Lecture 16:
Structure of random and quasi-random amino acid sequences

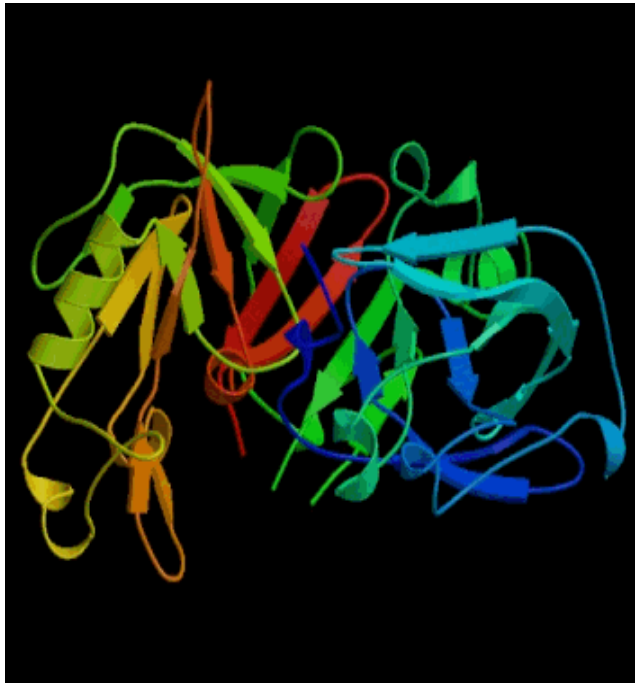
Lecturer:
Prof. Brigita Urbanc (brigita@drexel.edu)

What is the relationship between the protein stability and structural regularities?

- water-soluble globular proteins
- protein globule compact: α-helical & β-structure segments connected by irregular segments
- irregular segments slide over the surface: no inter-crossing regular structures
- high stability of the protein globule → "the multitude principle" (large # of sequences fit the same architecture)
- diverse amino acid sequences associated with globular protein structures
- typically, the relative number of hydrophobic versus hydrophilic residues about the same

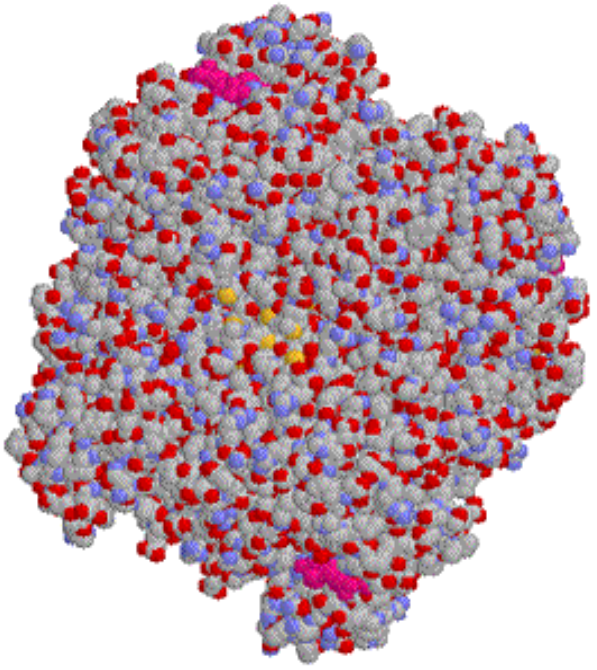
Some examples of globular proteins:

pepsin



luciferase

enzyme catalase



– the primary structure as a "random co-polymer" with a random sequence of hydrophobic/hydrophilic residues

Are random sequences compatible with compact fold into a protein globule?

- **consider a random sequence of hydrophobic & polar residues (i.e. random co-polymer)**
- **to include an α-helical or β-structural segment, the segment needs to have a continuous hydrophobic surface**
- **an α-helical surface: residues i–(i+4)**
- **a β-sheet surface: residues i–(i+2)**

What is the distribution of non-polar groups in the sequence?

- **p – a fraction of non-polar groups in the sequence**
- **(1-p) – a fraction of polar groups in the sequence**
- **r – the number of non-polar residues/groups in a row**

- **probability W(r) to have r non-polar residues with two polar residues at the ends:**

$$W(r) = (1-p)\, p^r\, (1-p)$$

- **the hydrophobic surface of an α- or β- segment forms if $r \geqslant 2$**
- **calculate the average value of r, <r>:**

$$\langle r \rangle = \{\Sigma_{r\geqslant 2}[rW(r)]\}/\{\Sigma_{r\geqslant 2}W(r)\}$$

- **the summation of series: $\Sigma_{r\geqslant 2}[rp^r]$ and $\Sigma_{r\geqslant 2}p^r$**

(see any mathematical handbook)

- **the result: $\langle r \rangle = 2 + p/(1-p)$**
- **at p = ½ (equal fractions of hydrophobic and hydrophilic aa):**

$$\langle r \rangle = 3$$

Open circle (hydrophilic), filled circle (hydrophobic):

○○●●○●●○●○●○○○○●●○●○●●○○●○●●

↓

Random sequences are capable of folding into at least a two-layer arrangement of secondary structures!

- **random sequence provides continuous hydrophobic surfaces that can form α-helices or β-structures**
- **these regular structures folded inside the hydrophobic core surrounded by short loops on th surface**
- **these results consistent for up to ~150 residue sequences**

- **the primary structure as a "random co-polymer" with a random sequence of hydrophobic/hydrophilic residues**

QUESTIONS:

- **why can an "energetic defect" of a few kcal/mol prohibit many protein architectures?**
- **how are "entropic defects" related to the almost fixed native protein structure?**

Quasi-Boltzmann Statistics of Small Elements of Protein Structures:

$$\text{occurrence} \sim \exp(-F/k_B T_C),$$

where T_C is somewhere between the room (300K) & melting (370K) temperature

Experimentally found free energies of transfer of residue side groups from non-polar solvent to water

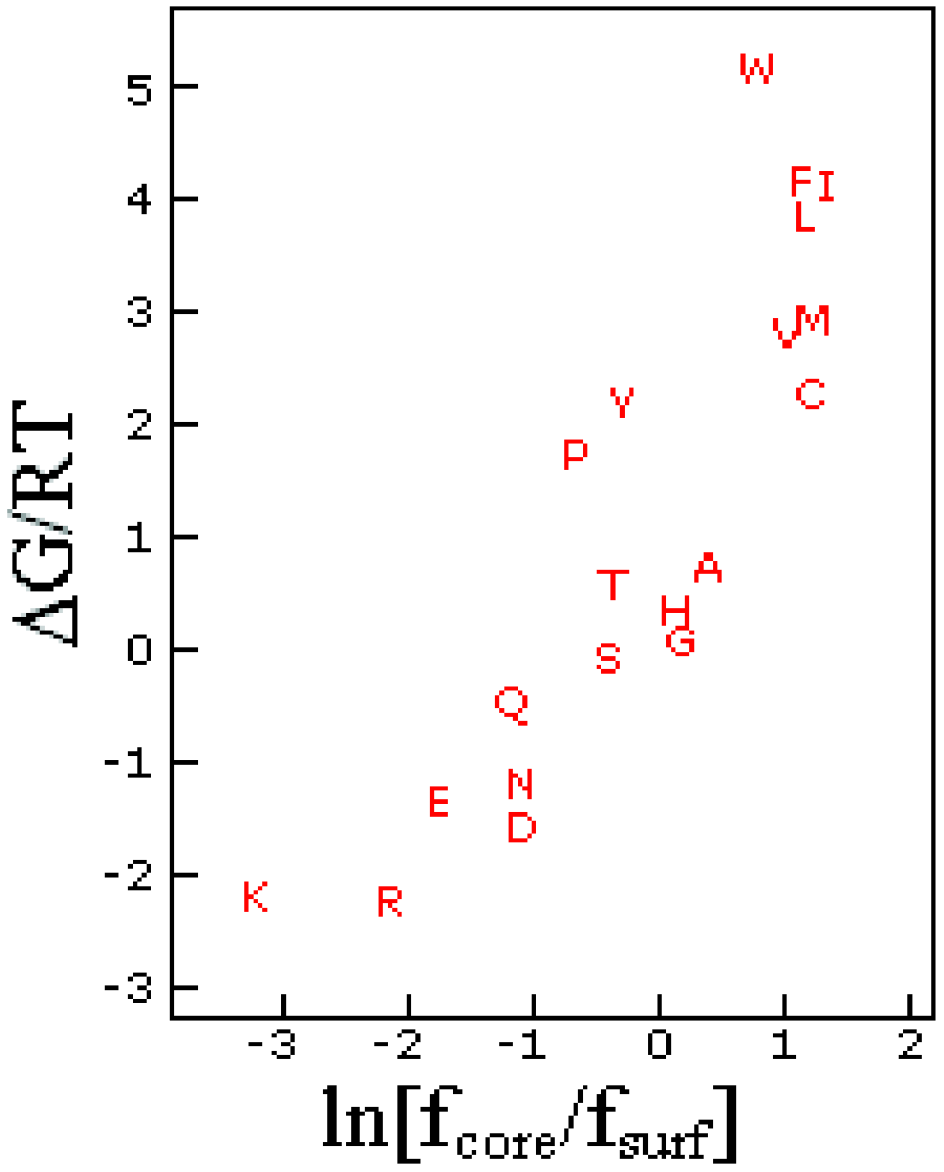
f_{core} (occurrence frequency in the core)

f_{surf} (occurrence frequency at the surface)

⇩

almost linear relationship with a slope of 1-1.5 ⇨

similarity to Boltzmann statistics



Because of the analogy to the Boltzmann statistics, the protein structure statistics are used to estimate the free energy of interactions between amino acids.

PHYSICS INTERPRETATION: the true Boltzmann statistics describes temporal fluctuations in a 3D space, In the native state of the globular protein one particular amino acid is at ALL TIMES in the same position relative to the center of mass (no moving from the core to the surface and back).

- **consider a protein with Leu in the interior of the stable globular structure**
- **how does the internal free energy the structural element (e.g. at the position of Leu) affect the # of protein sequences capable of stabilizing the protein**

How does Leu → Ser mutation in the protein interior change the # of fold-stabilizing sequences?

Assumptions:

1. **only one folded state exists**
2. **only hydrophobicity of residue X(Leu) relevant**
3. **internal a.a. 100% screened and external exposed to H_2O**
4. **Unfolded state: all a.a. 100% exposed to H_2O**

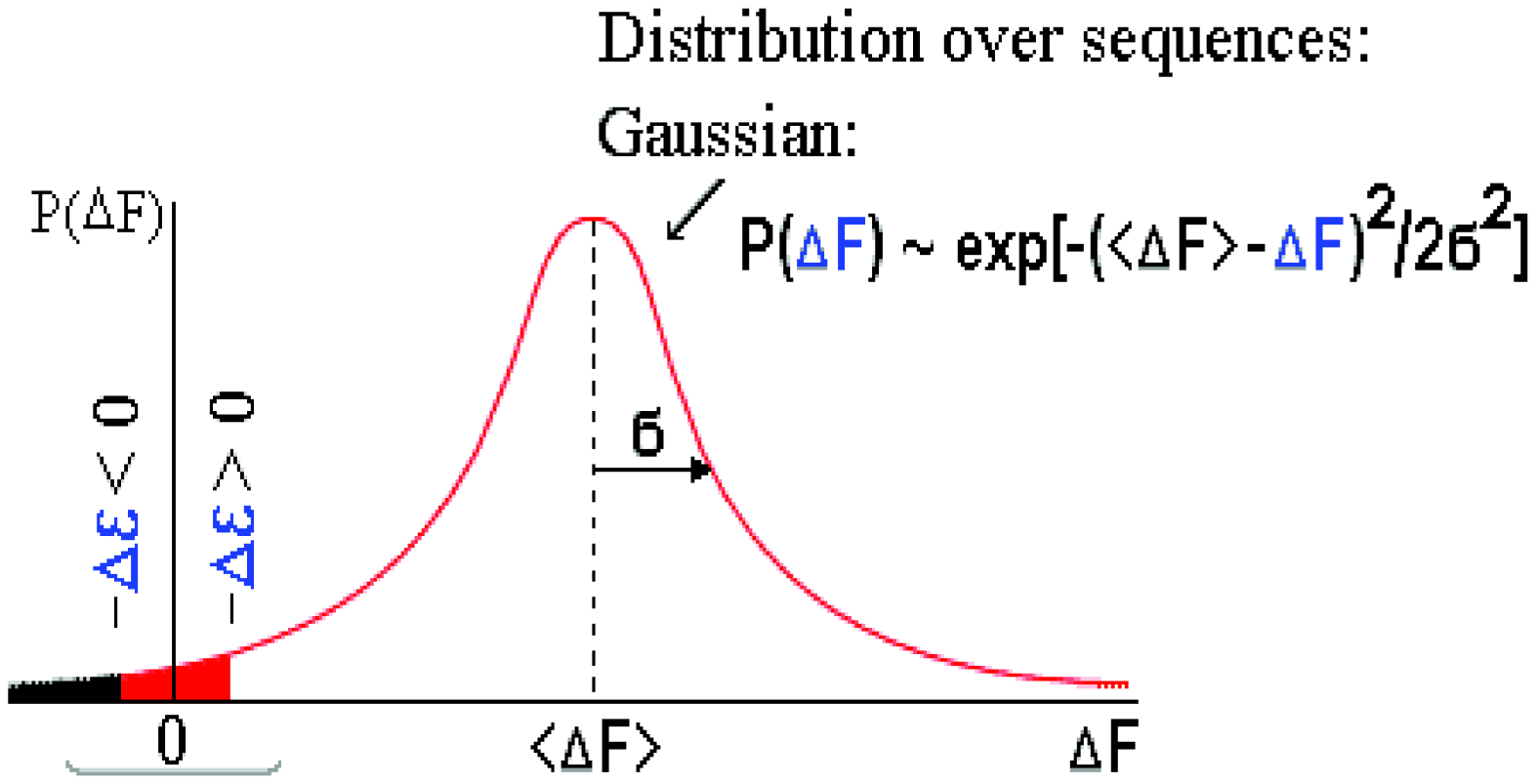
Leu → Ser mutation: What is the free energy transfer from hydrophobic environment to water for each?

- **Ser: 0 kcal/mol & Leu: ~2 kcal/mol**
- **the free energy different between the folded and unfolded states: $\Delta\varepsilon + \Delta F < 0$ ($\Delta\varepsilon$ – Leu contribution & ΔF – rest of the chain**
- **$\Delta\varepsilon$ & ΔF – depend on amino acid sequence**
- **the folded state stable if $\Delta F < -\Delta\varepsilon$**
- **the probability P* that $\Delta F < -\Delta\varepsilon$:**

$$P^*(\Delta F < -\Delta\varepsilon) = \int_{-\infty}^{-\Delta\varepsilon} P(\Delta F)d(\Delta F)$$

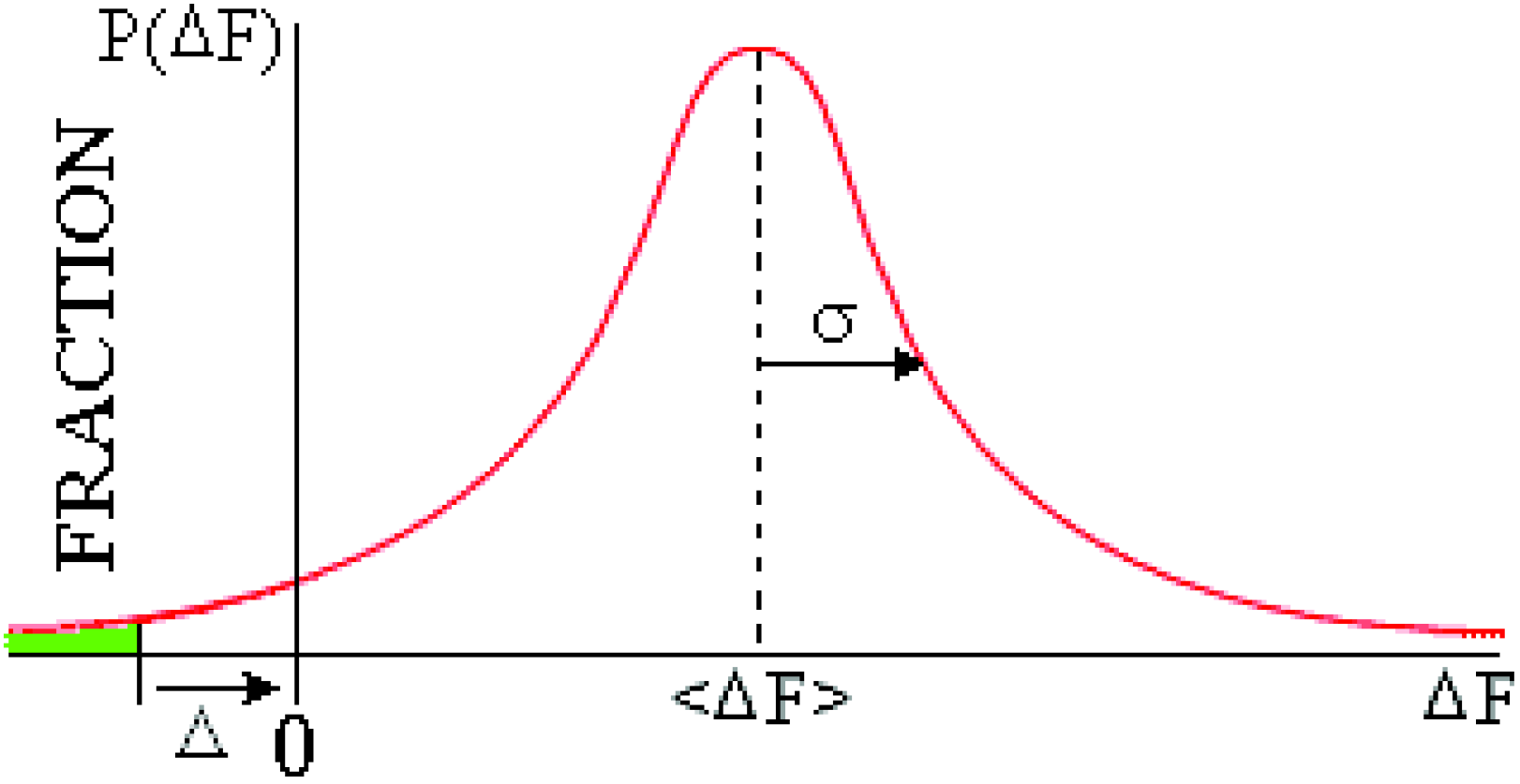
- **$P(\Delta F)$ – probability of ΔF for a random sequence**

Gaussian distribution for P(ΔF) [central limit theorem] & approximation for ΔF « <ΔF>:



Near 0: $P(\Delta F) \sim \exp[-\langle \Delta F \rangle^2/2\sigma^2] \cdot \exp[\Delta F \cdot (\langle \Delta F \rangle/\sigma^2)]$

The fraction of a.a. sequences that have a folded state:



$$P^*(\Delta F < -\Delta) \approx \text{const} \times \exp[-\Delta/(\sigma^2/\langle\Delta F\rangle)]$$

$$\text{const} \times \exp[-\Delta/k_B T_C]$$

Because both σ^2 and $\langle\Delta F\rangle$ are proportional to the protein size (# of amino acids in the sequence), $\sigma^2/\langle\Delta F\rangle$ is independent of the protein size: $\sigma^2/\langle\Delta F\rangle = k_B T_C$

$\sigma^2/\langle\Delta F\rangle = k_B T_C$ – average energy of non-covalent INTs per residue in the sequence

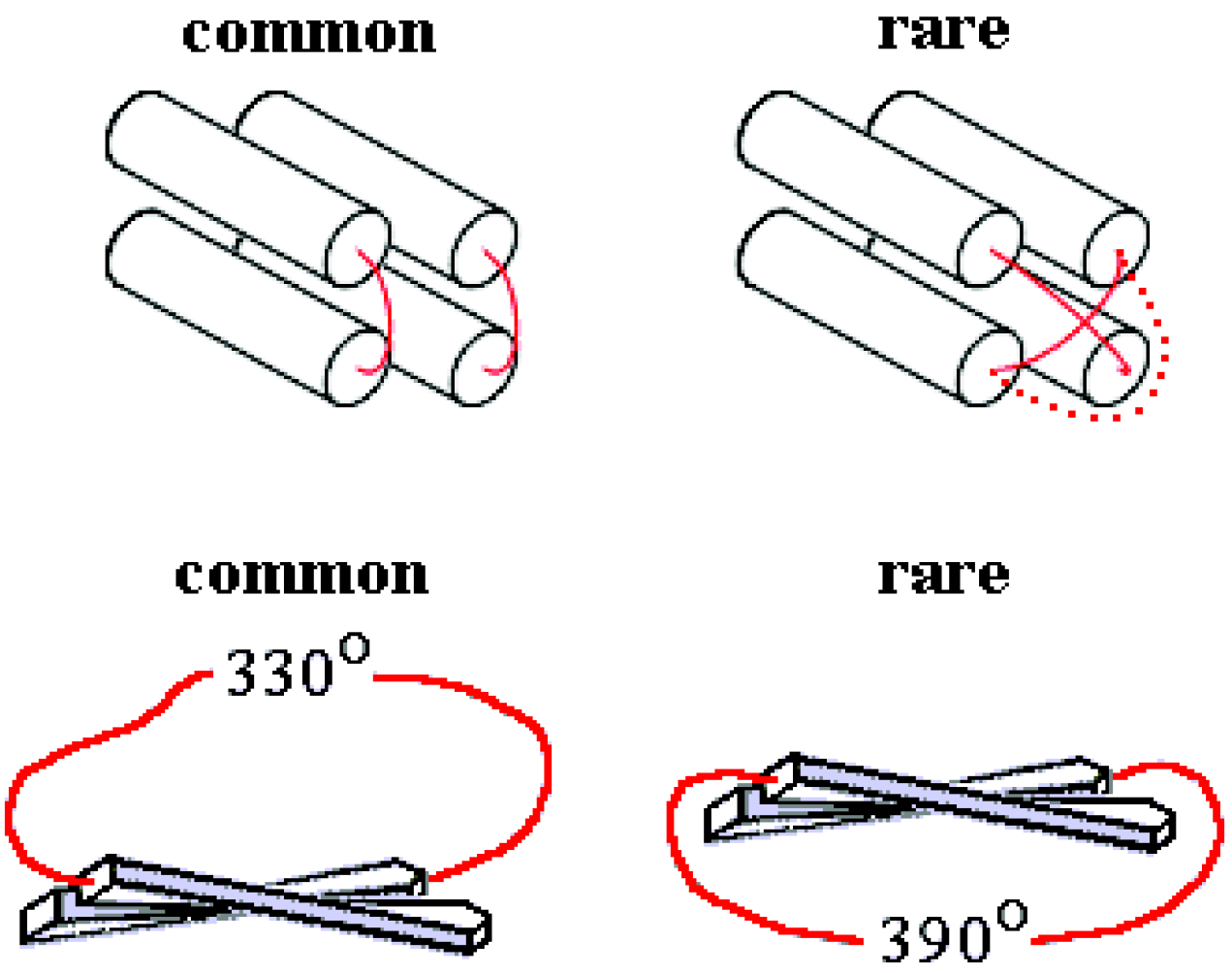
~ 0.5 – 1 kcal/mol ($300K < T_C < 370K$)

Any energetic defect of several kcal/mol needs to be compared to $\sigma^2/\langle\Delta F\rangle = k_B T_C$ =0.5 – 1 kcal/mol.

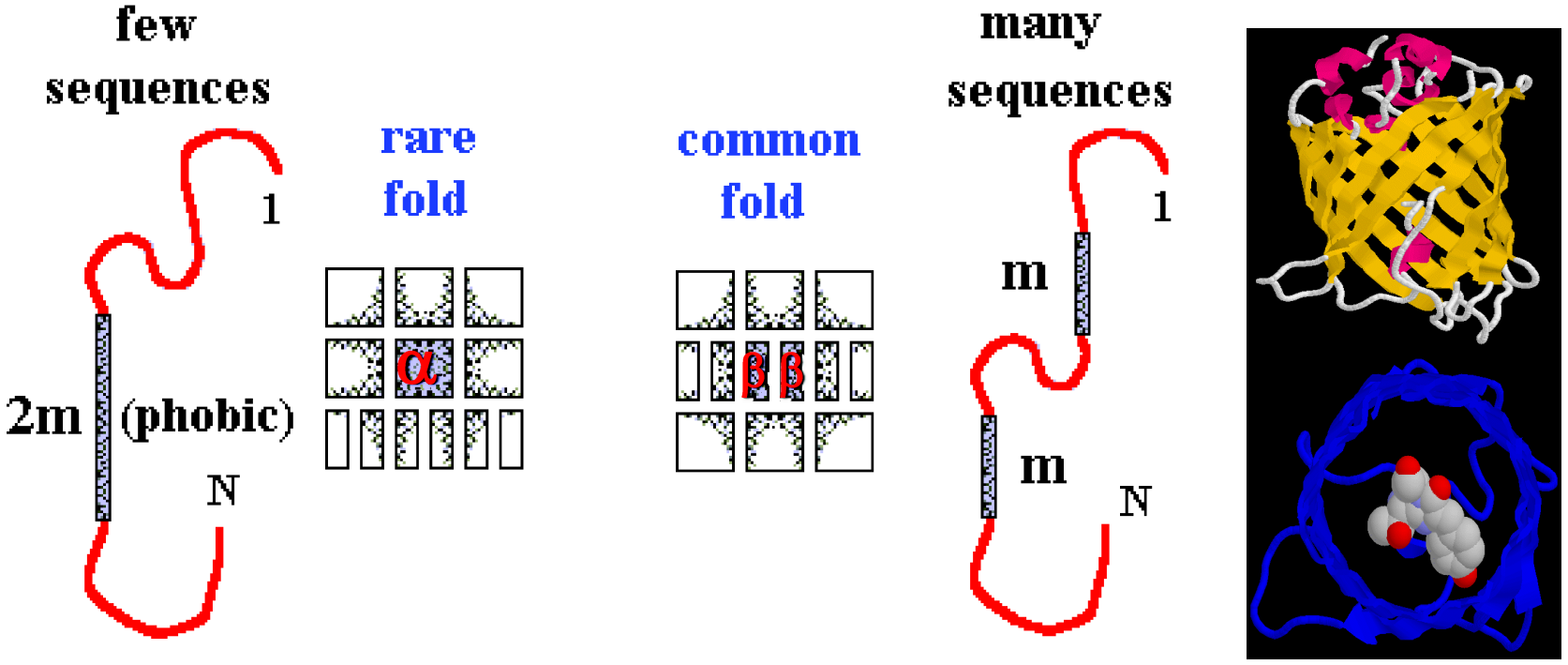
Defect of

- **1 kcal/mol decreases the # of possible sequences by 5**
- **2 kcal/mol decreases the # of possible sequences by 20**

Defects: Loop Crossing & Left Handed Twist are Rare:



A multiple-layer packing with an α-helix in its center: RARE (exception: green fluorescent protein, GFP)



The center needs to include hydrophobic amino acids

(per length, α-helix contains 2×residues than β-strand)

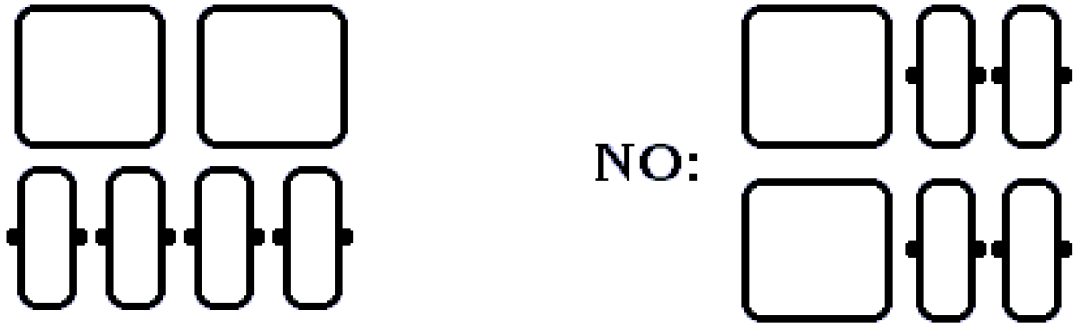
(a) p^{2m} – probability of one 2m–residues long HP block

(b) $p^{m} \times p^{m}$ – probability of two m-residues long HP blocks

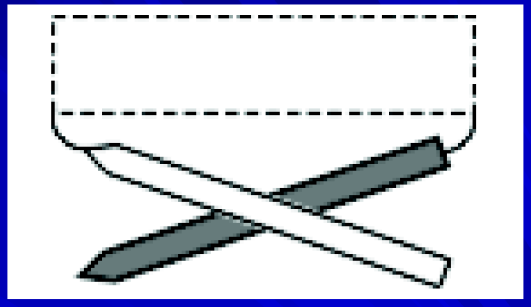
→ BUT: the # of realizations of (a) ~N and (b) ~N × N/2

EMPIRICAL RULES for FREQUENT FOLDS

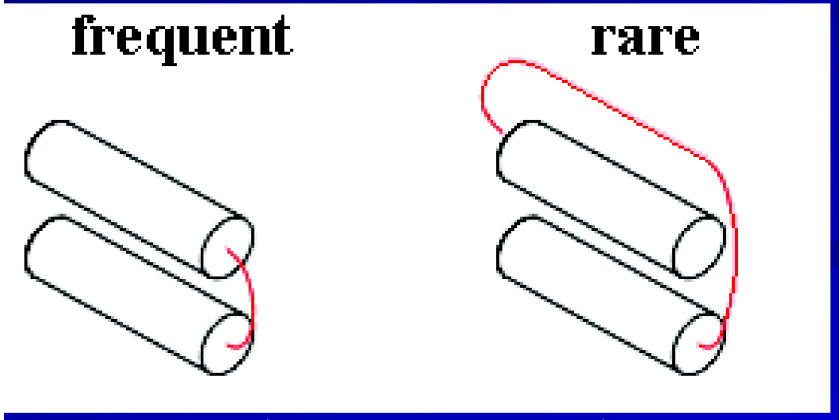
α and β structures,
separate α and β layers



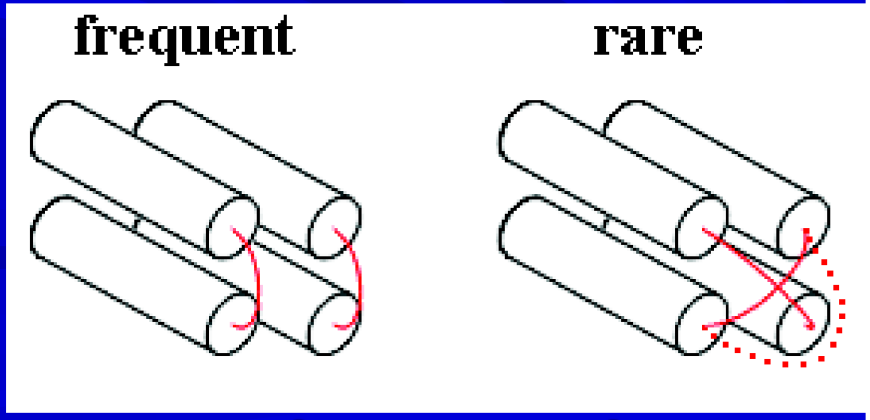
right-handed
superhelices



Lost H-bonds: defect!



no large (360-degree) turns



no loop crossing

The Physics Protein Structures – Summary

- **protein structure classified into only limited number of folding patterns**

- **globular, water-soluble proteins, made of "random" amino acid sequences: –HPPHHPHPPPPHHPHHH–**

- **rare folding patterns associated with energetic or entropic defects of only ~1–5 kcal/mol**

- **occurrence of a given structural element in a stable fold:**

$$\sim \exp(-\Delta\varepsilon/k_B T_C),$$

where $k_B T_C \sim$ 0.5–1 kcal/mol ($\Delta\varepsilon$ compared to the free energy of melting : several 100 kcal/mol)