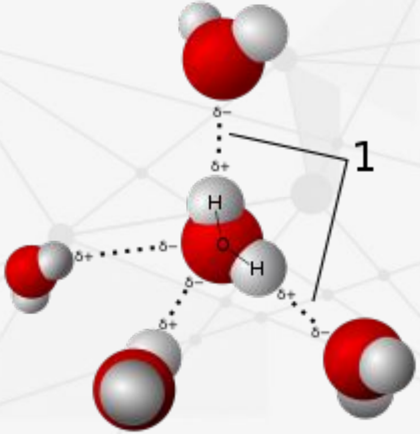






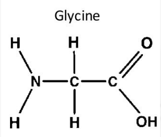
Hydrogen bonds

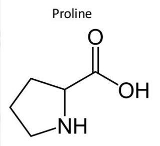
- **Electrostatic force** of attraction
- Between a hydrogen (**H**) atom bound to a more electronegative atom or group (**N, O, F**) - the donor (**Dn**)...
- … and another electronegative atom bearing a lone pair of electrons - the acceptor (**Ac**)
- **Dn–H···Ac**

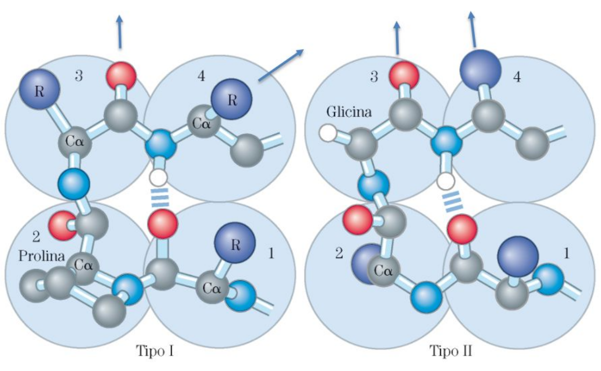


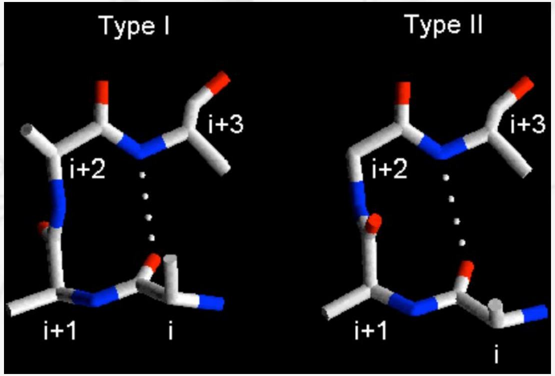
β-bulge loops

- Give rise to chain reversal
- Proline and glycine are the most frequent
- **Type I**, CO of residue i and the NH of residue i+3 (a β-turn)
- **Type II**, CO of residue i+4 and the NH of residue i

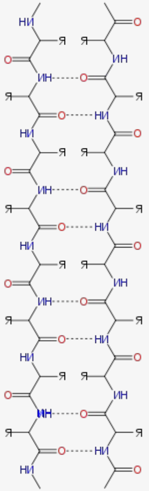

Glycine


Proline

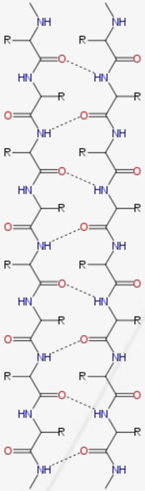




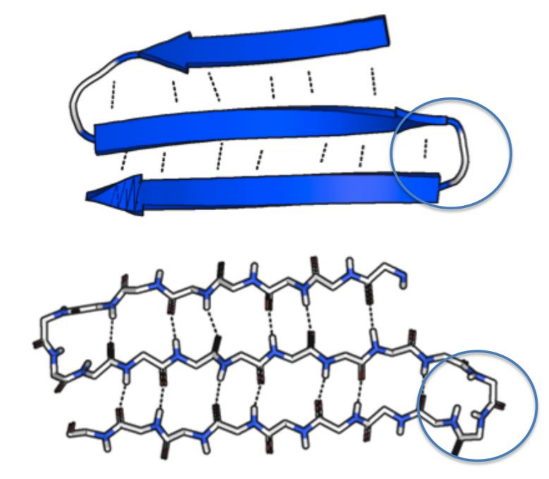
β-sheets



Anti-parallel



Parallel



β-Sheet (3 strands)

Conformation patterns - Secondary structure

Patterns of **hydrogen bonds**

Specific **Φ, Ψ angles**

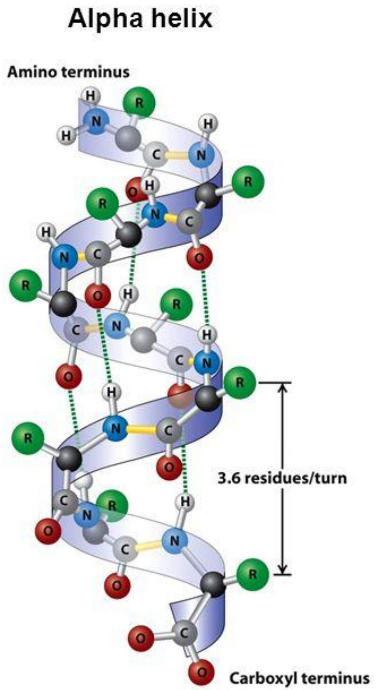


Figure 3-4
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

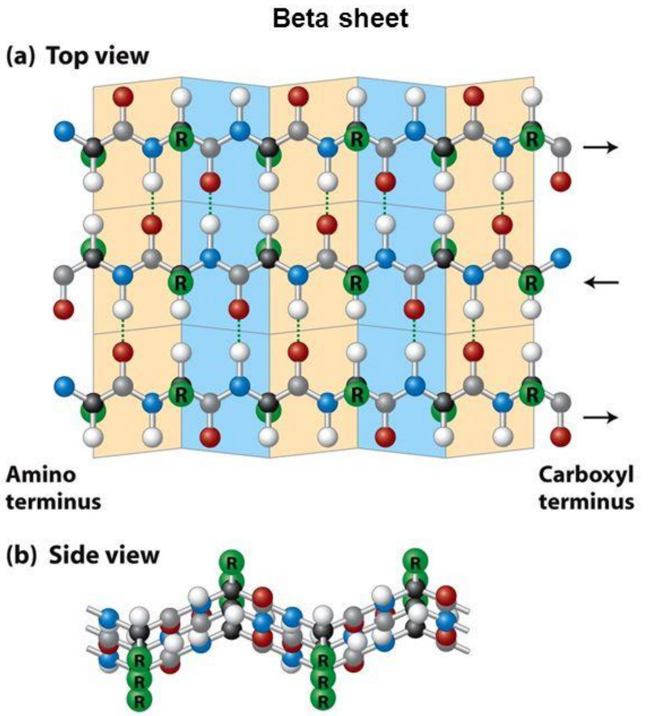
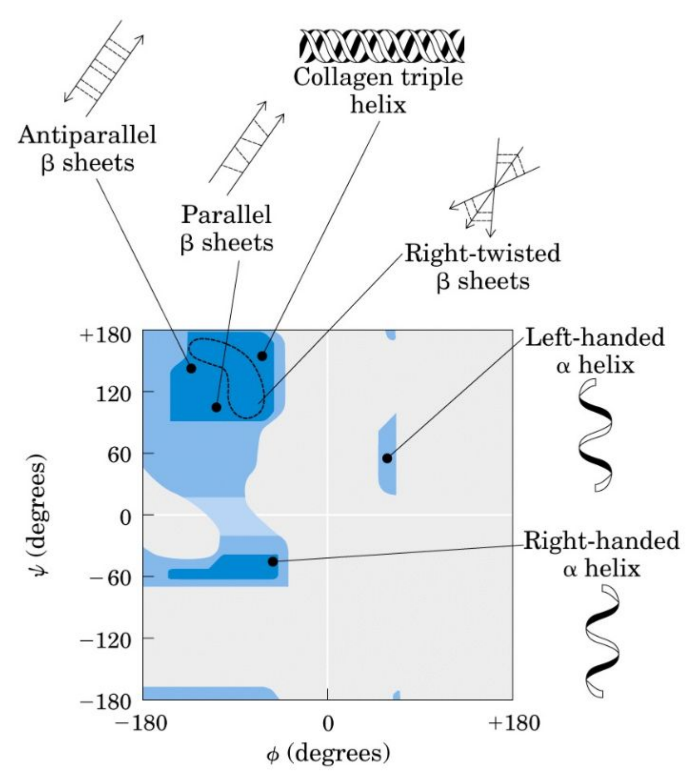
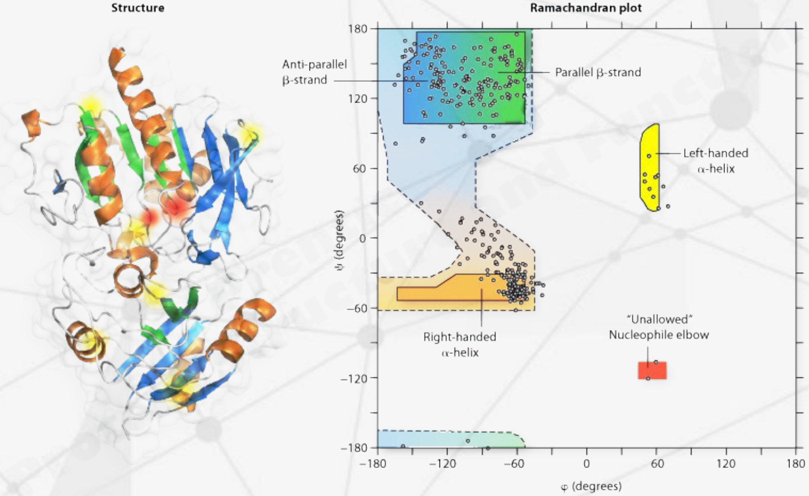


Figure 3-5
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

Conformational preferences
Ramachandran plot





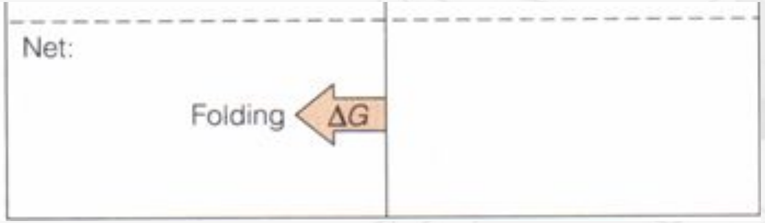
Folding driving forces

Computer Science - Structural bioinformatics

2020

Folding energy

$$\Delta G_{fold} = G_{native} - G_{unfold}$$

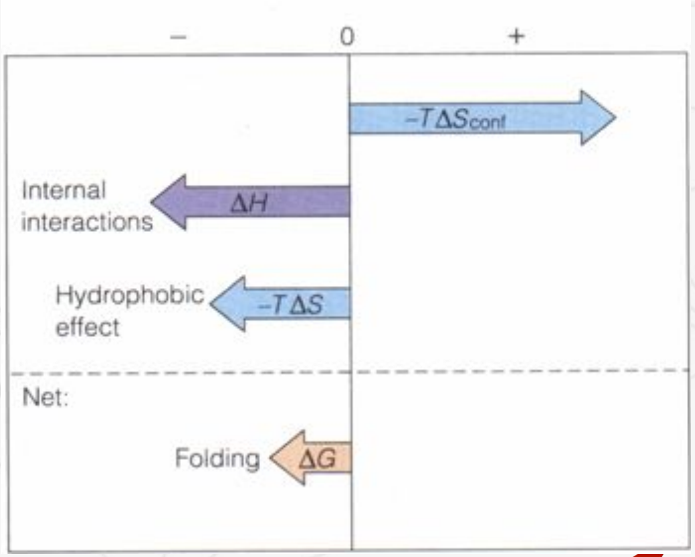


- *G,* energy of Gibbs
- Spontaneous processes have negative *G*
- Proteins are marginally stable ca. **-5 / -15 kcal/mol**

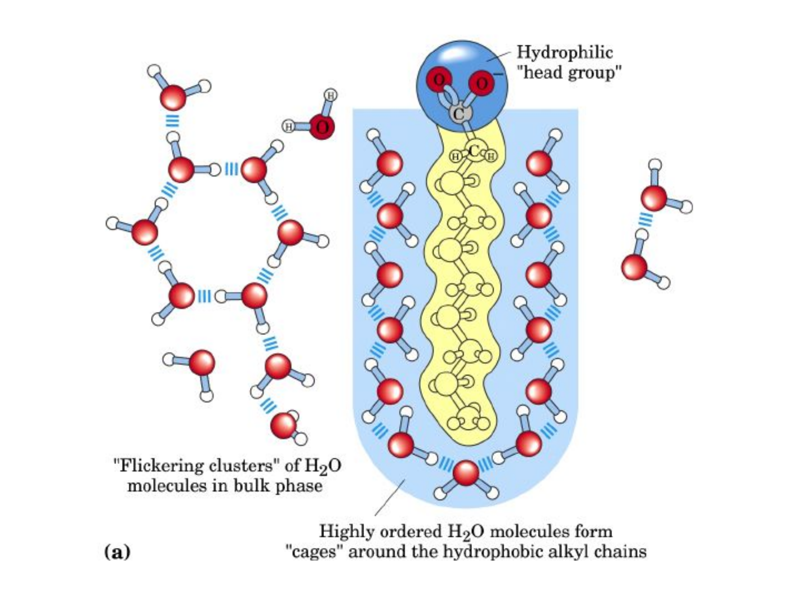
Folding energy

$$\Delta G_{fold} = \Delta H_{fold} - T\Delta S_{fold}$$

- ***ΔH_{fold}*** →**enthalpy gain**, the contribution of novel interactions formed in the folded configuration
- ***$-T\Delta S_{conf}$*** →**entropy loss**, the cost of reducing the degree of freedom generated by adopting a fixed conformation
- **Hydrophobic effect?**

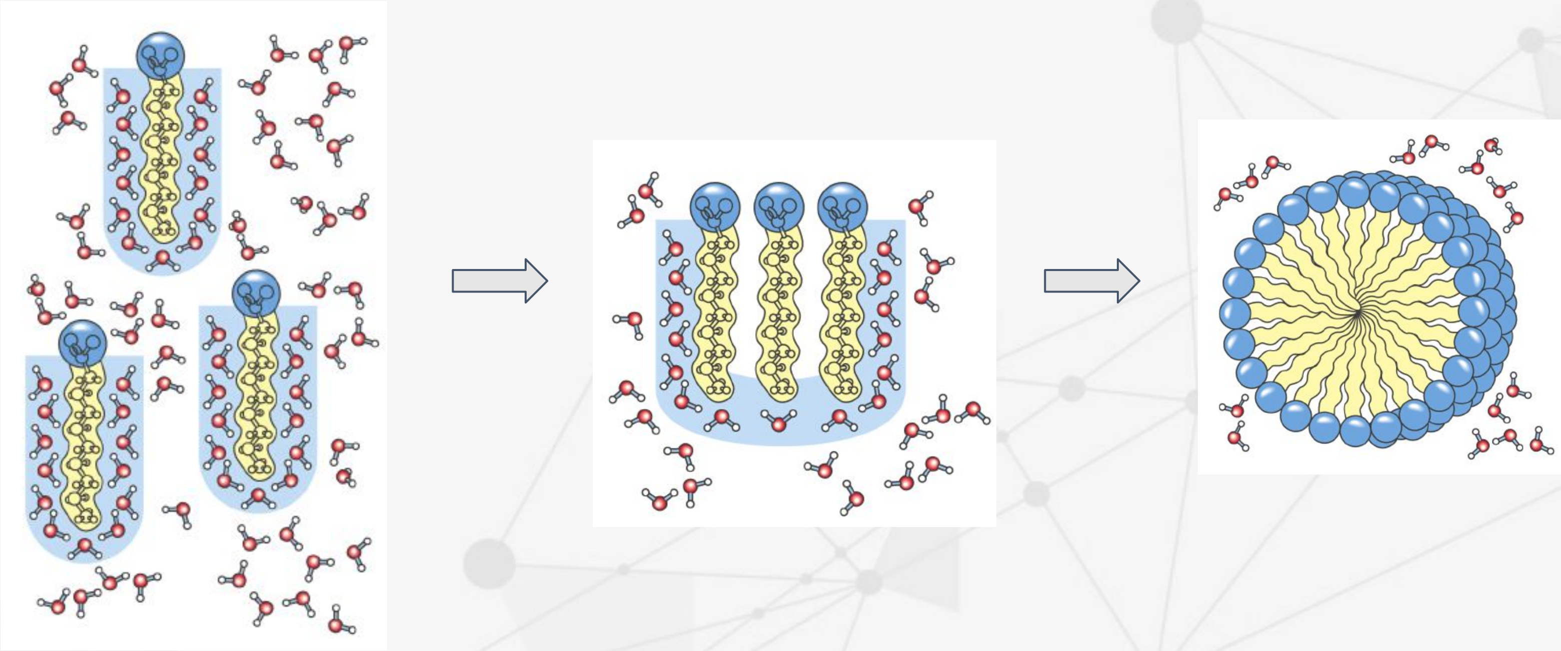


Hydrophobic effect


Hydrophilic "head group"
O
O
C
H
C
H
H
O
H
"Flickering clusters" of H_2O molecules in bulk phase
Highly ordered H_2O molecules form "cages" around the hydrophobic alkyl chains
(a)

- Water molecules form a cage-like structure around the non-polar molecule
- Positive ΔH → the cage has to be broken to transfer the nonpolar molecule
- Positive ΔS → water molecules are less ordered when the cage is broken

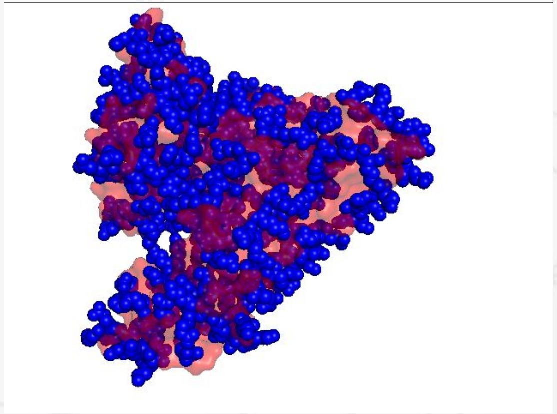
Burial of hydrophobic tails

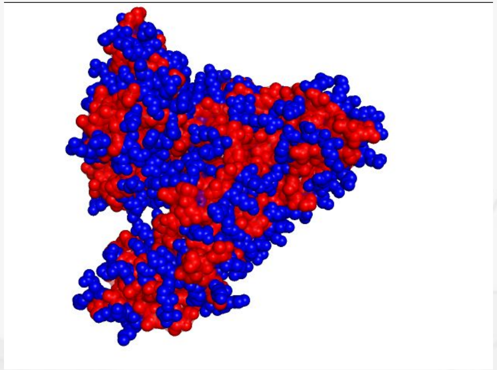


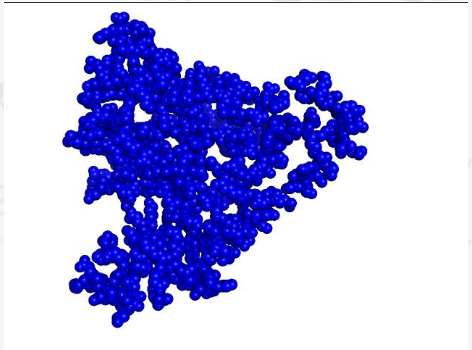
Hydrophobic core

Hydrophobic residues

(cys, ala, gly, val, ile, leu, phe, met, thr, ser, trp, tyr, pro)







PDB 1AO6
HUMAN SERUM ALBUMIN

DIPARTIMENTO
MATEMATICA

Folding pathway

Computer Science - Structural bioinformatics

2020

Levinthal's paradox

Assumptions (wrong)

- A protein sample all possible conformations (random walk)
- The conformation of a residue is independent of the rest

Statement

- The protein will never fold to its native structure

How it is possible that proteins fold in milliseconds / seconds range?

Example: Millisecond protein folding, NTL9

https://www.youtube.com/watch?v=gFcp2Xpd29I

Protein “frustration”

- A single conformation that optimizes al the interactions at the same time does not exists

Degrees of freedom

- Rotamers

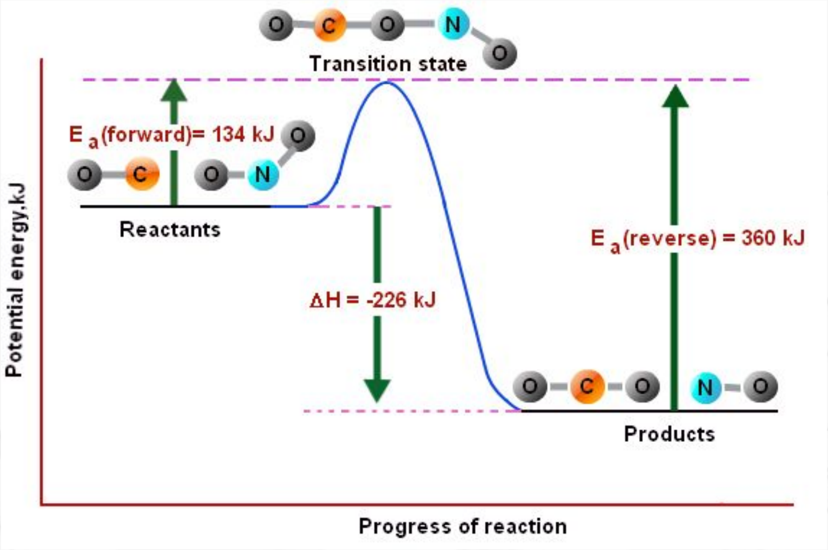
↔

Constraints

- Chain connectivity
- Different affinities of the residues for their neighbours and the environment

Simple chemical reactions

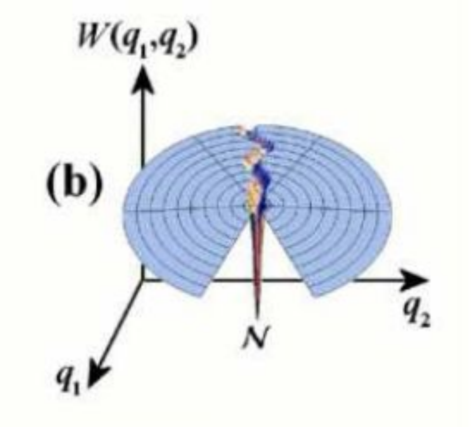
In simple chemical reactions there are steep well defined energy paths (reaction coordinate)



- *E (or G),* energy of Gibbs
- Spontaneous processes have negative *G*
- The transition state is reached when substrate molecules collide with enough kinetic energy

Protein folding pathway - “Old view” (1969)

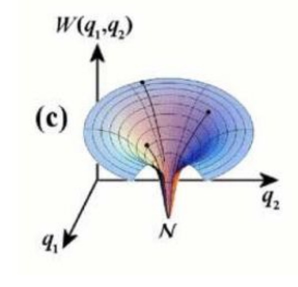
Hypothesis: as for simple chemical reactions, there are steep well defined energy paths leading to the native conformation



- q_1 and q_2 represent configuration coordinates
- $W(q_1,q_2)$ is the potential energy (Gibbs)

However, in protein folding...

- Driving **forces are weaker** and comparable to RT (unit of energy)
- Short-lived **transient interactions form randomly** and the system describes stochastic trajectories that are never the same
- The native state may be reached in **many ways**, there is not a single minimum energy path dominating over the others



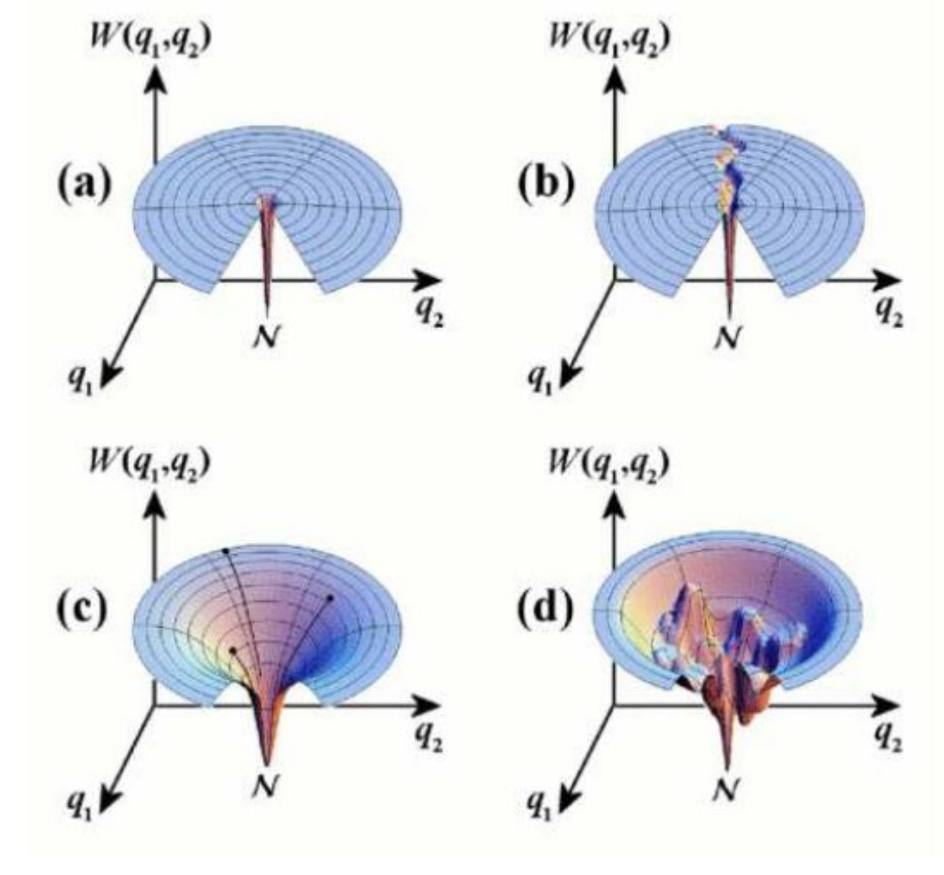
“New view” (late 80s)

- Statistical treatment in which folding is a **heterogeneous reaction** involving broad **ensembles of structures**
- Each molecule follow a partially **stochastic trajectory** determined by the intrinsic energetics of the system
- However, the probability of going towards the **native basin** is very high (+99%) and the only explanation is a **“funneled” energy landscape**
- The “old view” is a particular case of the “new view”

Conformation energy landscape

Flat golf course
(Levinthal's paradox)

Ant trail
(Old view, Levinthal's solution)



Smooth funnel
(New view)

Rugged funnel
(Realistic)

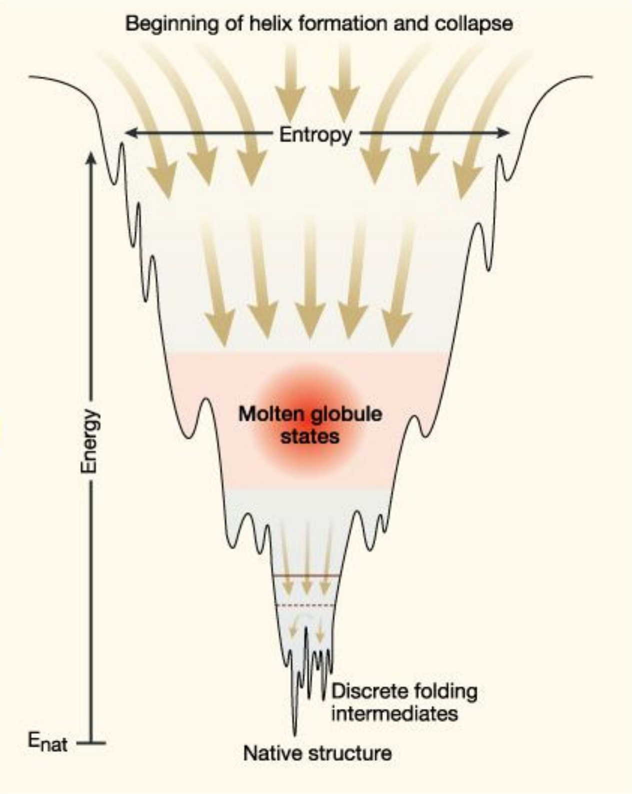
Principle of minimal "frustration"

Proteins are **not random polymers**

- They are selected and improved by **natural selection**. Random sequences will never fold
- The **score function** is the ability to fold into a **native structure** in a biologically **reasonable time**

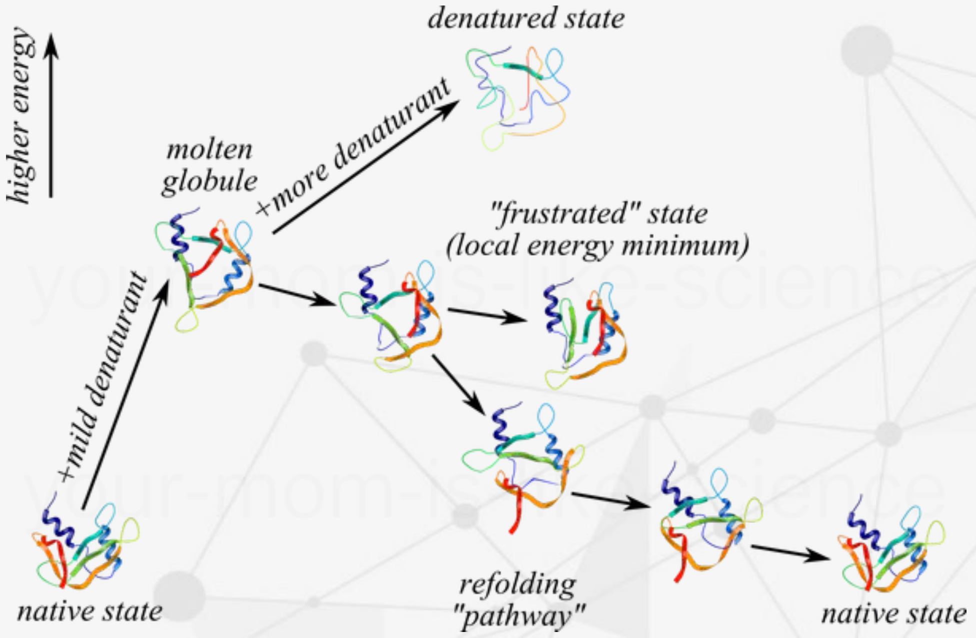
Protein sequences satisfy the **principle of minimal frustration**

- In every point of the conformational space it is more stabilizing (less energy) to form "**native contacts**"
- **Native conformation** is at **global minimum**, but proteins are marginally stable

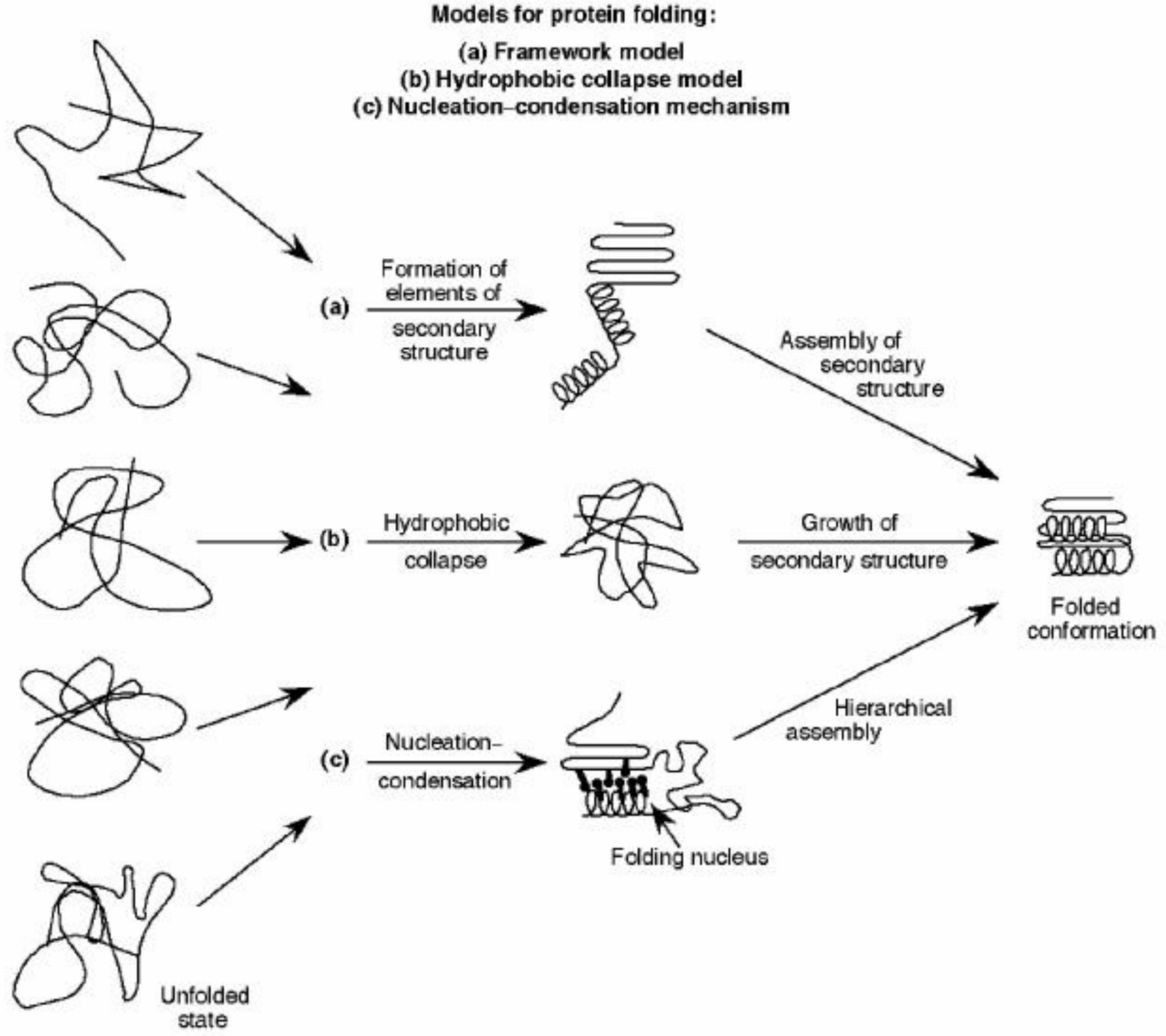

Beginning of helix formation and collapse
Entropy
Energy
Molten globule states
Discrete folding intermediates
E_{nat}
Native structure

Native conformation

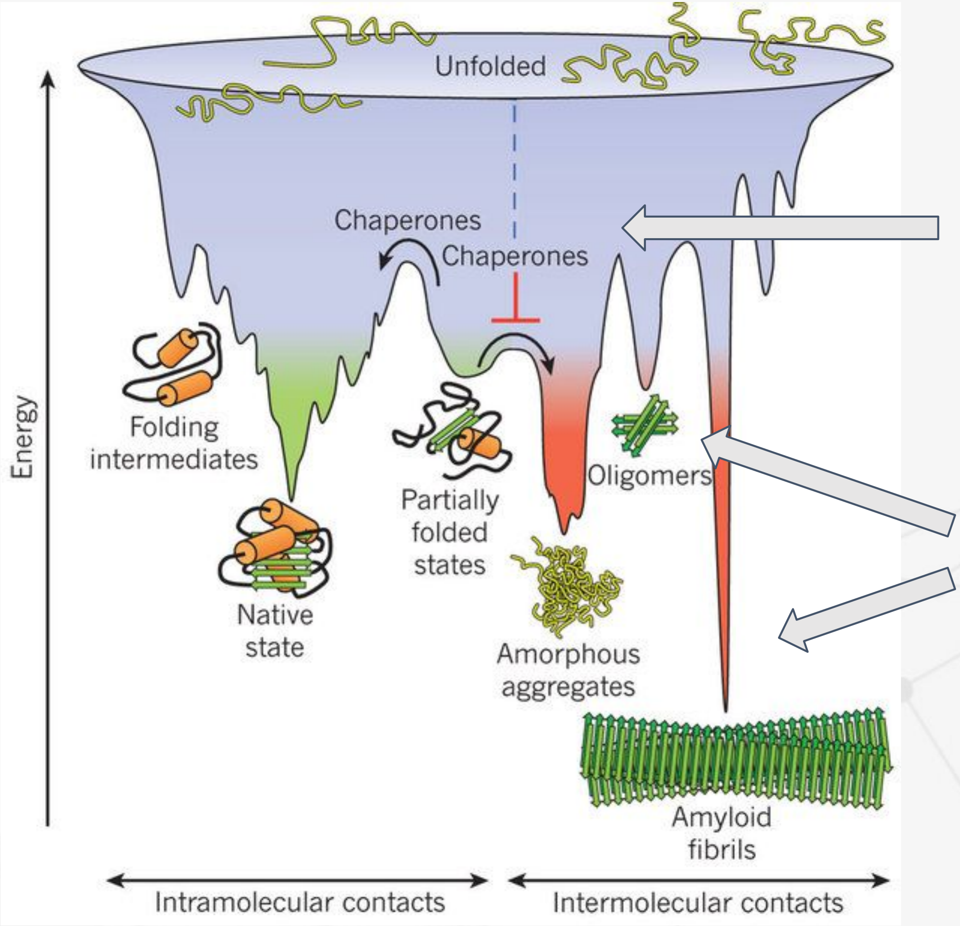
- Process driven by **non-covalent interactions** (low energy, many interactions)
- The **energy landscape** of natural sequences is **funneled** → random movements (trajectories) have high probability to make stabilizing contacts
- Random sequences will never fold → natural sequences have been selected to satisfy the **principle of “minimal frustration”**
- **Native conformation** is at **global minimum**, but proteins are **marginally stable**


higher energy
denatured state
molten
globule
+more denaturant
"frustrated" state
(local energy minimum)
+mild denaturant
native state
refolding
"pathway"
native state

Folding pathways variants


Models for protein folding:
(a) Framework model
(b) Hydrophobic collapse model
(c) Nucleation–condensation mechanism
(a)
Formation of elements of secondary structure
Assembly of secondary structure
(b)
Hydrophobic collapse
Growth of secondary structure
Folded conformation
(c)
Nucleation–condensation
Hierarchical assembly
Folding nucleus
Unfolded state

BioComputing



Chaperones are proteins that help other proteins to fold properly and prevent errors

Pathological conditions (e.g. Alzheimer)

DIPARTIMENTO MATEMATICA

References & Links: Protein folding

Introduction to protein folding for physicists

Pablo Echenique

2007, arxiv.org

https://arxiv.org/abs/0705.1845

TMP Chem (Trent Parker’s YouTube channel)

https://www.youtube.com/user/TMPChem

PlayLists: PChem Math, Chemical thermodynamics, Computational Chemistry