Outline

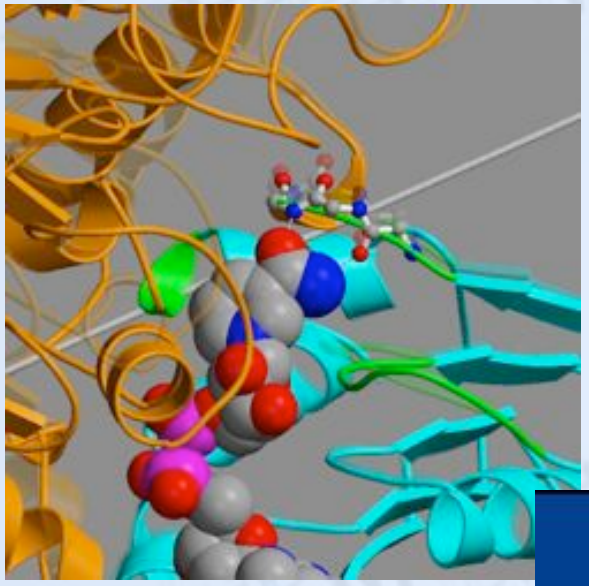
- Why Biomolecular Simulation?
  - 50 % working on biomolecules
  - Physicist difficult to interpret what biologist my find interesting or obvious
- Case studies of lattice proteins
  - simple model
  - provides a lot of the issues that are relevant
  - Introduction to practicum
- Biology
  - What are proteins?
  - What physical properties make them unique?
- Lattice proteins
  - Design
    - Discussion on evolution
  - Folding
  - Potential
    - Discussion ob evolution
- Case studies
  - Binding & Folding (?)
  - Disordered flanks
  - Chaperones
- Lessons for full atom MD on proteins

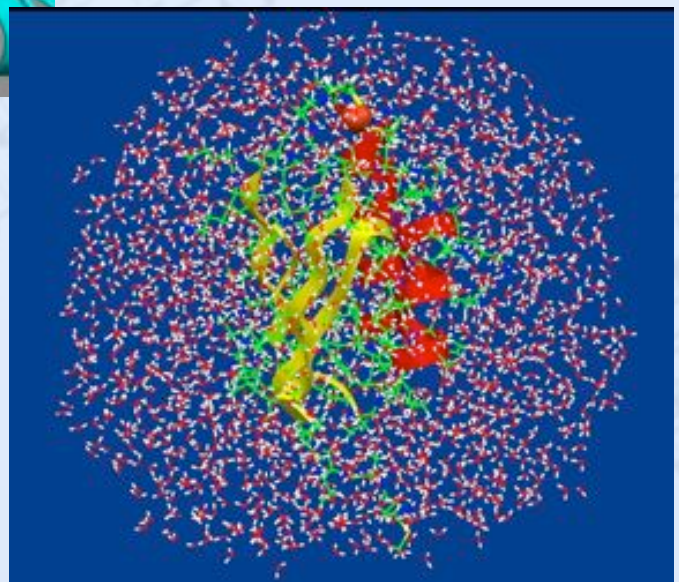
Biomolecular Simulation

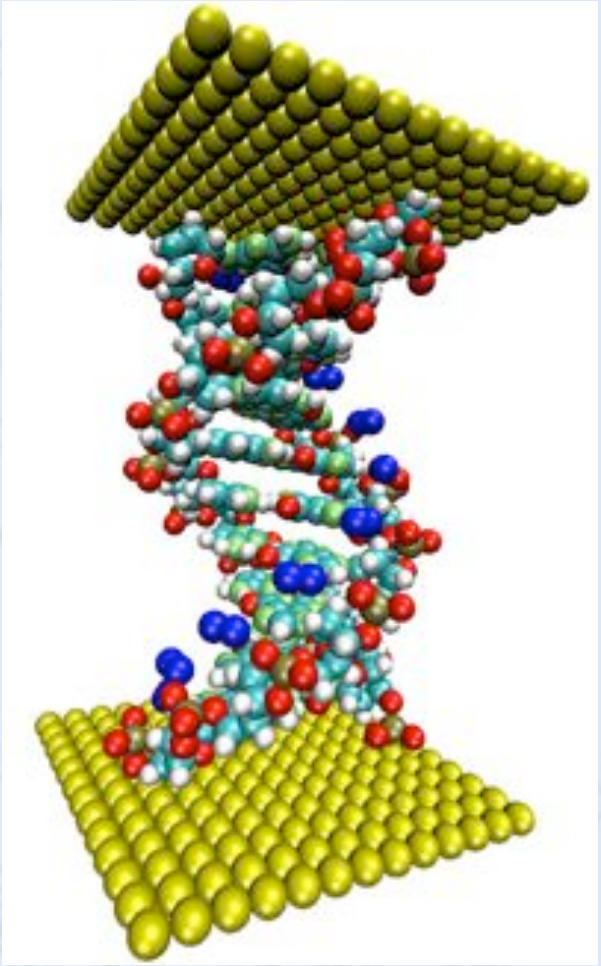
Should we consider evolution?

Sanne Abeln

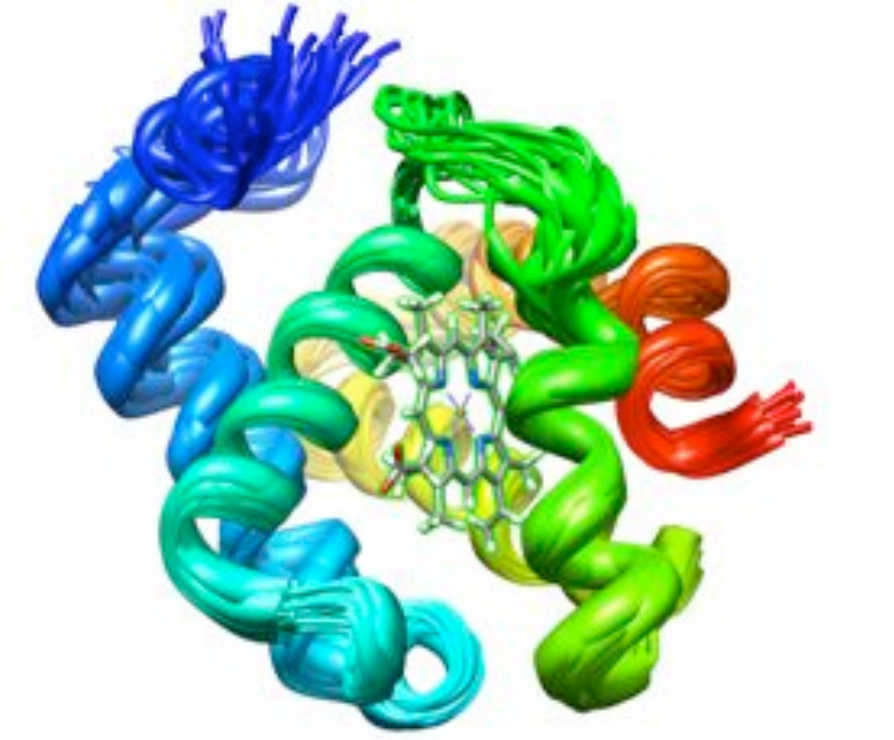
Why “Biomolecular Simulation”?





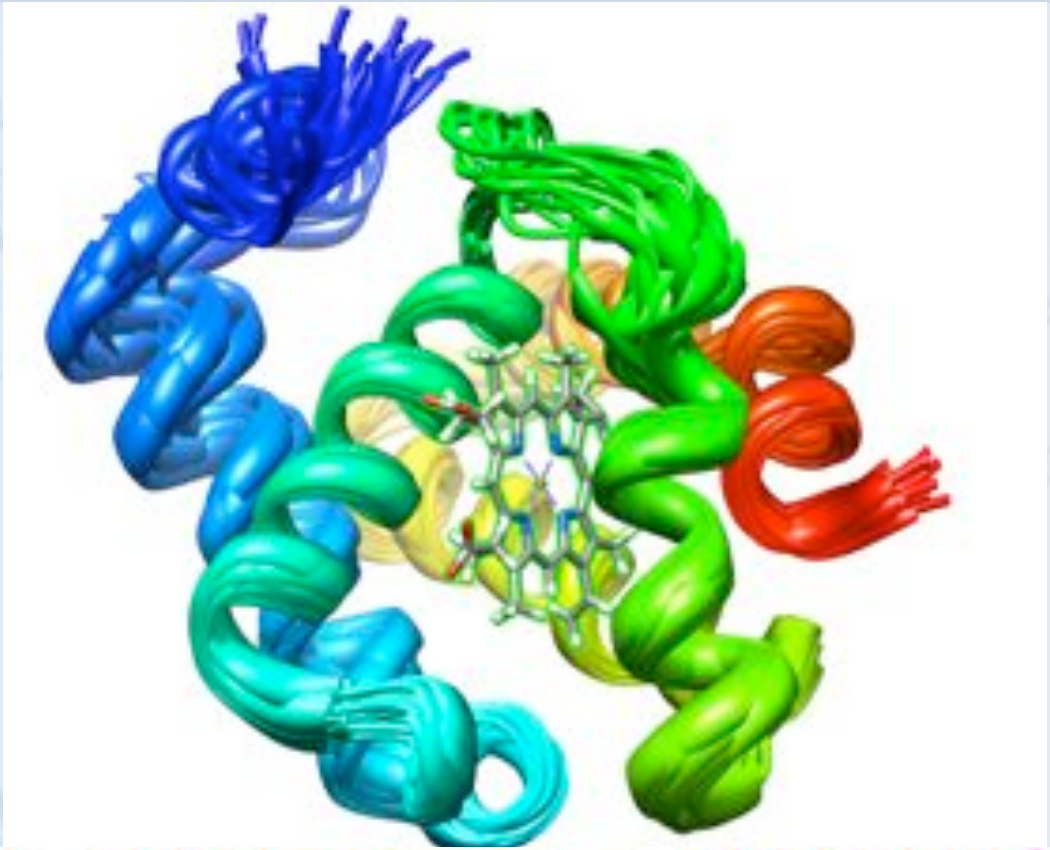


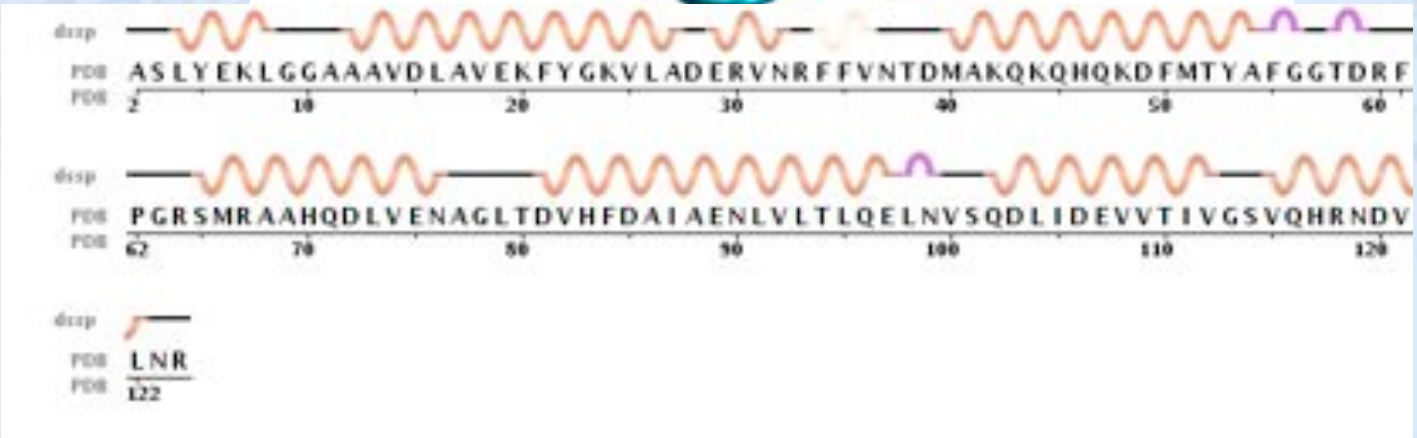
Why “Biomolecular Simulation”?



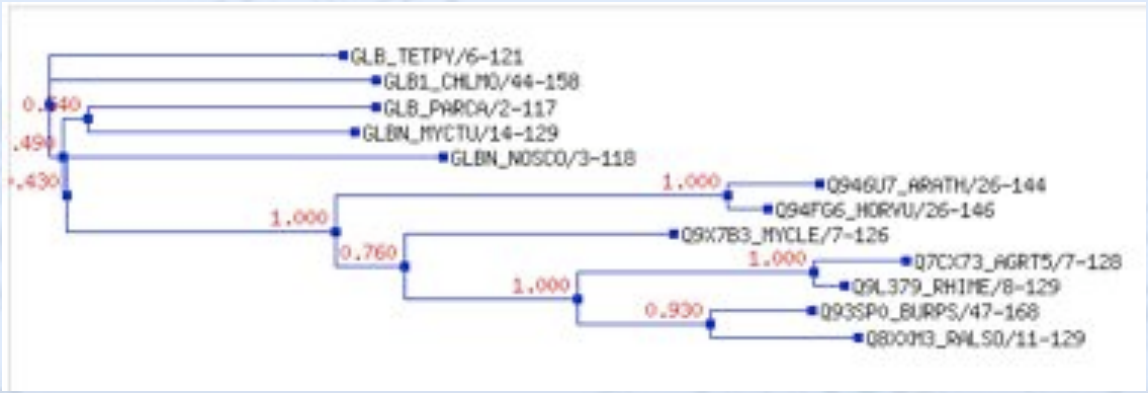
- Physical questions:
  - How stable is this protein
  - Under which conditions will this protein fold?
  - How strong is the binding to a substrate?
- Biological questions:
  - What is the function of this protein in the cell?
  - What happens if we changes the sequence of the protein?
  - Where does the substrate bind?
  - Do evolutionary related proteins bind the same substrate?

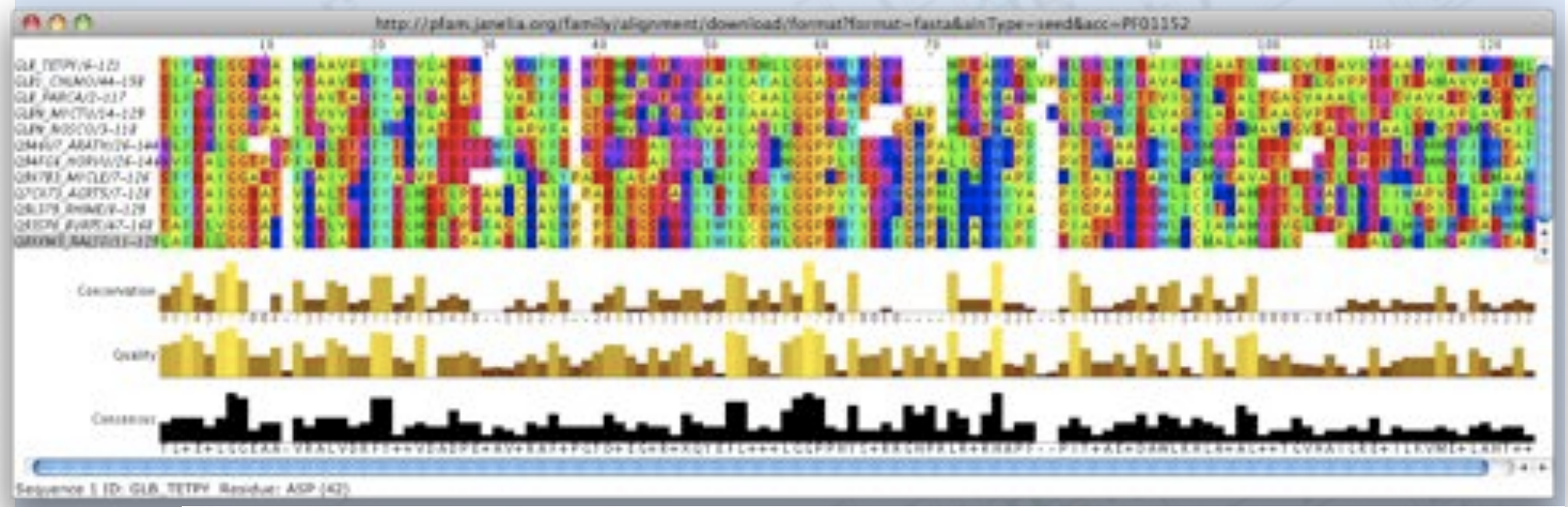
The biology: a sequence

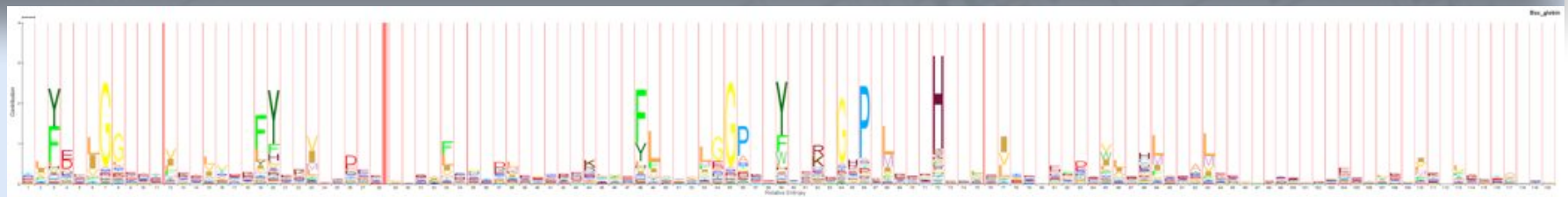




The biology: a tree

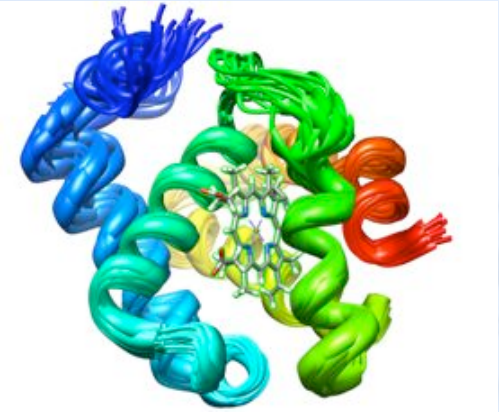


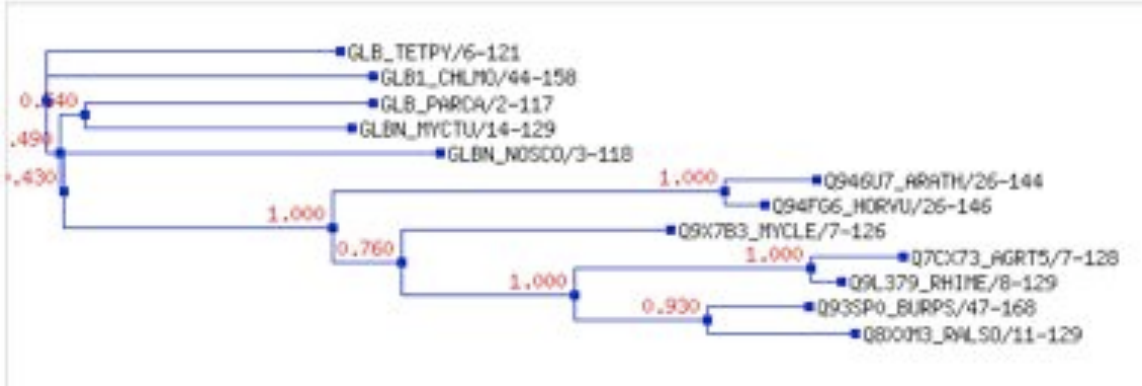


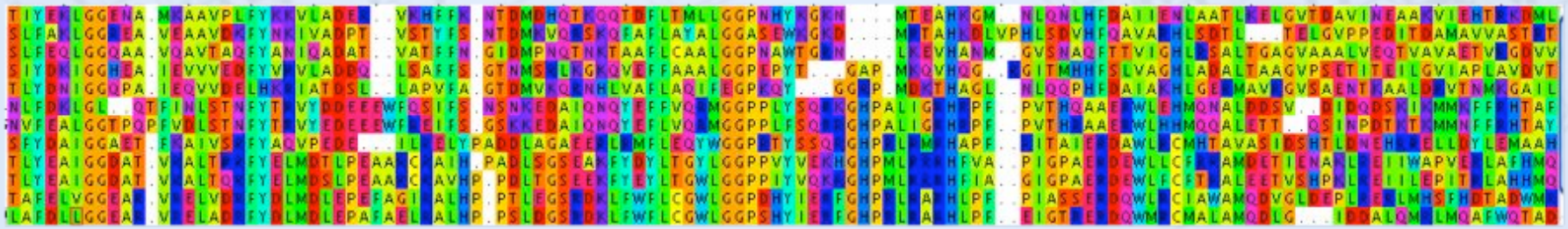


Biology & Physics:

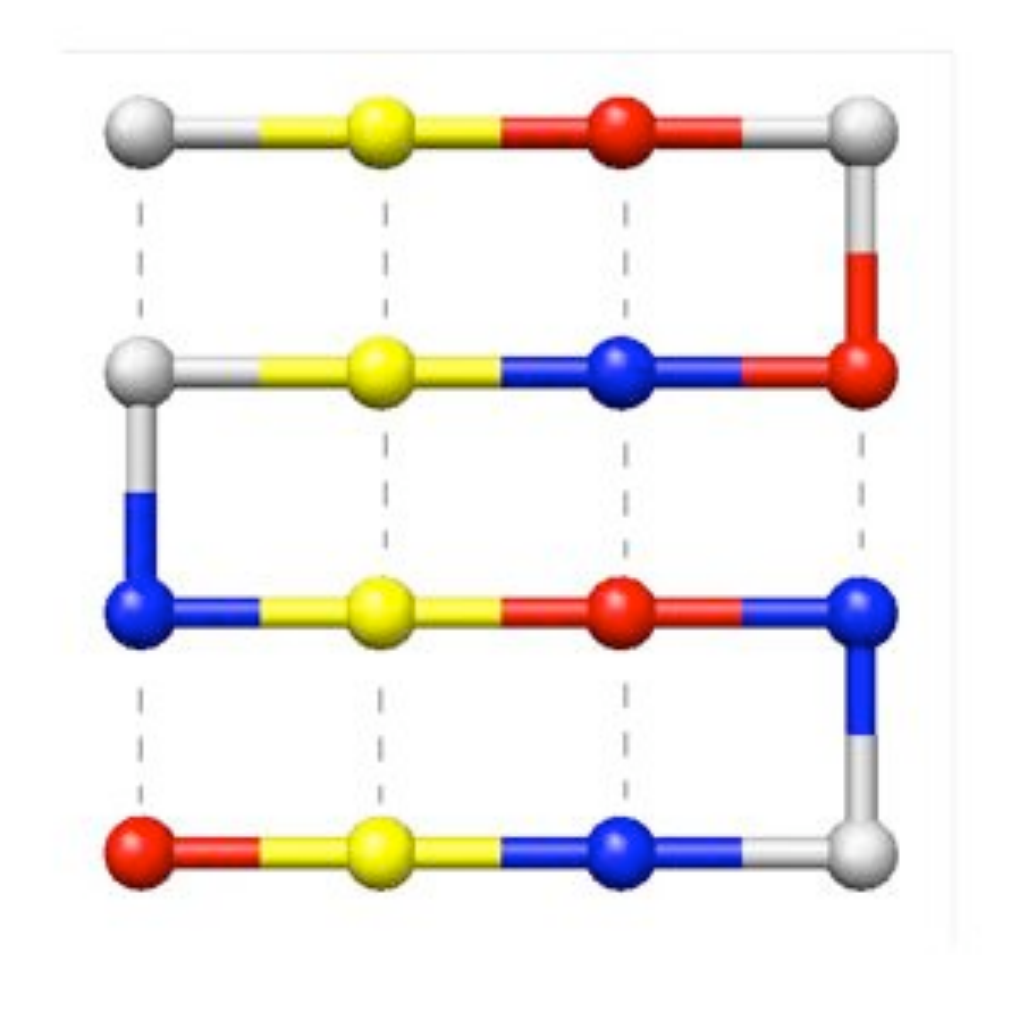
- What consequences does the biology have for our physical questions?
- How can the biological context help to answer our questions?



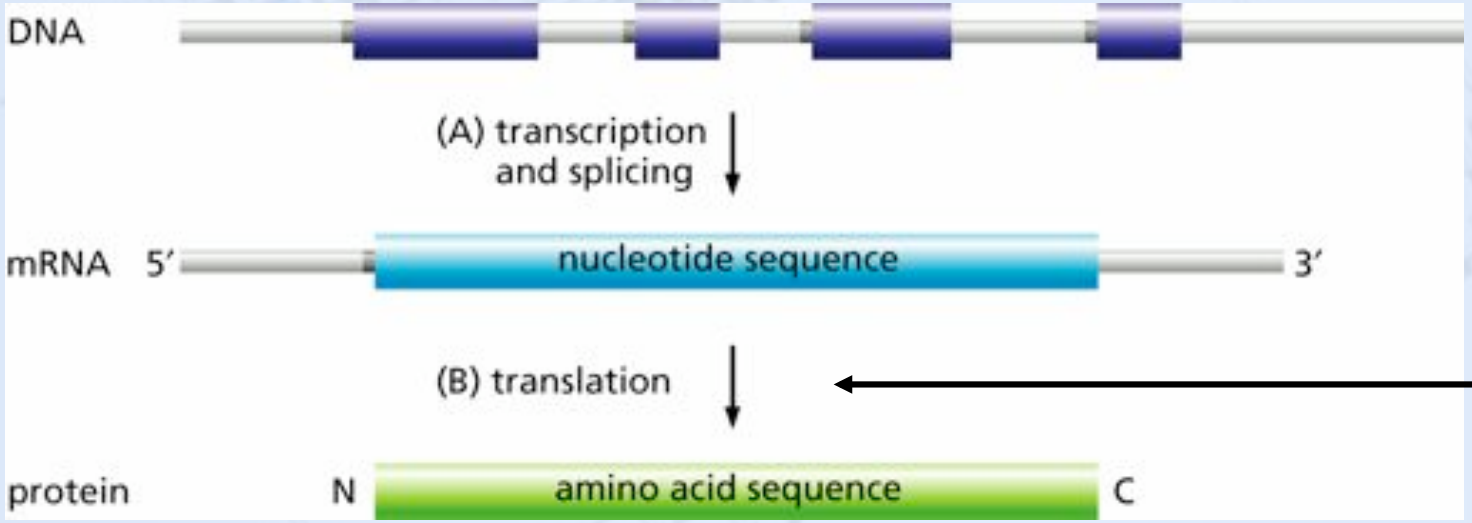




Practical - folding of a simple lattice protein:



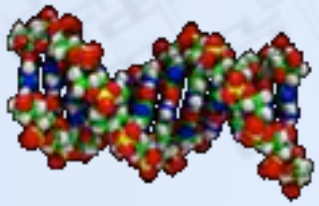
DNA - RNA - Proteins



First Position	Second Position: U		Second Position: C		Second Position: A		Second Position: G		Third Position
	code	Amino Acid	code	Amino Acid	code	Amino Acid	code	Amino Acid	
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

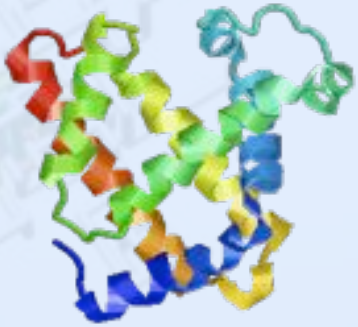
DNA - RNA - Proteins - function

DNA →

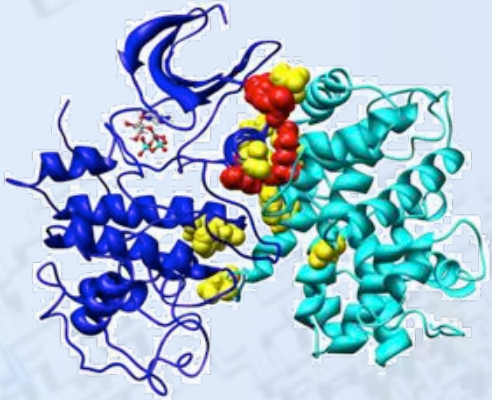


RNA codon

amino acid sequence



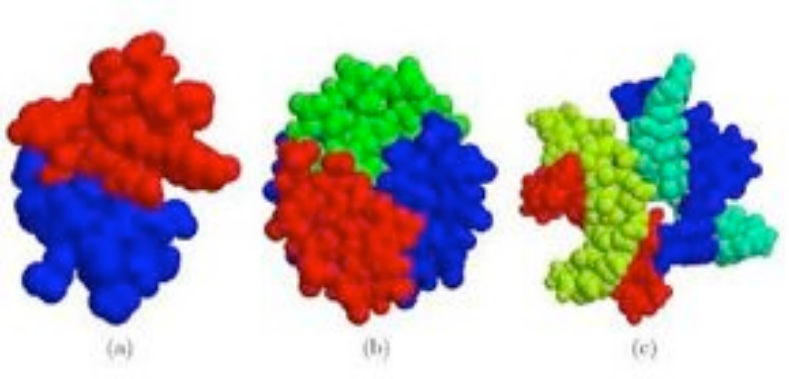
folded protein →



function

Structure & Function

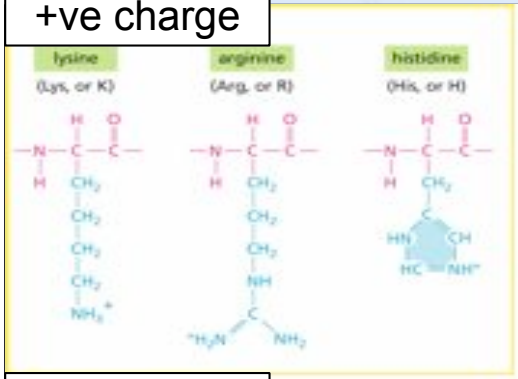
- Protein binding sites:
  - Catalyze metabolic reactions (enzymes)



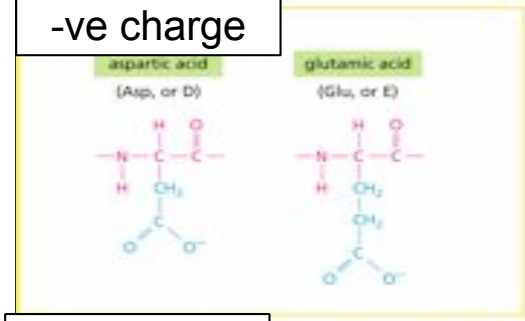
- Protein-protein interactions
  - Signaling / Complex formation

Primary Protein Structure

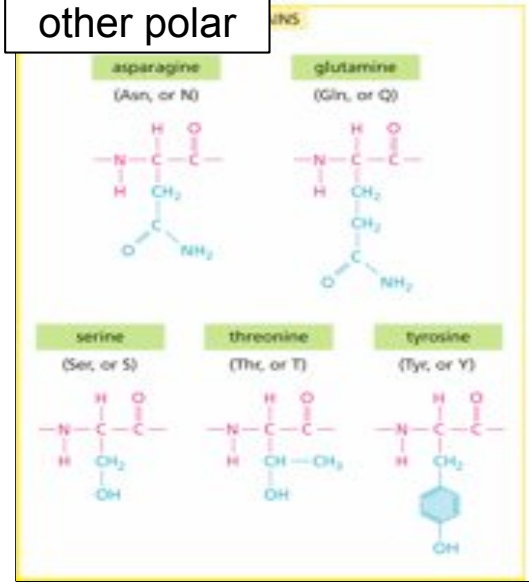
+ve charge



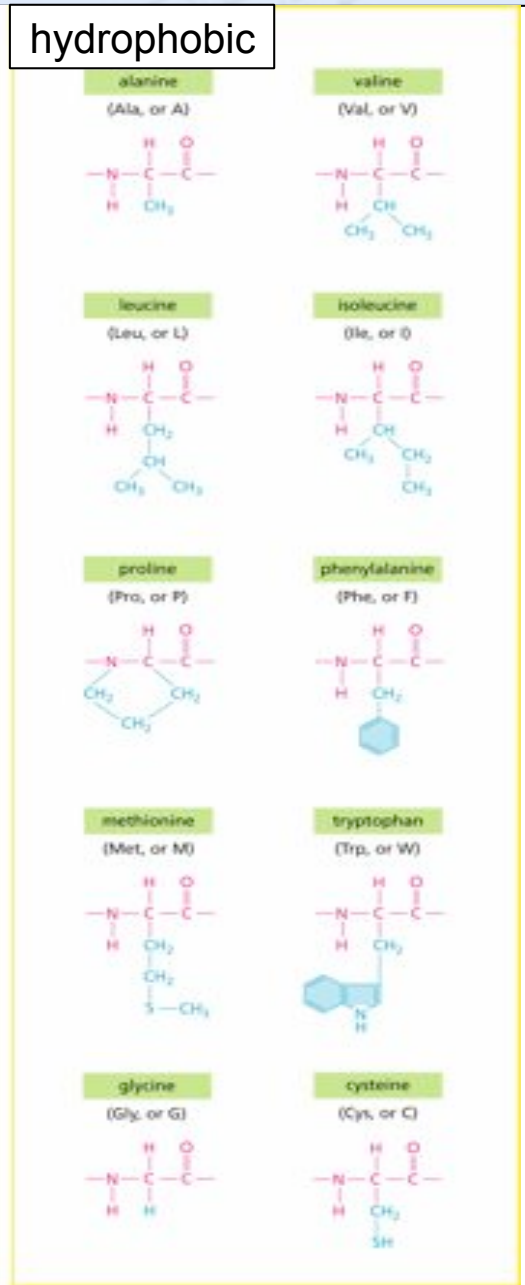
-ve charge



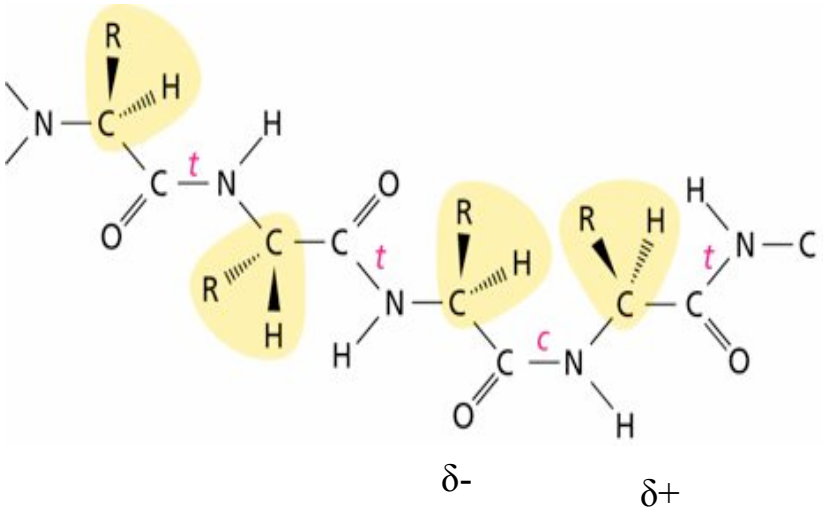
other polar

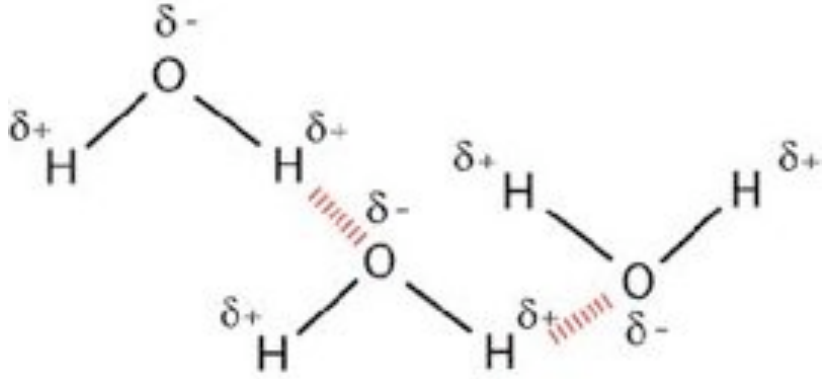


hydrophobic



Proteins live in water





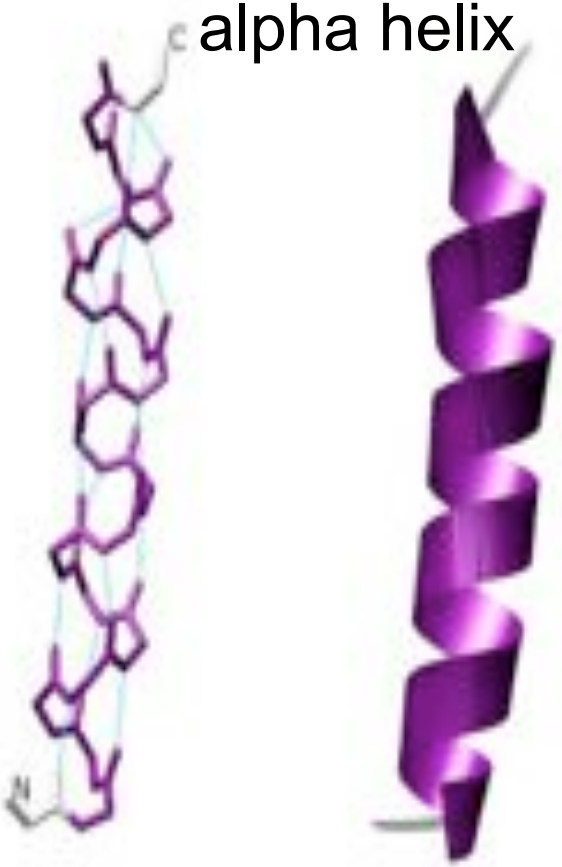
Why do proteins fold?

1 Hydrophobicity (oil in water)

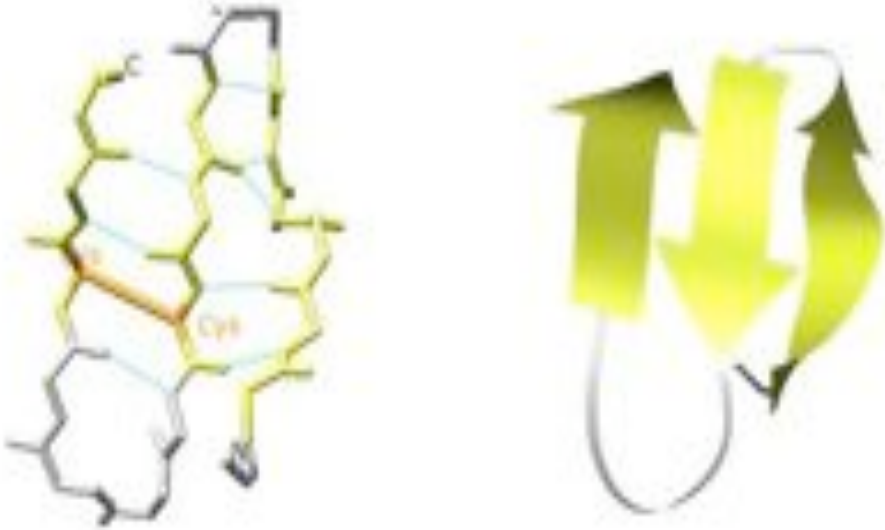
2 Hydrogen bonds form secondary structure

Hydrogen Bonds & Secondary Structure

alpha helix

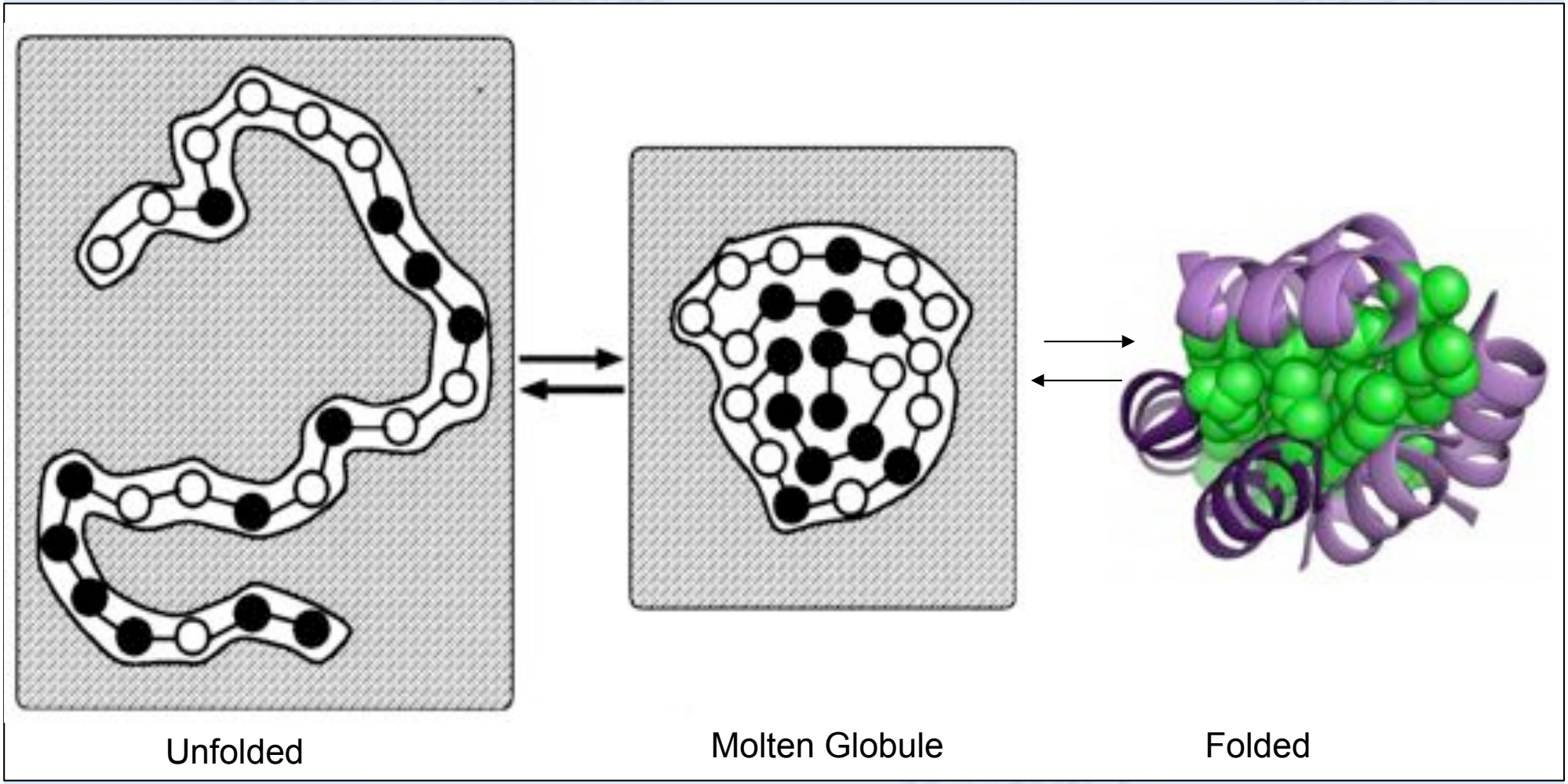


beta sheet

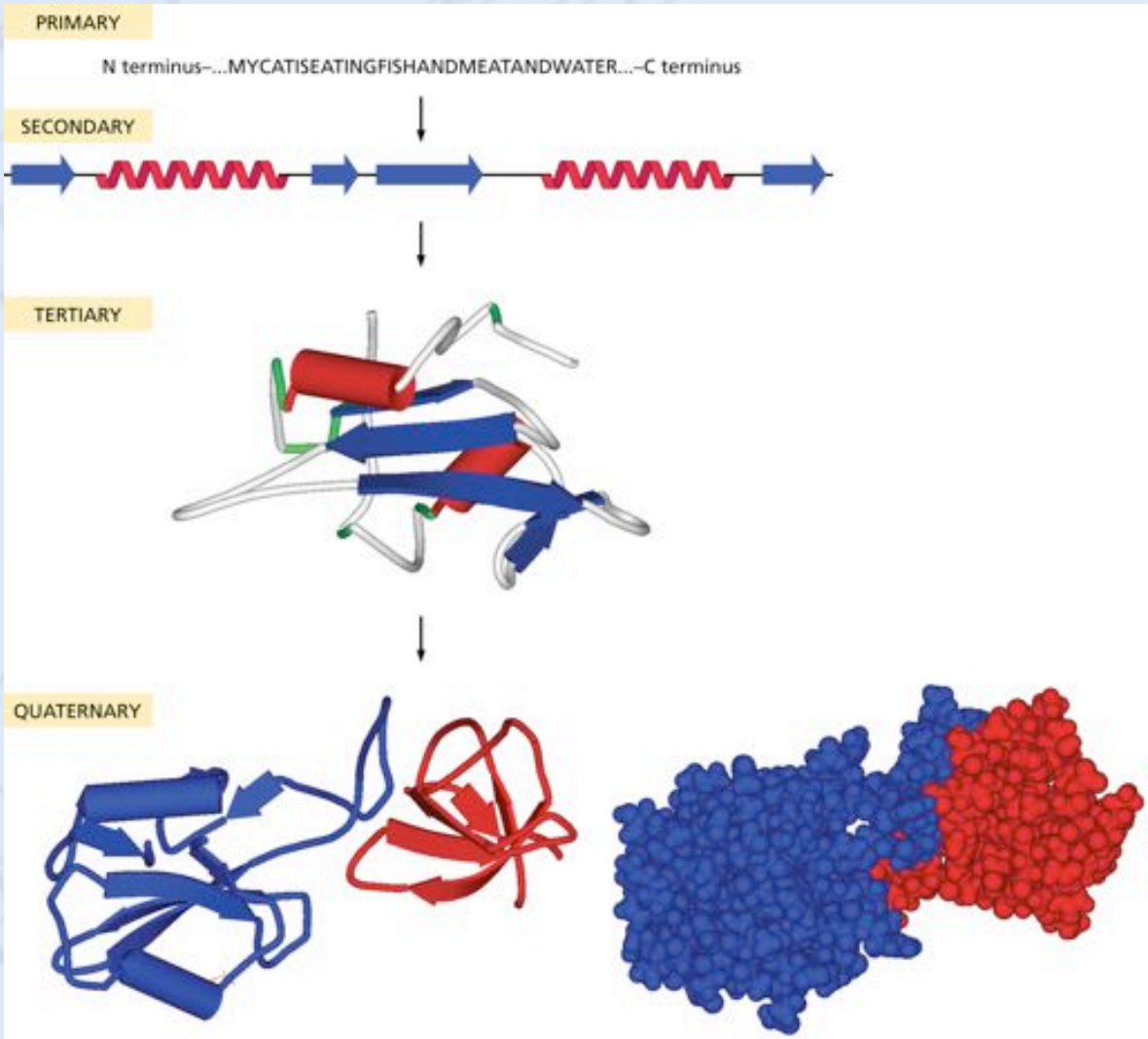


We will ignore this in the practical exercises!

Hydrophobic Collapse



Levels of protein structure



primary

amino acid sequence

secondary

helix, loop, beta strand

tertiary

3D orientation of secondary structure (topology)

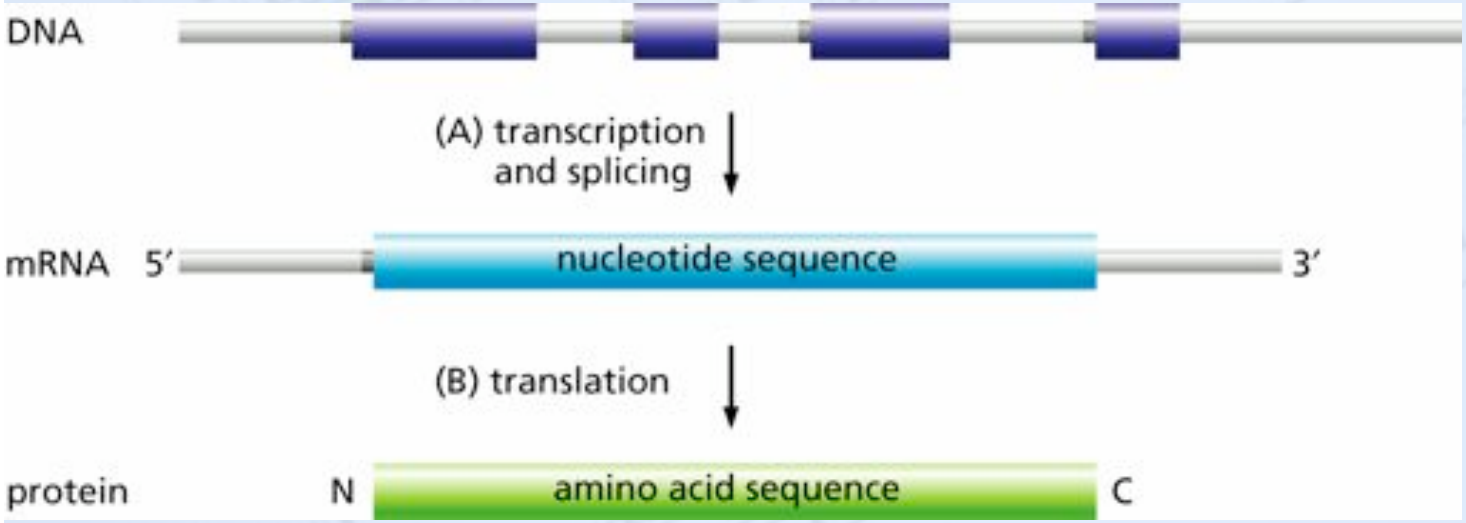
quaternary

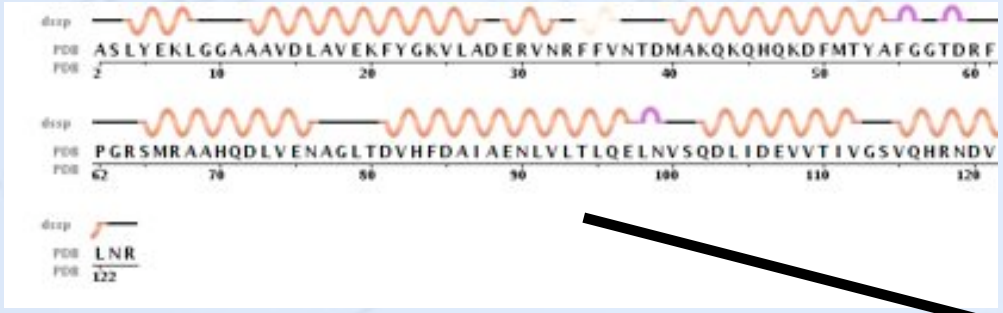
Interaction with other molecules, and peptide chains

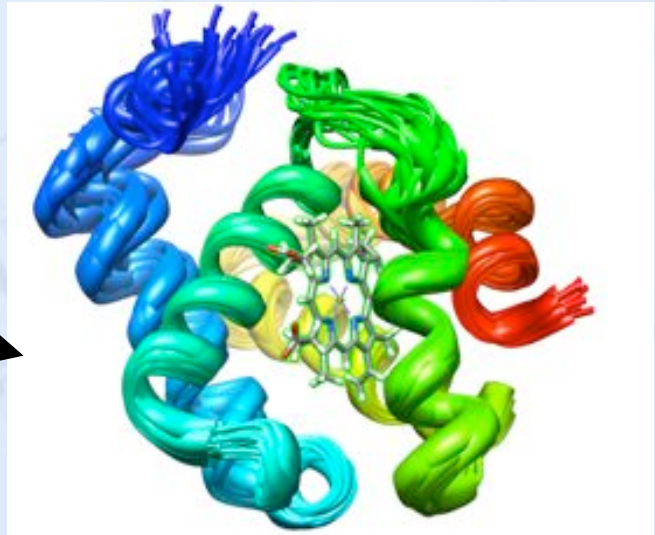
Physics & Biology: folding specificity

Unique characteristic:

The sequence of a protein determines its structure







Biology: structure is more conserved than sequence

1L9H:A(size=349) vs 1GUE:A(size=239)
Structure Alignment

Rmsd = 4.0Å Z-Score = 5.5
Sequence identity = 7.4%
Aligned/gap positions = 202/92

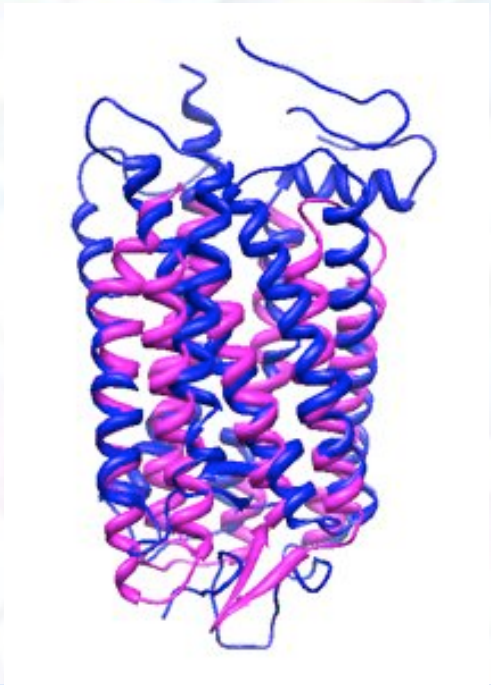
Sequence alignment based on structure alignment.

Sequence alignment based on structure alignment. Position numbers according to sequence (starting from 1) and according to PDB are given as SSSS/PPPP, SSSS - sequence, PPPP - PDB.

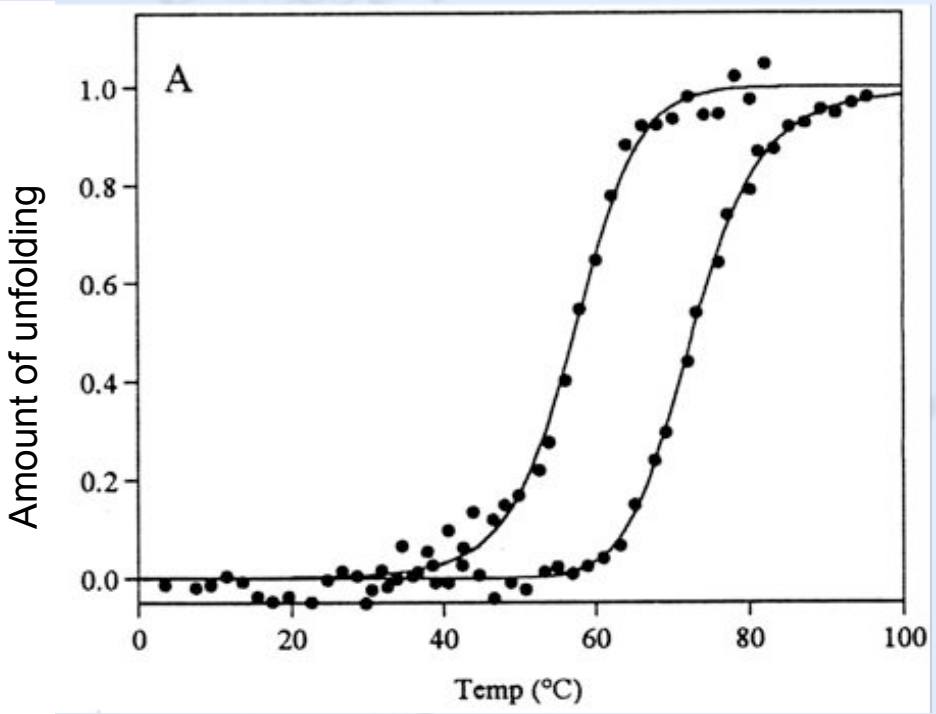
1L9H:A - MOL_ID: 1; MOLECULE: RHODOPSIN; CHAIN: A, B

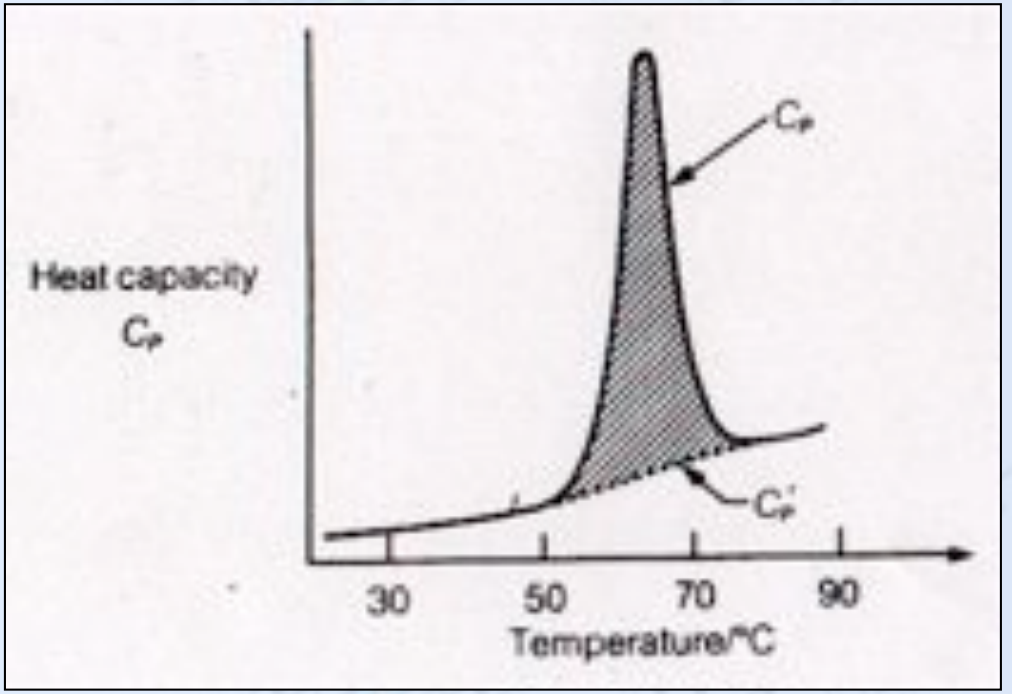
1GUE:A - MOL_ID: 1; MOLECULE: SENSORY RHODOPSIN II; SYNONYM: SR-II; CHAIN: A; ENGINEERED: YES; OTHER_DETAILS: K-STATE, REFINED WITH EXTRAPOLATED DATA

```
1L9H:A    33/33    AEPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGF
1GUE:A     3/4     GLTTLFWLGAIGMLVGTLAFAWAGRDA--GSG-----------ERRYYVTLVGISGIAA

1L9H:A    93/93    TTTLYTSLHGYFVFG--------PTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPM
1GUE:A    49/50    VAYVAMALG---VGWVPVAERTVFADPRYIDMILTTPLIVYFLGLLAG-----------

1L9H:A   145/145   SNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIPEGMQCSCGIDYYTPHEETNNESF
1GUE:A    93/94    -----------LDSREFGIVITLNTVVM--LAGFA---------GAMVPGIERYALFGM

1L9H:A   205/205   VIYMFVVHFIIPLIVIFFCYGQLVFTVKEAAAQQQESATTQKAEKEVTRMVIIMVIAFLI
1GUE:A   130/131   GAVAFLGLVYYLVGPMTESASQ----------------------RSSGIKSLYVRLRNL

1L9H:A   265/265   CWLPYAGVAFYIFTH---QGSDFG----PIFMTIPAFFAKTSAVYNPVIYIMMN
1GUE:A   167/168   TVI-LWAIYPFDWLLGPPGVALLTPTVDVALIVYLDLVTKVGFGFIALDAAATL
```



Physics: folding specificity - perfect self assembly





What kind of biological data is available?

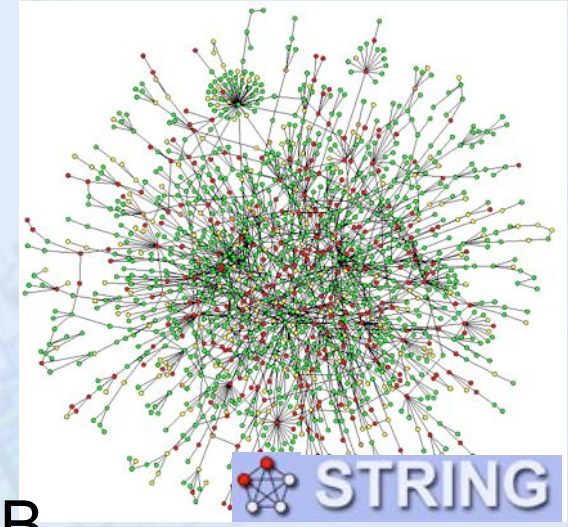
Structures

Sequences

Pfam

NCBI

Function

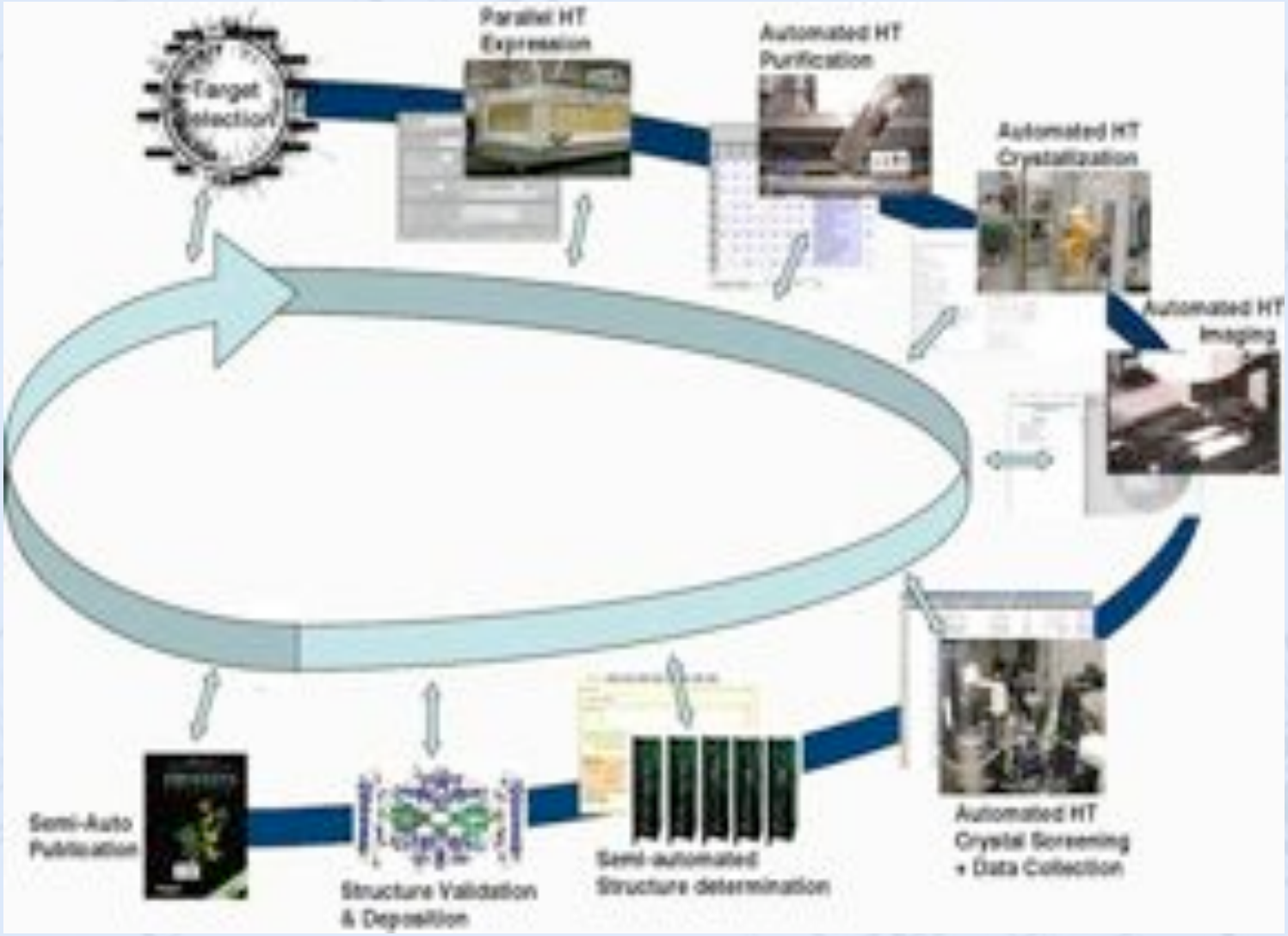


We will use protein structures from the PDB in the practical exercises

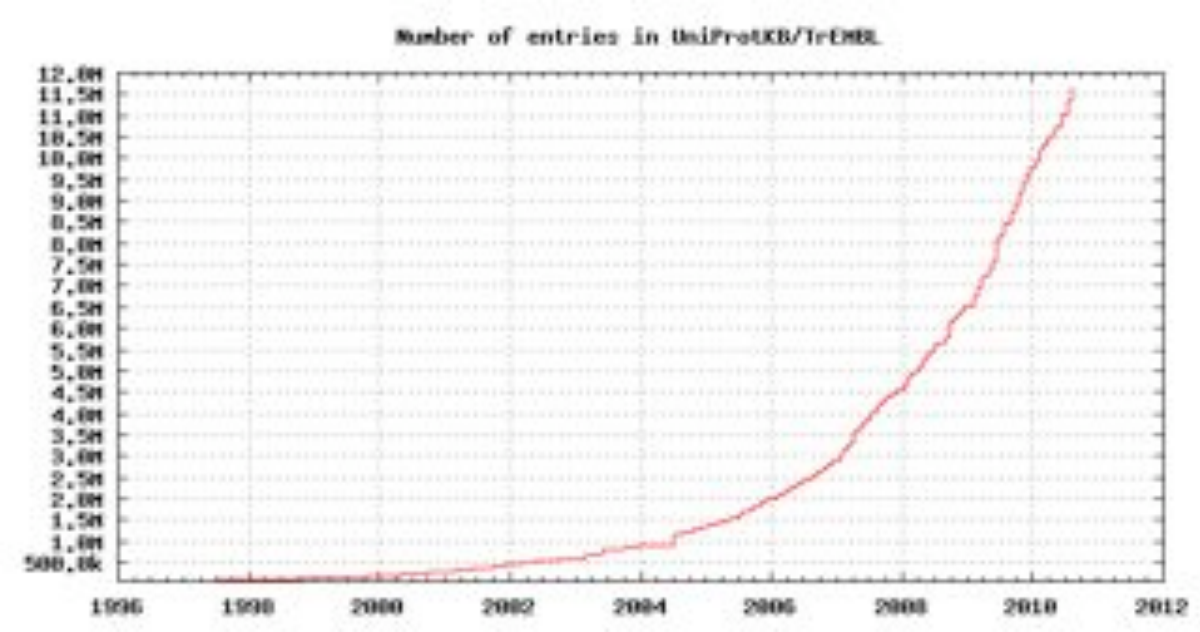
PDB

- Protein DataBank
  - X-ray structures
  - NMR structures
  - cryo-electron microscopy
- Biases in PDB
  - proteins that we can:
    - purify
    - crystalize
    - stabilize in solution
  - Sequence bias
- TM proteins

Structural Genomics



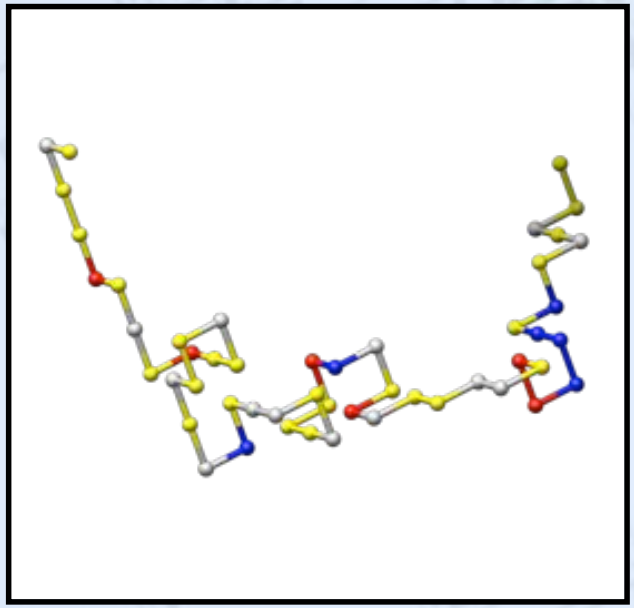
Sequence Structure 'Gap'



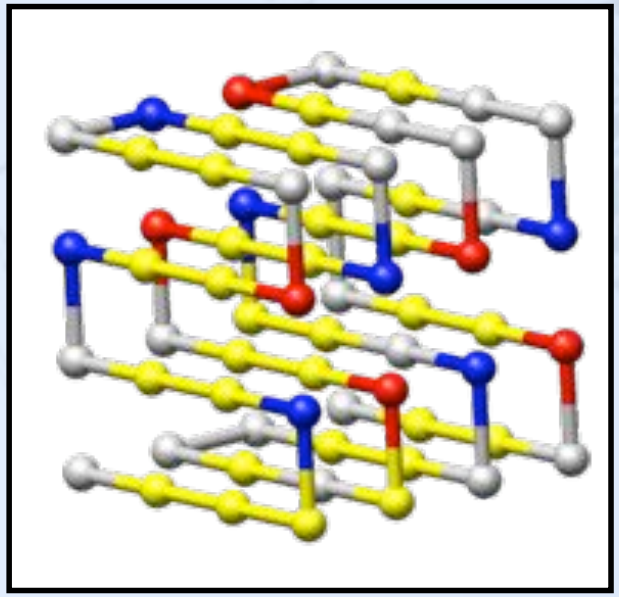
- Against 65.000 protein structures
- If only we could predict structure from sequence...

Lattice Model

Sequence



Structure



hydrophobic

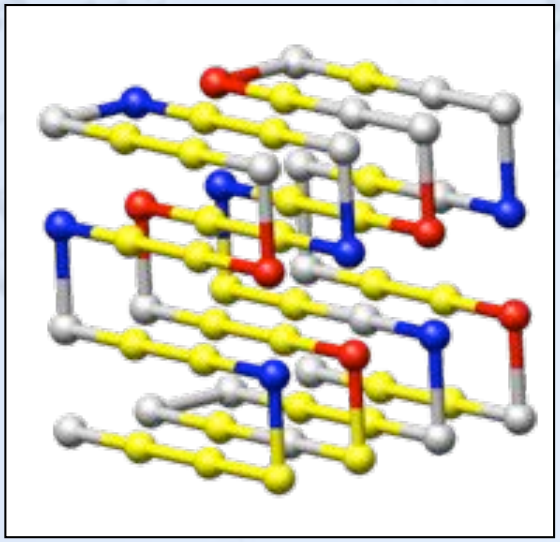
polar (hydrophilic)

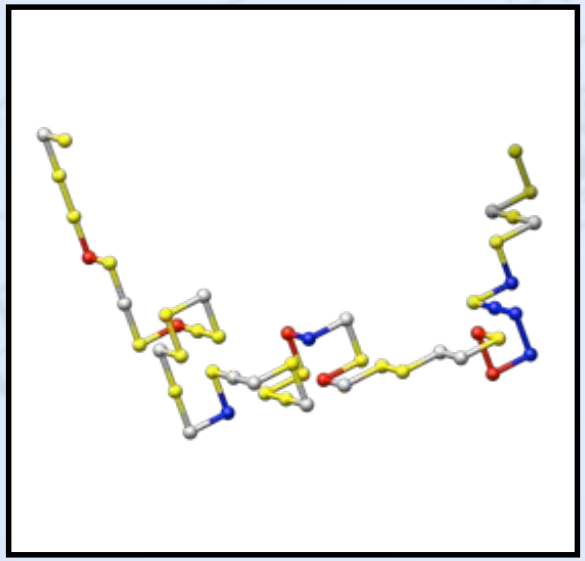
negative charge

positive charge

	--	+	+	+
	+			
	+		+	-
	+		-	+

Cubic Lattice Model





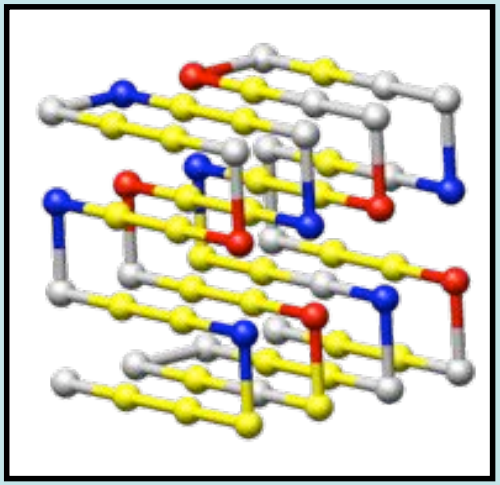
- Cheap & simple
  - Use for right purpose
- Can model:
  - General trends
  - Folding specificity
  - Heat capacity
  - Binding and unbinding
- Not captured:
  - Secondary structure
  - Hydrophobic effect (cold denaturation)
  - Specific proteins

Shakhnovich & Gutin 1993 PNAS 90

Coluzza et al 2003 Phys Rev E 68

Lattice Model, Potential, Design & Simulation

Sequence Design



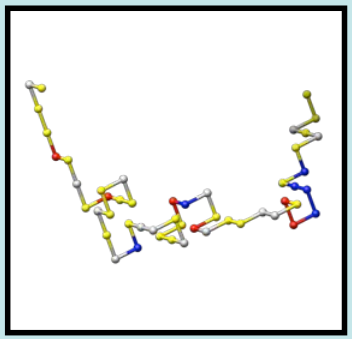
hydrophobic

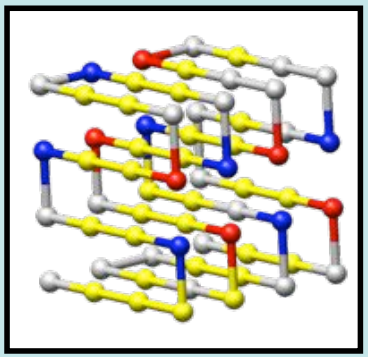
polar (hydrophilic)

negative charge

positive charge

Folding Simulation





Interaction Potential

	--	+	+	+
	+			
	+		+	-
	+		-	+

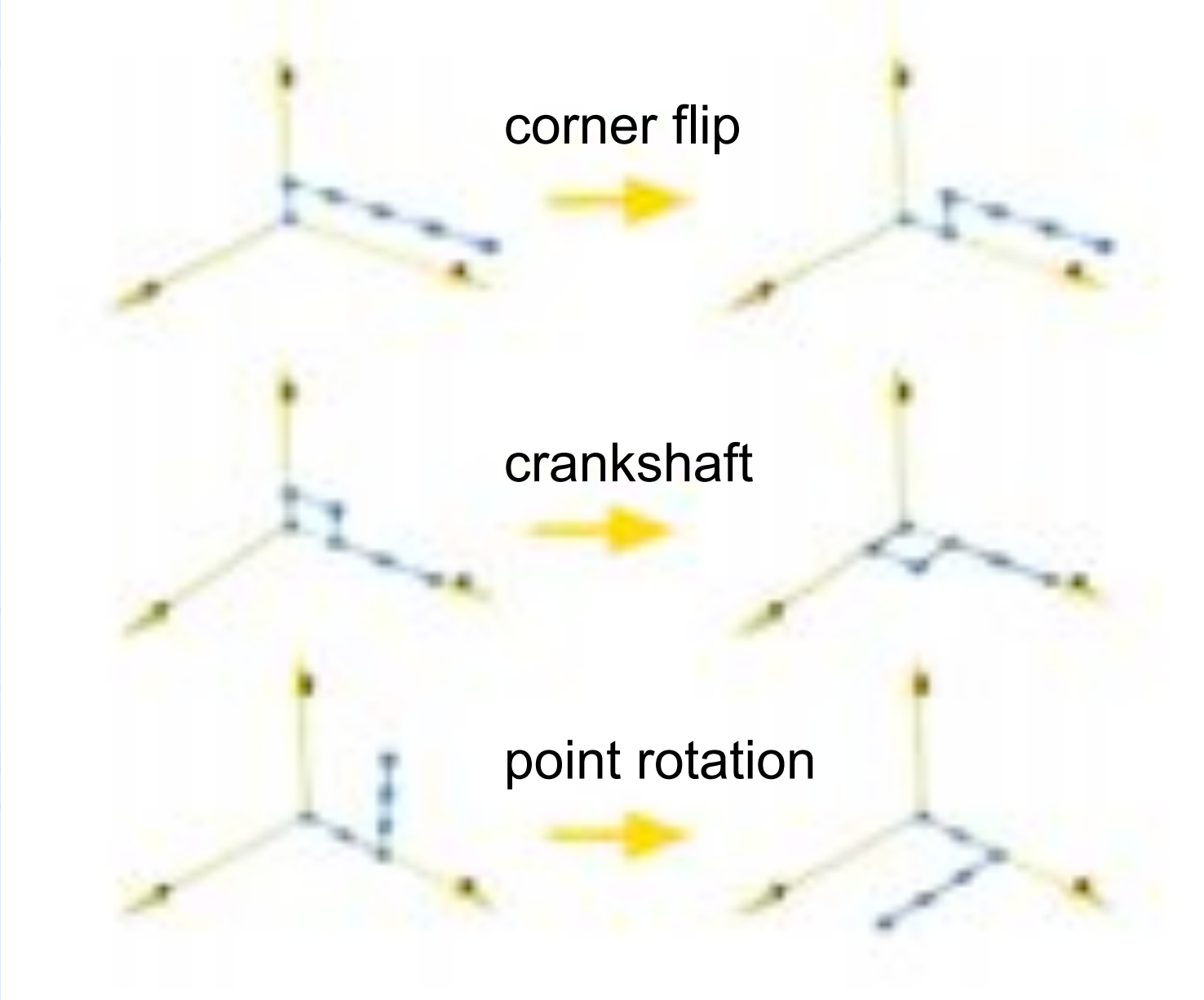
hydrophobic

polar (hydrophilic)

negative charge

positive charge

Simulation: Lattice Moves

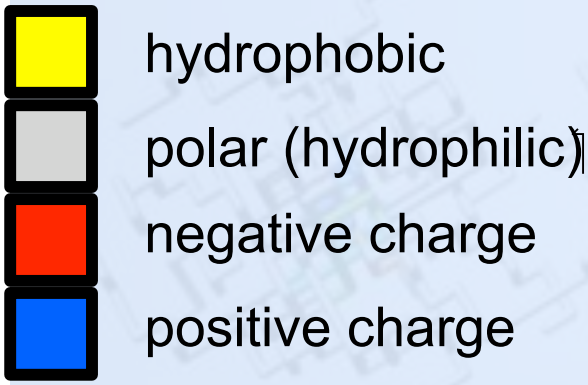


Simulation: interaction potential

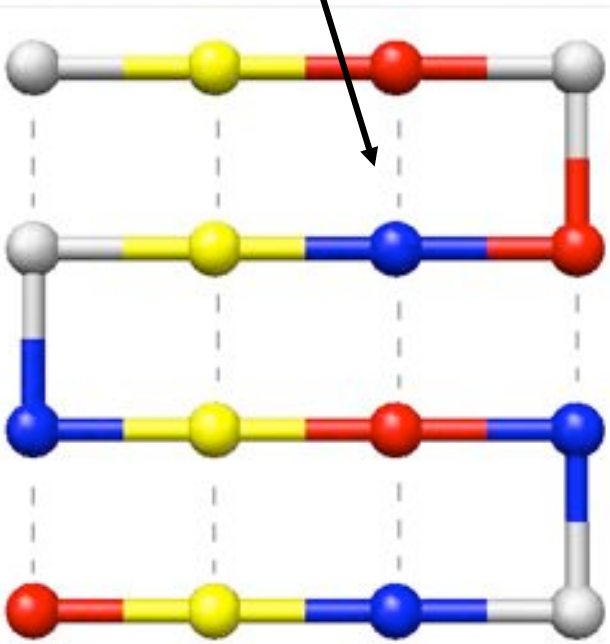
$$E = \frac{1}{2}\sum_{i}^{N}\sum_{j}^{N}\epsilon_{a(i),a(j)}C_{ij}$$

energy

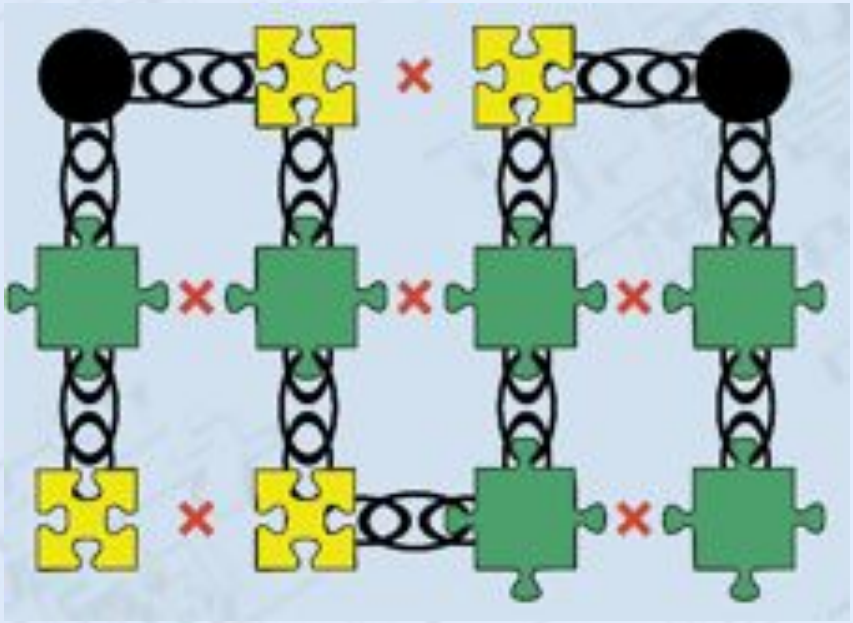
contact

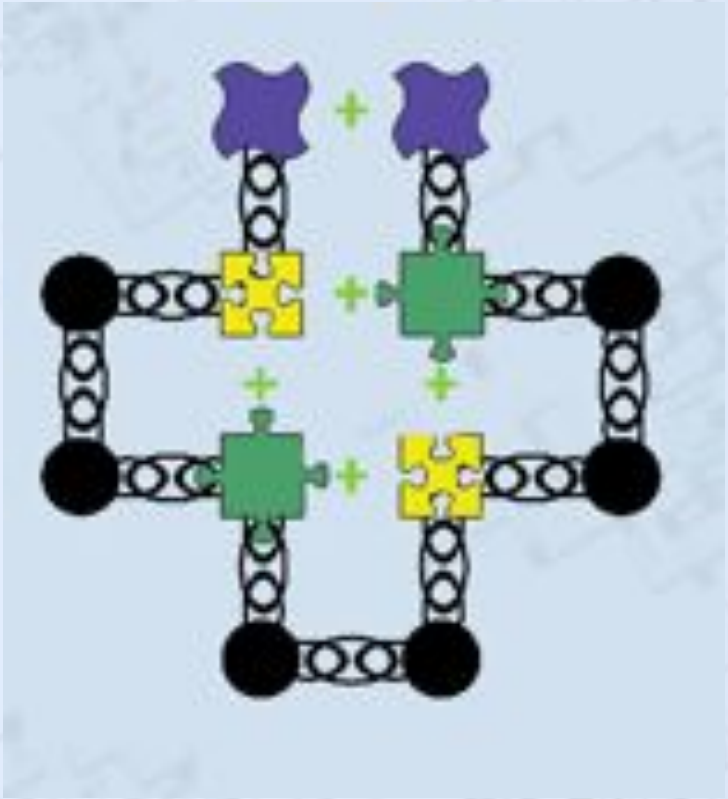


	--	+	+	+
	+			
	+		+	-
	+		-	+

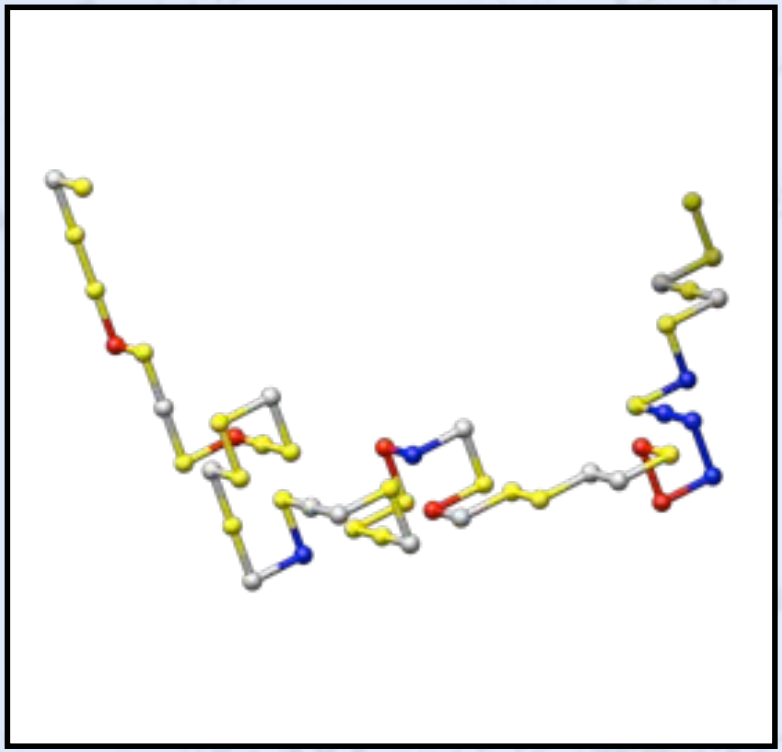


Interactions: example 2D





Simulation: Monte Carlo



Monte Carlo:

- Choose a residue (or region)
- Change its position
- Calculate new interaction energy
- Accept with Monte Carlo criterion

$$P_{\text{acc}} = \min\left\{1, \exp\left(\frac{E_{\text{old}} - E_{\text{new}}}{kT}\right)\right\}$$

Shakhnovich & Gutin 1993 PNAS 90

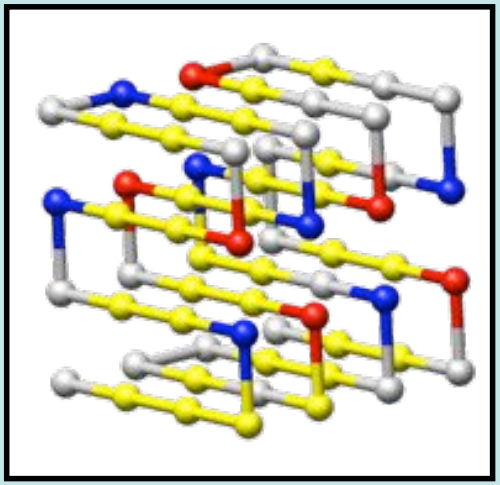
Coluzza et al 2003 Phys Rev E 68

Betancourt & Thirumalai 1999 Protein Sci 8

Simulation: 2D example

Sequence Design

Design



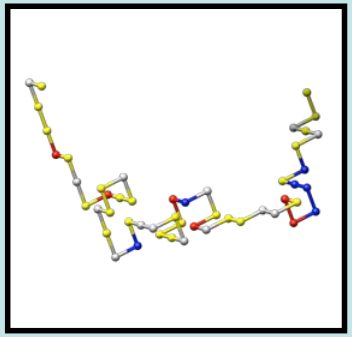
hydrophobic

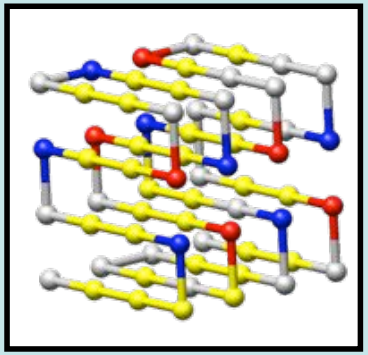
polar (hydrophilic)

negative charge

positive charge

Simulation





Potential

	--	+	+	+
	+			
	+		+	-
	+		-	+

hydrophobic

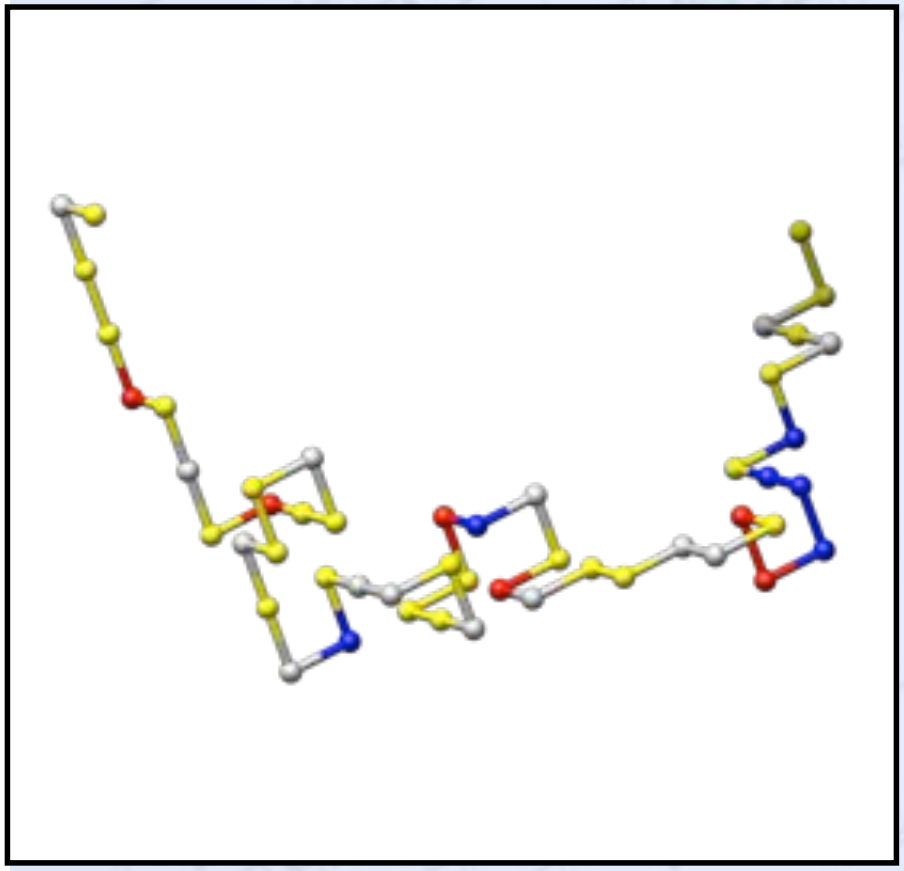
polar (hydrophilic)

negative charge

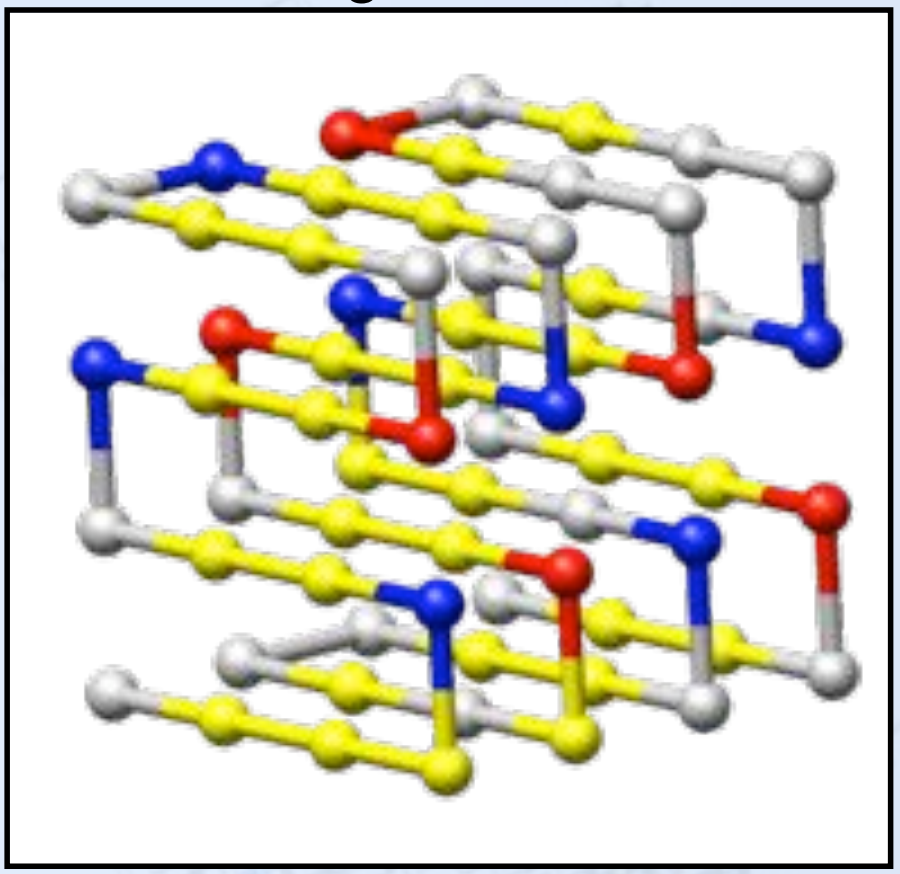
positive charge

Problem: how to create a folding sequence?

In nature evolution ensures folding...

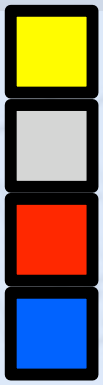


?



we can simulate evolution by changing sequence
with random substitutions

Lattice Model: design



hydrophobic

polar (hydrophilic)

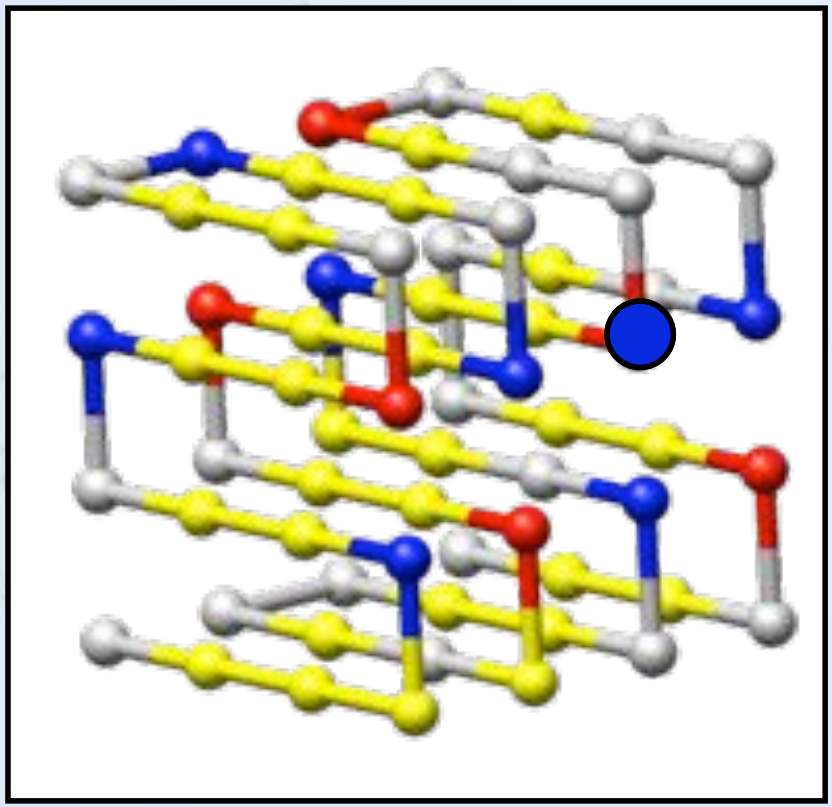
negative charge

positive charge

	--	+	+	+
	+			
	+		+	-
	+		-	+

Miyazawa & Jernigan 1993 Protein Eng 6

Betancourt & Thirumalai 1999 Protein Sci 8



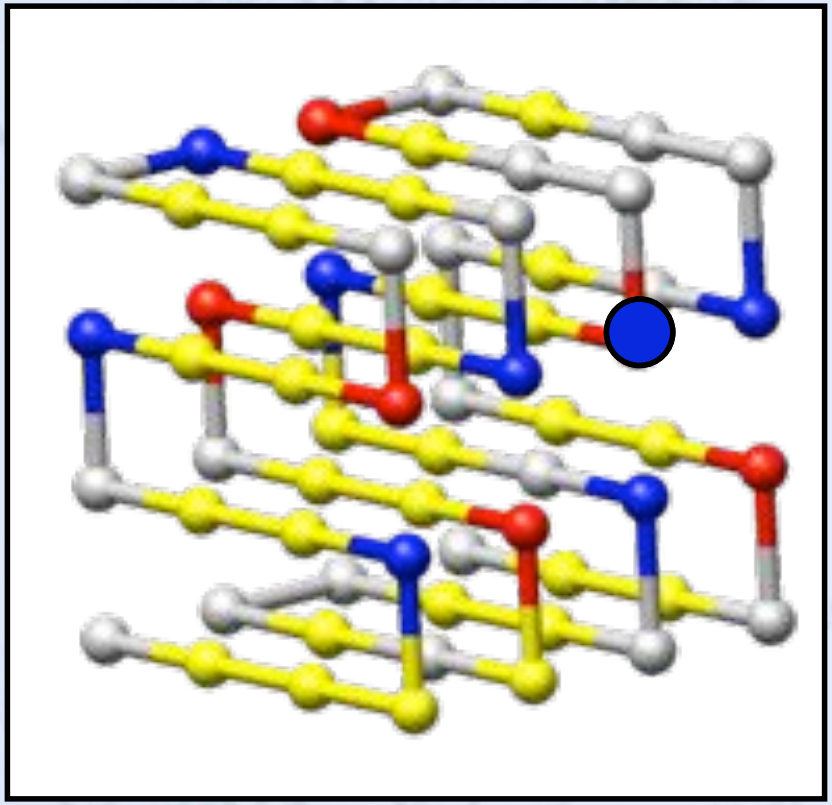
Design loop:

- Choose a residue
- Change the amino acid
- Calculate new interaction energy
- Accept with Monte Carlo criterion based on energy and variance

Shakhnovich & Gutin 1993 PNAS 90

Coluzza et al 2003 Phys Rev E 68

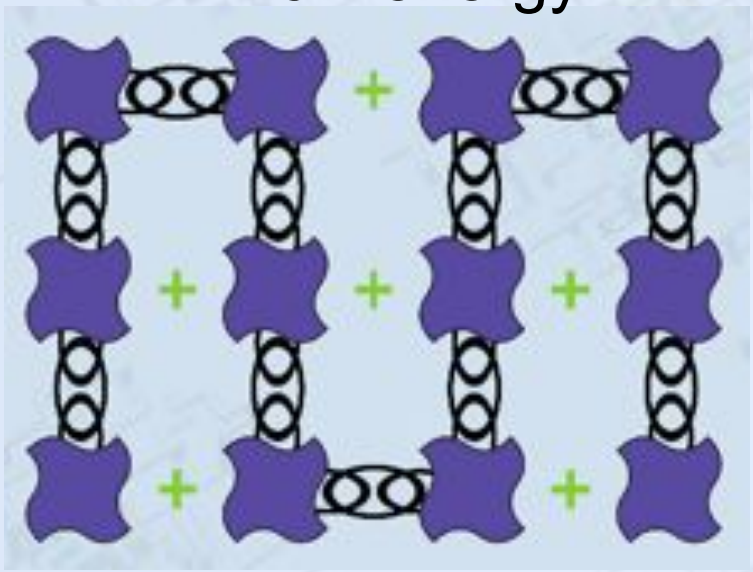
Sequence Design: Energy



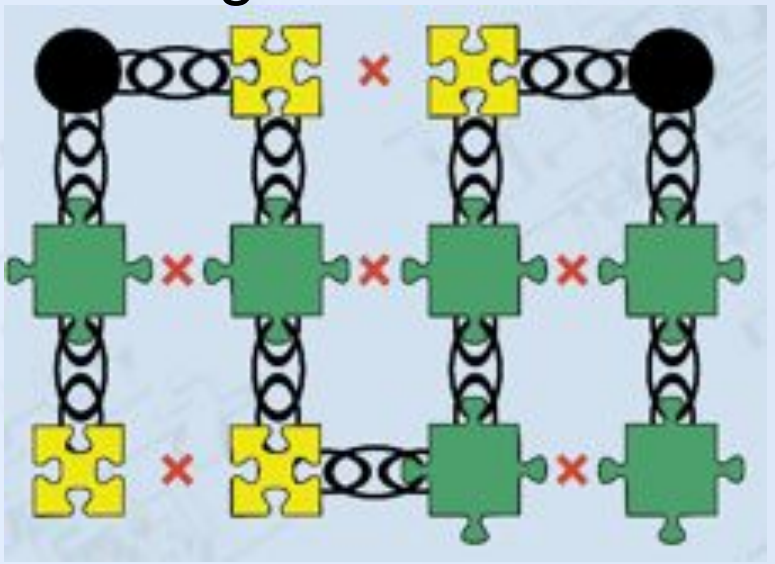
$$P_{\text{acc}} = \min\left\{1, \exp\left(\frac{E_{\text{old}} - E_{\text{new}}}{kT}\right)\right\}$$

Sequence Design

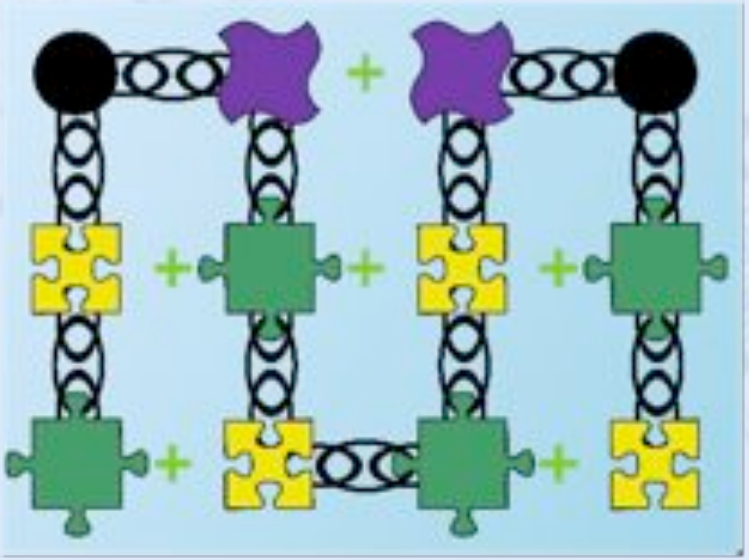
Low energy



High variance



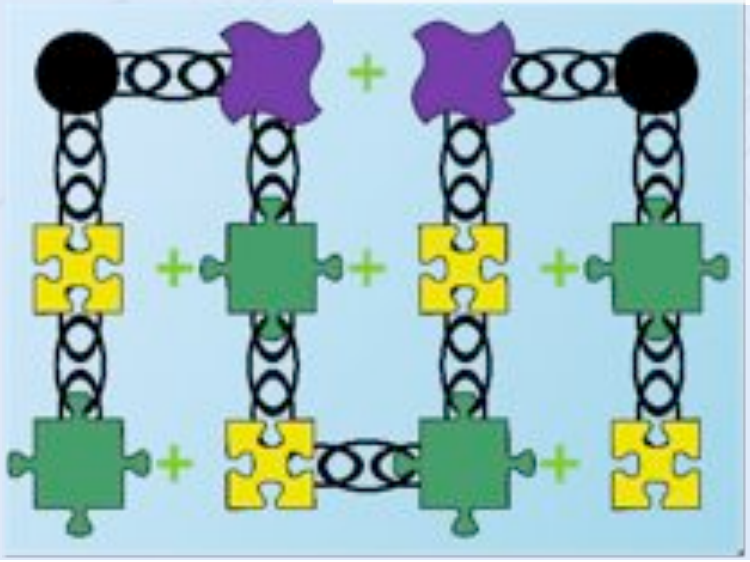
Good folder

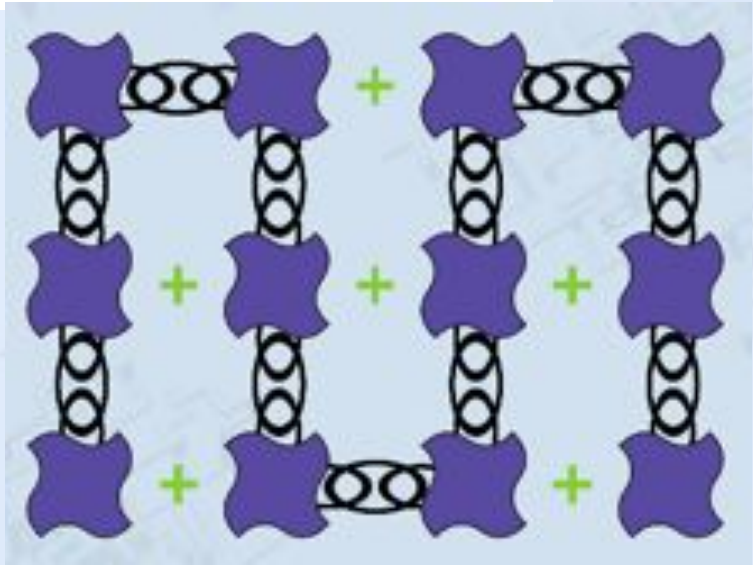


Sequence Variance

$$N_p = \frac{N!}{n_1!n_2!...n_{N_A}!}$$

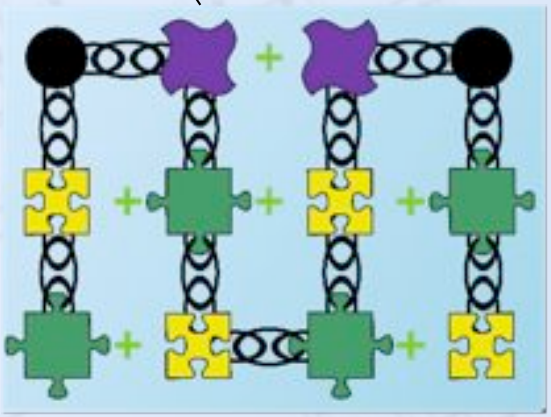
$$P_{\text{acc}} = \min\left\{1, \left(\frac{N_p^{\text{new}}}{N_p^{\text{old}}}\right)^{1/q}\right\}$$





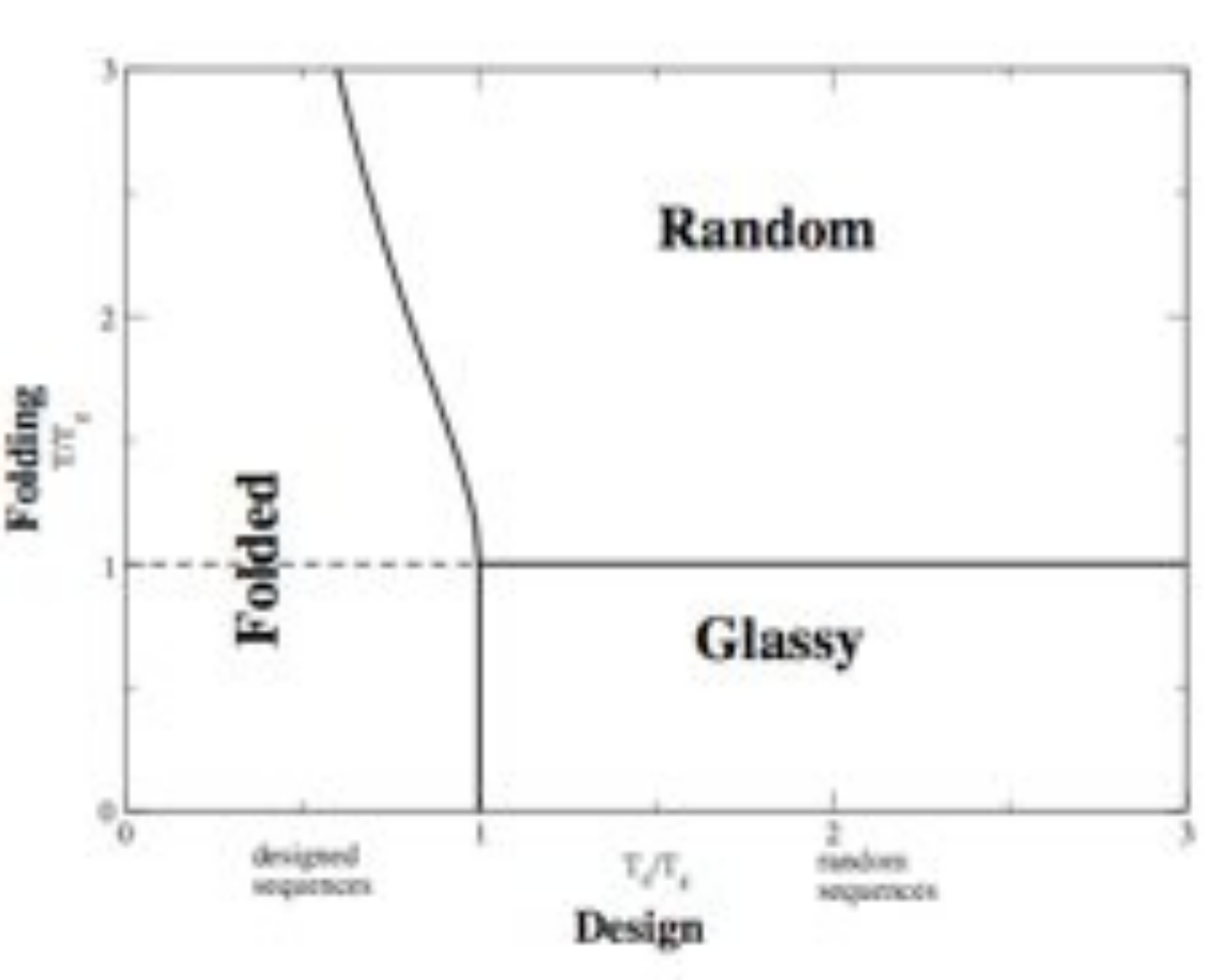
Sequence Variance & Biology

$$d = \sum_{a=0}^{a=20} (p_a - s_a)^2$$

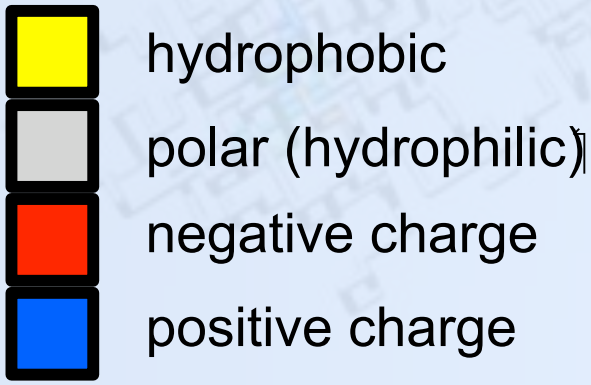


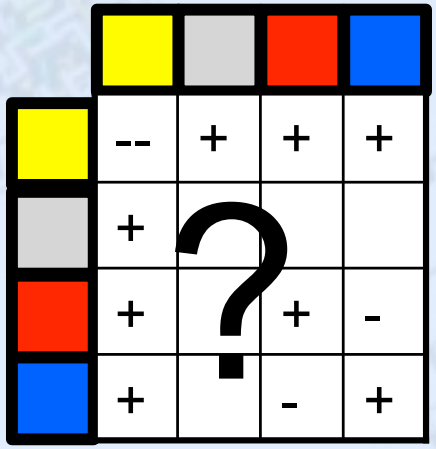
$$P_{\text{acc}} = \min\left\{1, e^{-1/q(d_{\text{new}} - d_{\text{old}})}\right\}$$

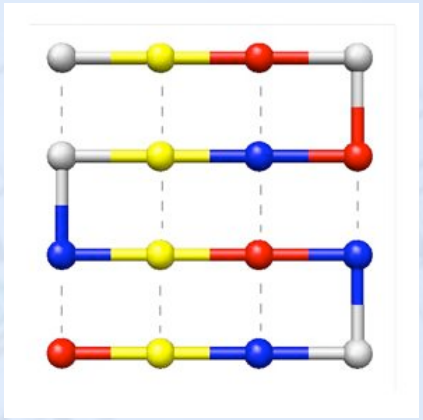
Design Temperature



How to derive the parameters?



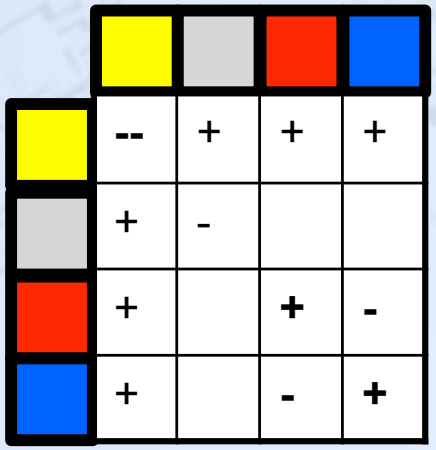


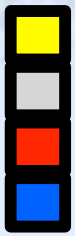


$$E = \frac{1}{2}\sum_{i}^{N}\sum_{j}^{N}\epsilon_{a(i),a(j)}C_{ij}$$

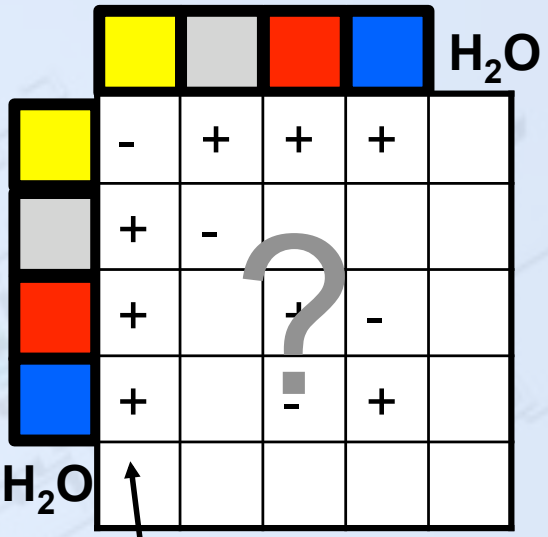
Can we use experimental biological data?

“Knowledge Based” Amino Acid Pair Potentials





hydrophobic
polar (hydrophilic)
negative charge
positive charge

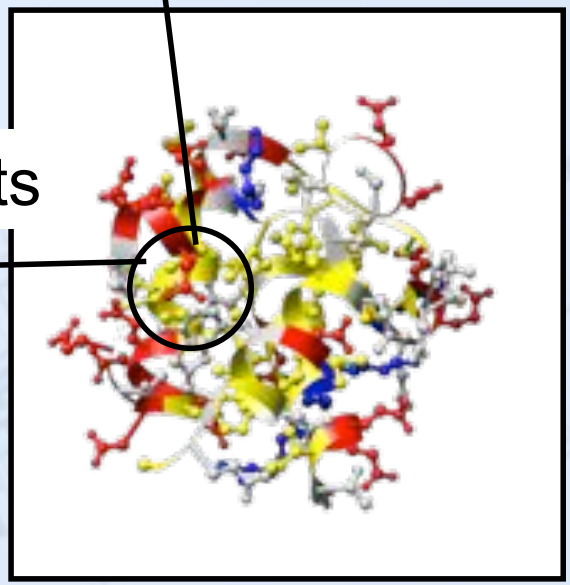


Miyazawa & Jernigan 1993 Protein Eng 6
Betancourt & Thirumalai 1999 Protein Sci 8

observed contacts

$$\epsilon_{i,j} = -kT \ln \left(\frac{c_{i,j}}{\omega_{i,j}} \right)$$

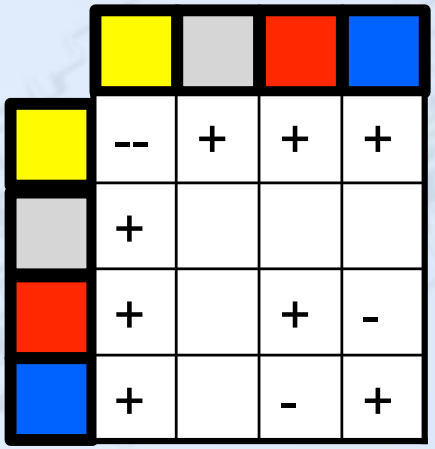
expected contacts

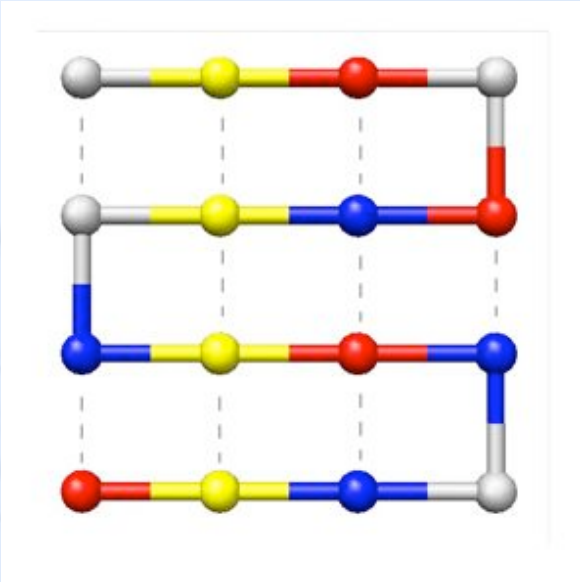


“Knowledge Based” Amino Acid Pair Potentials

$$\epsilon_{i,j} = kT \ln \left(\frac{c_{i,j}}{\omega_{i,j}} \right)$$

$$\omega_{i,j} = \frac{n_i q_i n_j q_j}{\sum_k q_k n_k}$$



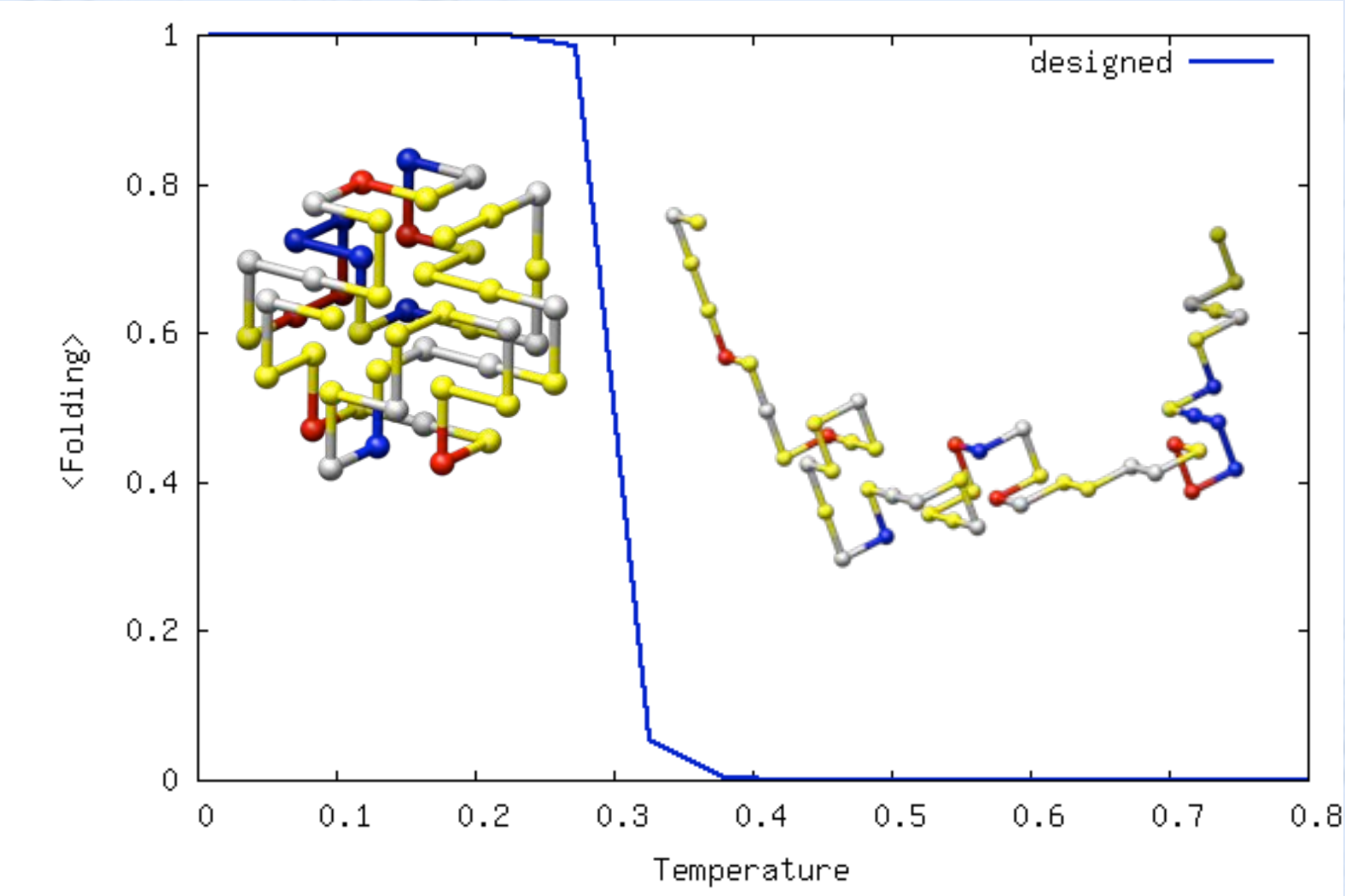


“Knowledge Based” Amino Acid Pair Potentials

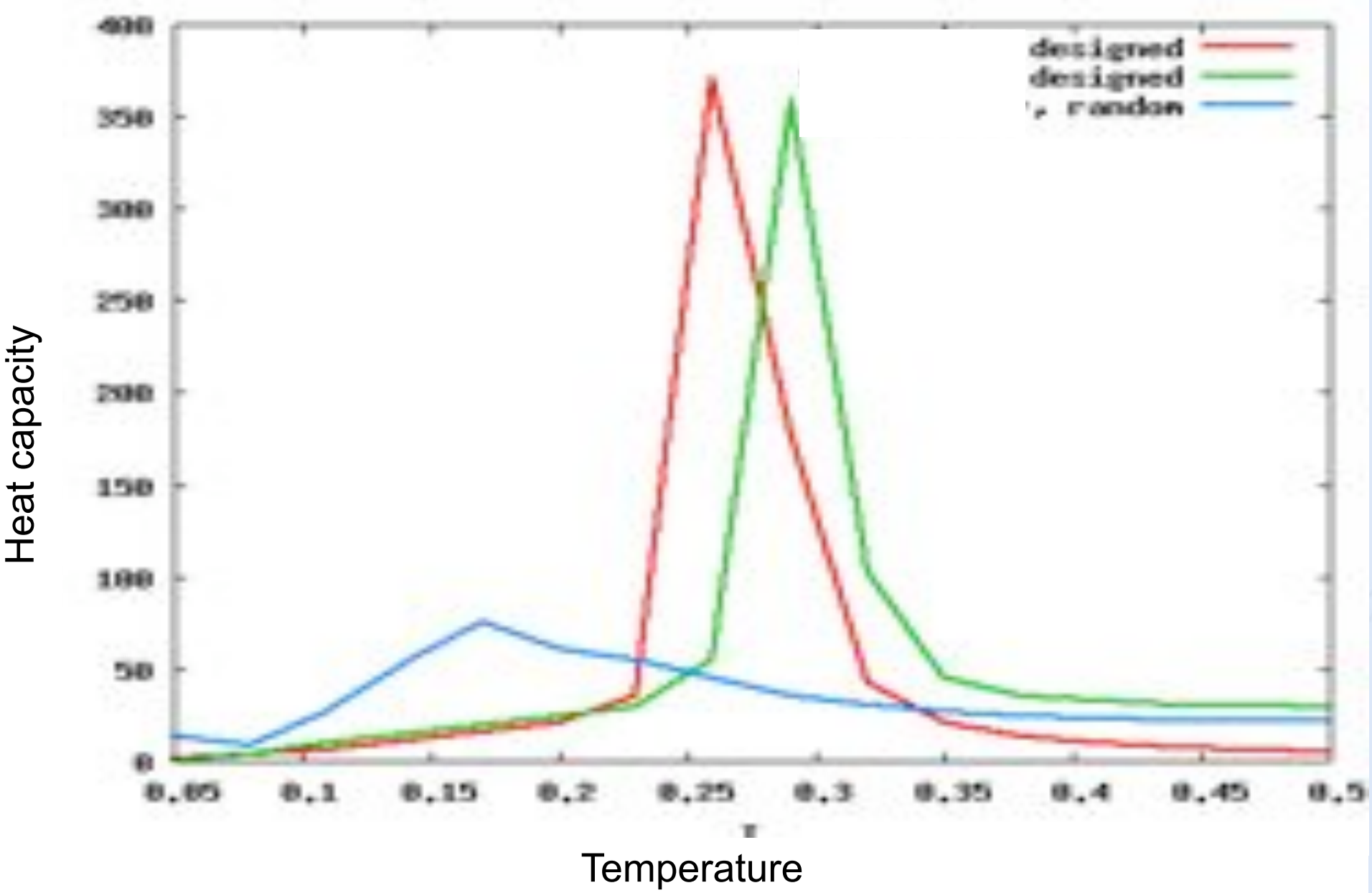
- Sample contacts in the PDB
- Assumption:
  PDB is a representative ensemble of well mixed amino acids

- How could biology (evolution) affect these results?
- How could we prevent this?

Folding Specificity on the Lattice

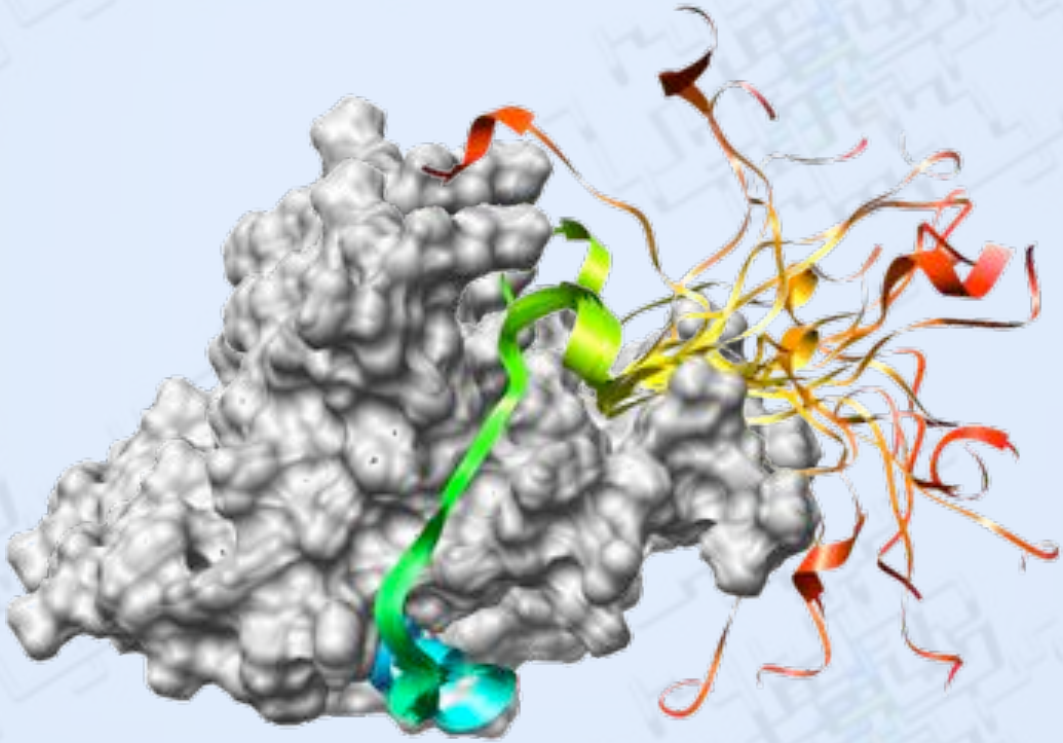


Foldable, with high specificity



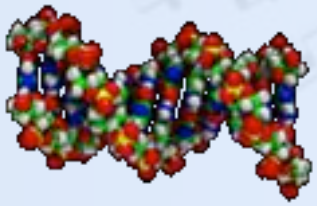
Case study:
Disordered flanks prevent peptide aggregation

Why binding regions are embedded in disordered flanks



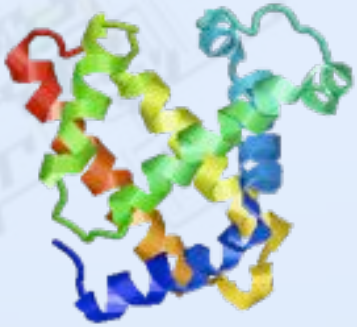
Protein function depends on folded protein structure

DNA →

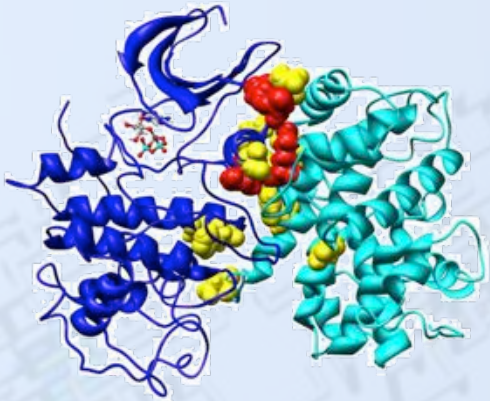


RNA codon

amino acid sequence

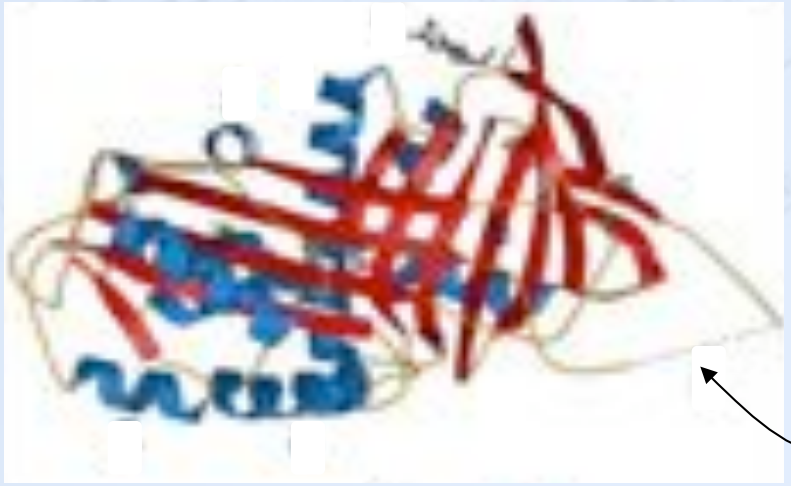


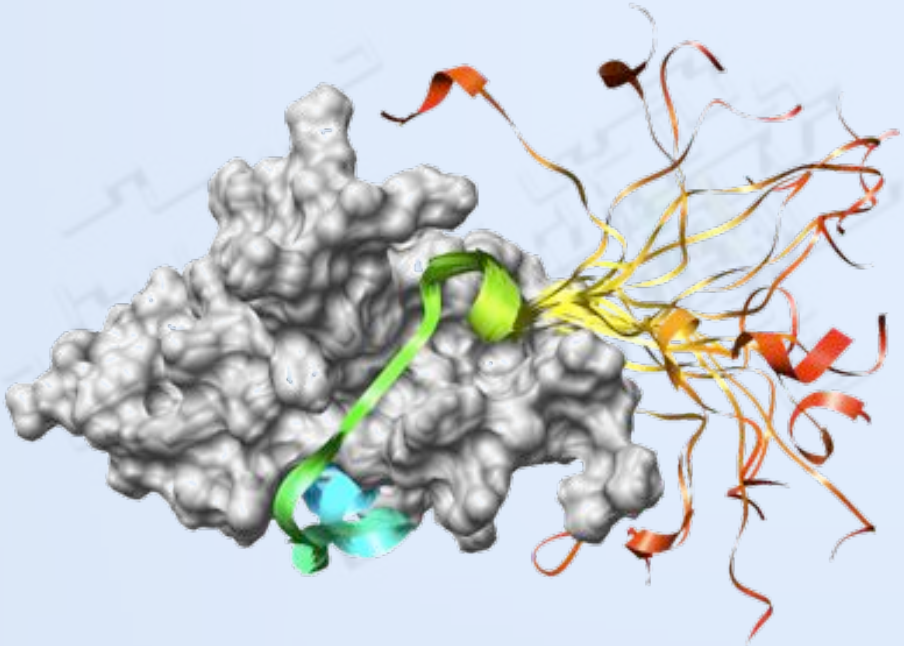
folded protein →



function

Disordered regions are common





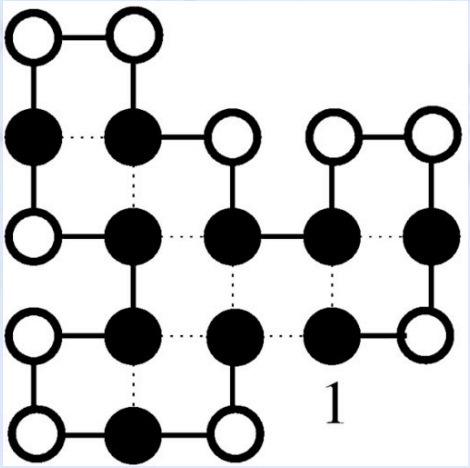
Disordered regions:

- Missing in X-ray structures
  - Typically removed for crystallization
- 33% of eukaryotic proteins contain large disordered segment *Ward et al. J Mol Biol (2004)*
- Associated functions
  - Signalling
  - Regulatory
- Disordered flanks found next to binding motifs
  - Hydrophobic binding motif
  - Hydrophilic flanks

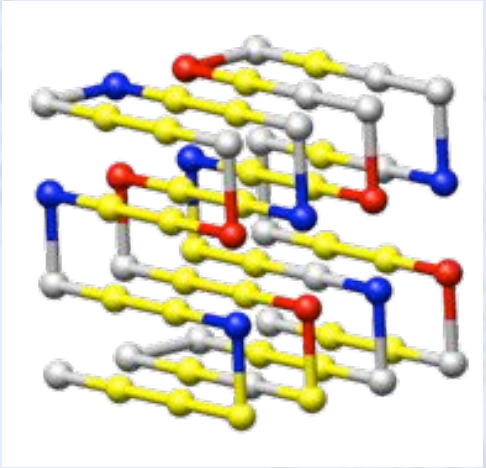
Fuxreiter et al. Bioinformatics (2007)

Why use simple models?

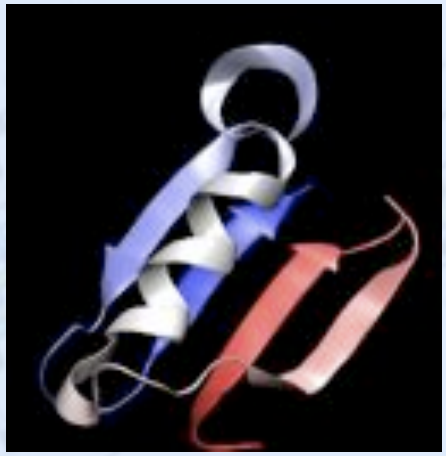
HP model - minute



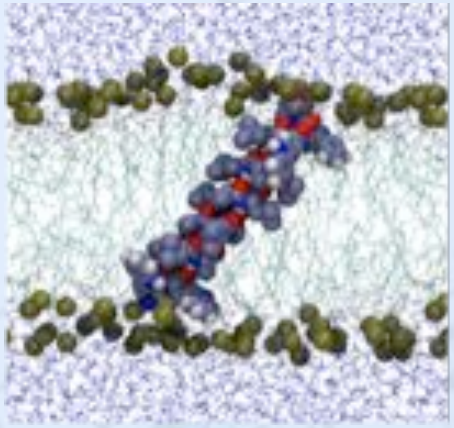
cubic lattice model - hour



backbone model - week

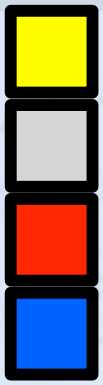


full atomistic model - year



- Sampling lowest free energy state
- Different conditions
- Larger systems

Lattice Model: Monte Carlo Simulation

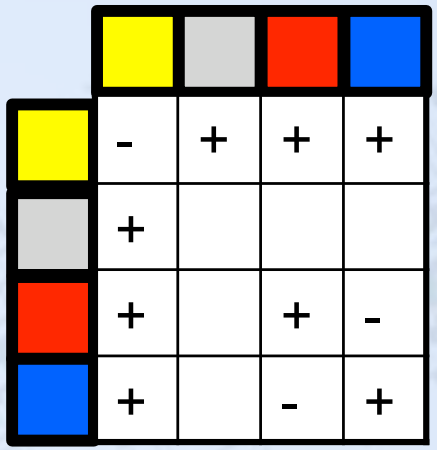


hydrophobic

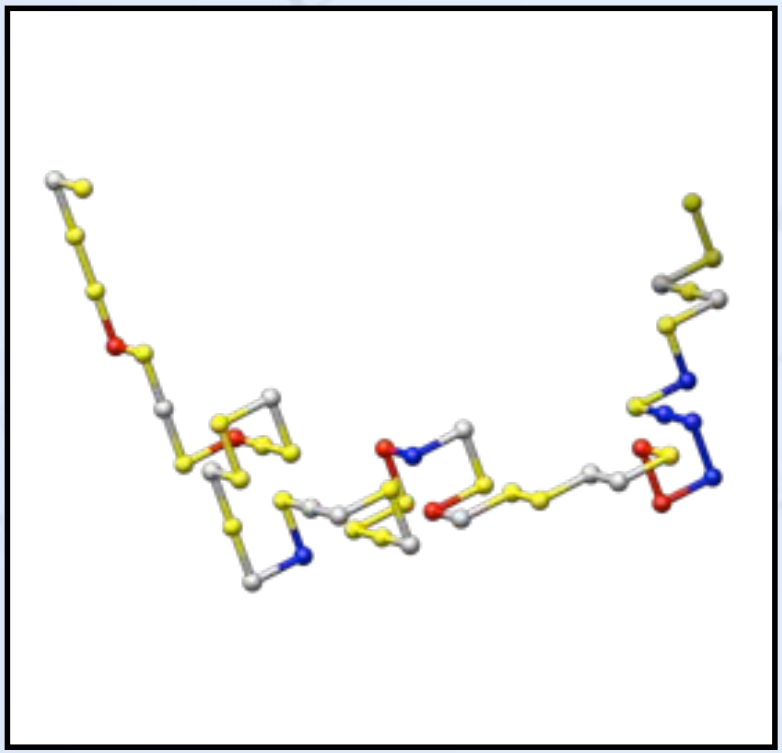
polar (hydrophilic)

negative charge

positive charge



Betancourt & Thirumalai 1999 Protein Sci 8



Monte Carlo:

- Choose a residue (or region)
- Change its position
- Calculate new interaction energy
- Accept with Monte Carlo criterion based on energy

Shakhnovich & Gutin 1993 PNAS 90

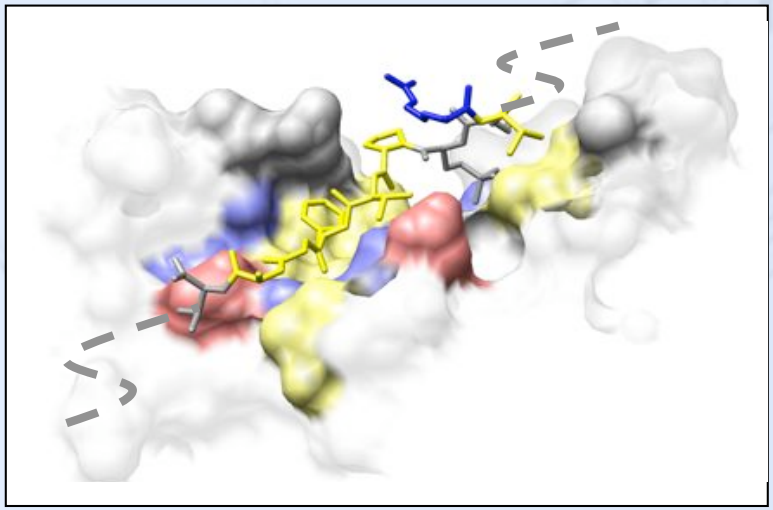
Coluzza et al 2003 Phys Rev E 68

Modelling the binding region & disordered flanks

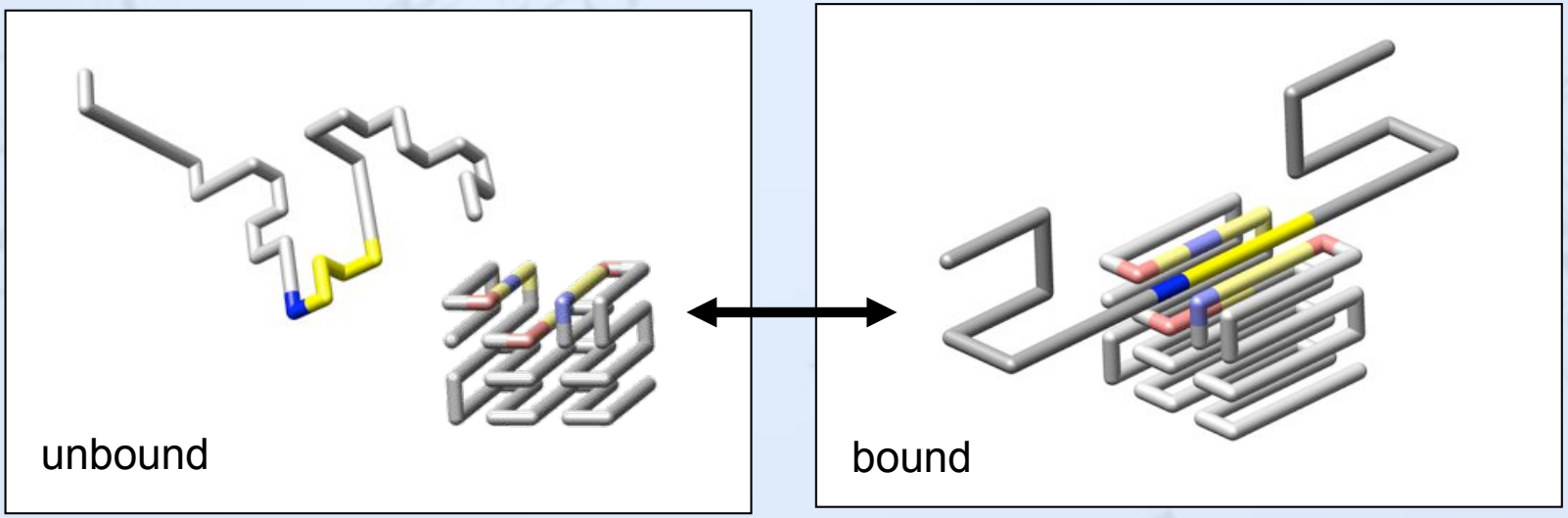
General trends:

- Hydrophobic binding motif
- Disordered hydrophilic flanks
- Designed binding pocket (specific binding)
- Difficult to observe unbound state in experiment

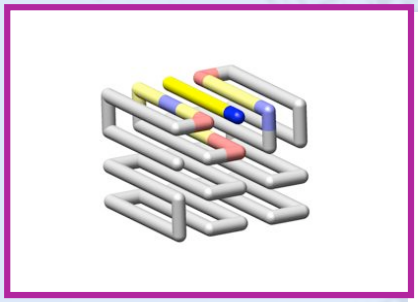
X-ray structure

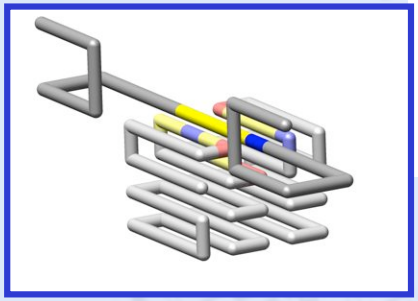


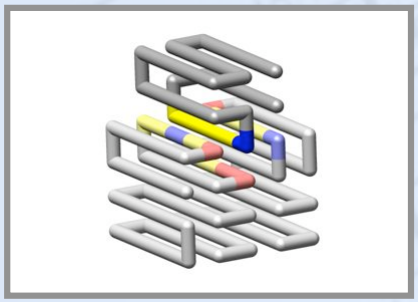
Lattice model

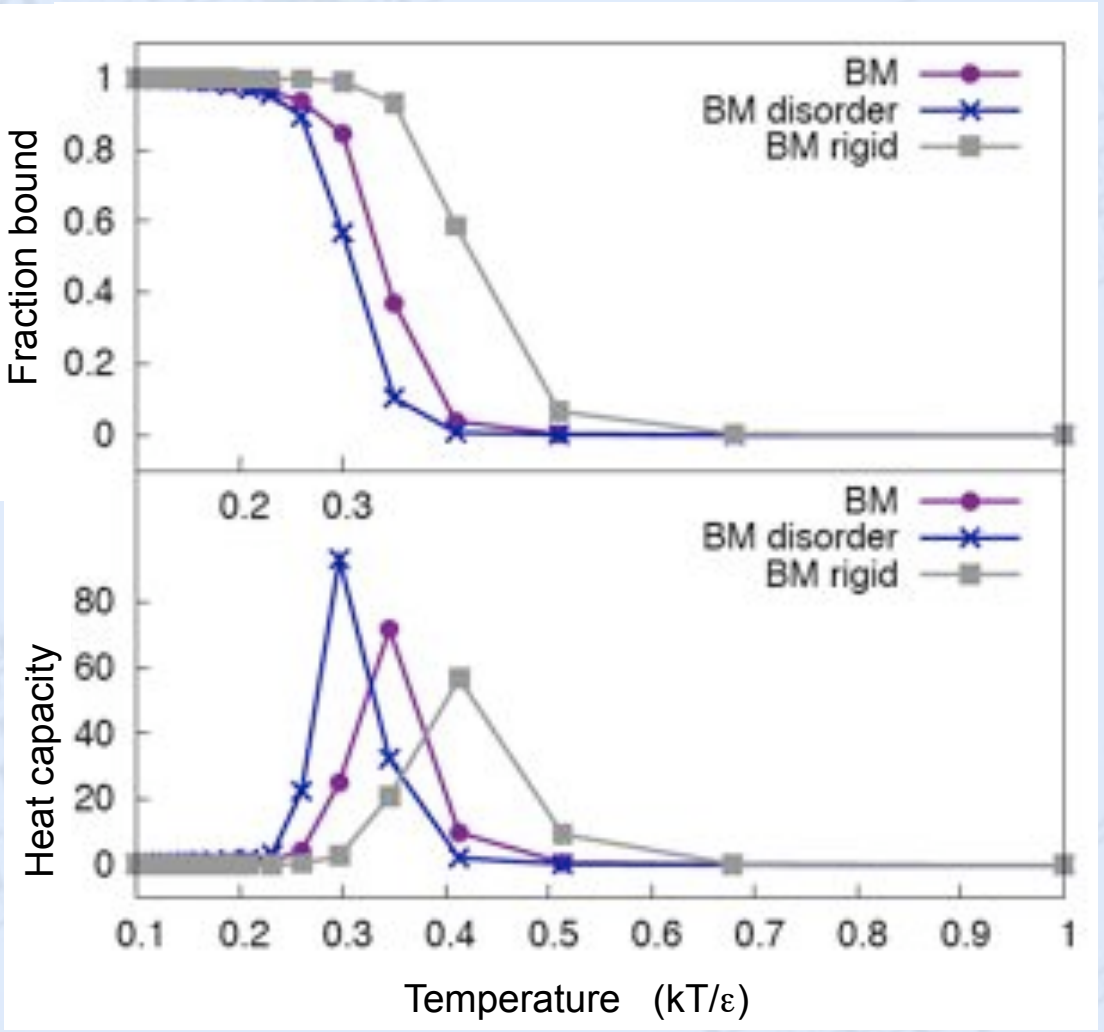


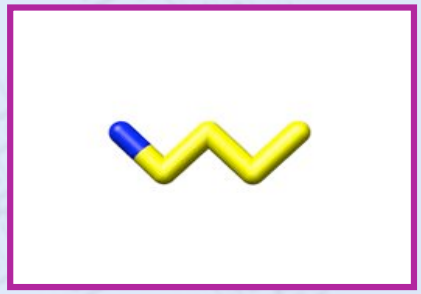
Lower binding strength flexible binding region

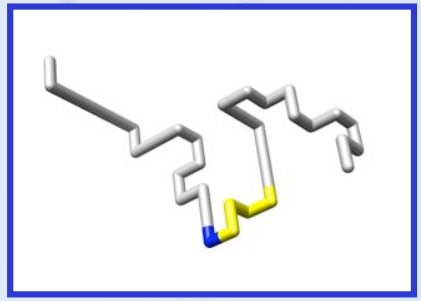


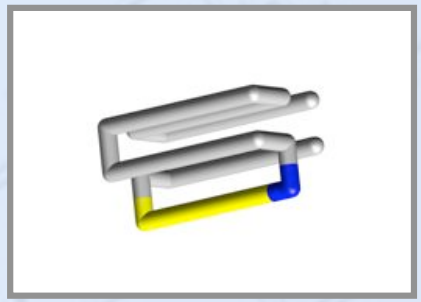








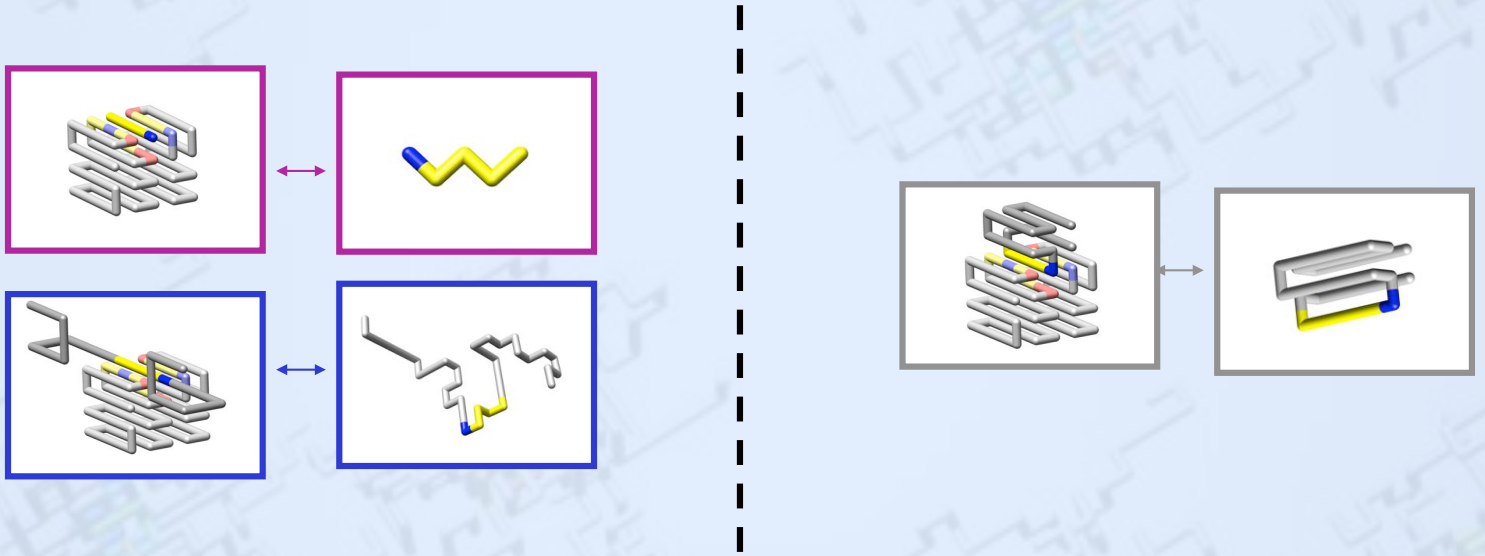




Low binding strength, with high specificity due to coupled binding and folding: *Coluzza & Frenkel Biophys. J. (2007)*

Flanks have no effect on binding

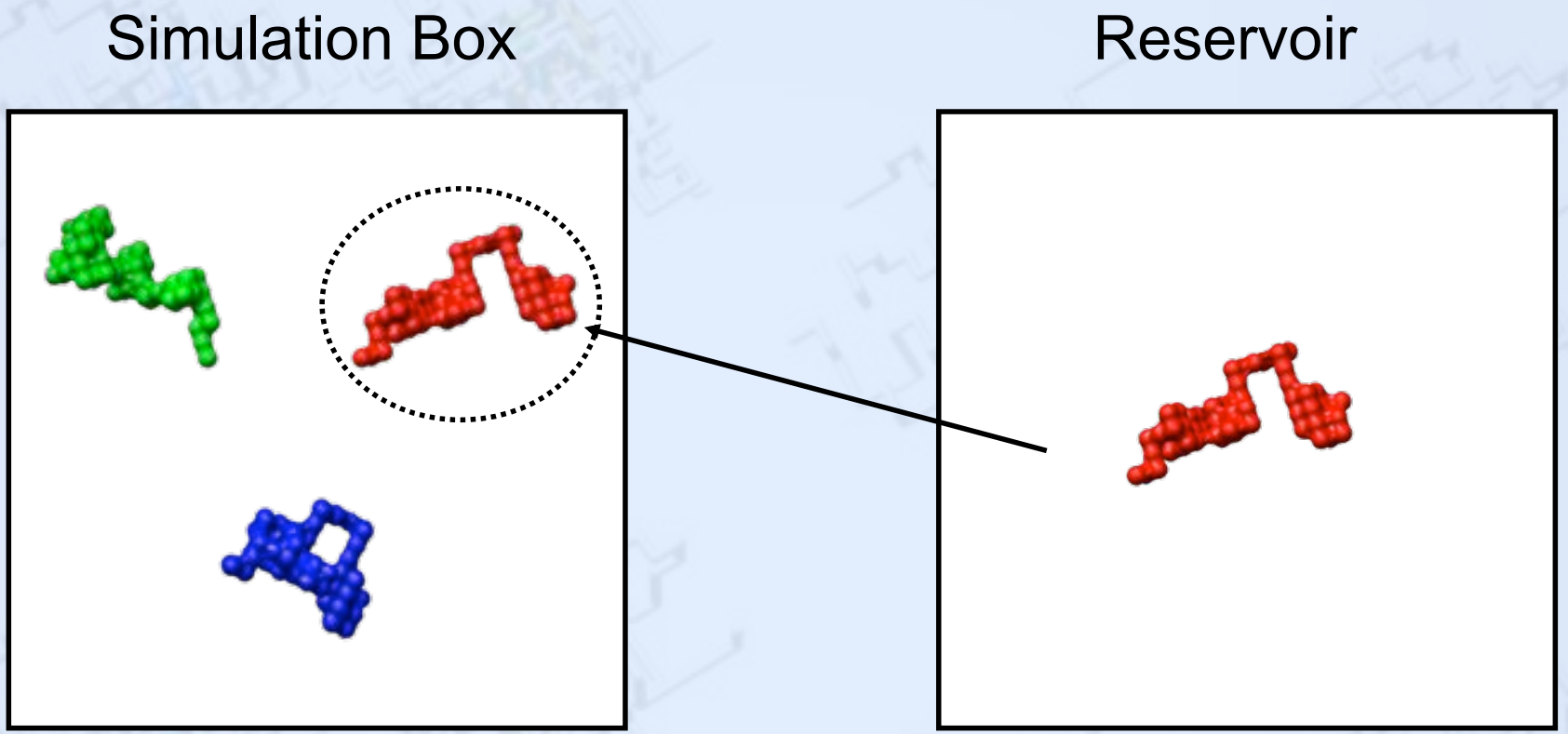
- Binding of a flexible region loses entropy upon binding
  - Reversible binding is important for signaling functions



- Disordered flanks have little effect on the binding strength

Low binding strength, with high specificity due to coupled binding and folding: *Coluzza & Frenkel Biophys. J. (2007)*

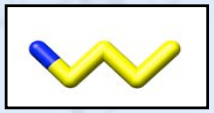
Simulating multiple proteins

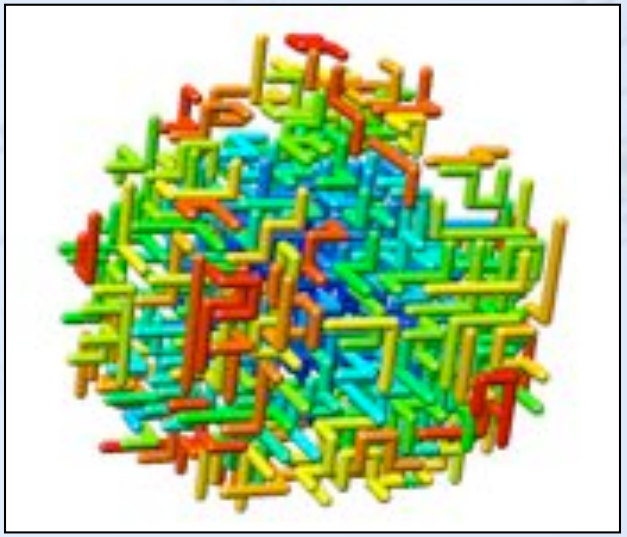


- Grand Canonical Simulation
  - at constant low concentration

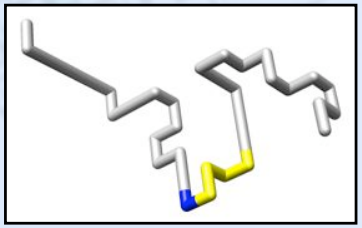
Collective effect of disordered flanks

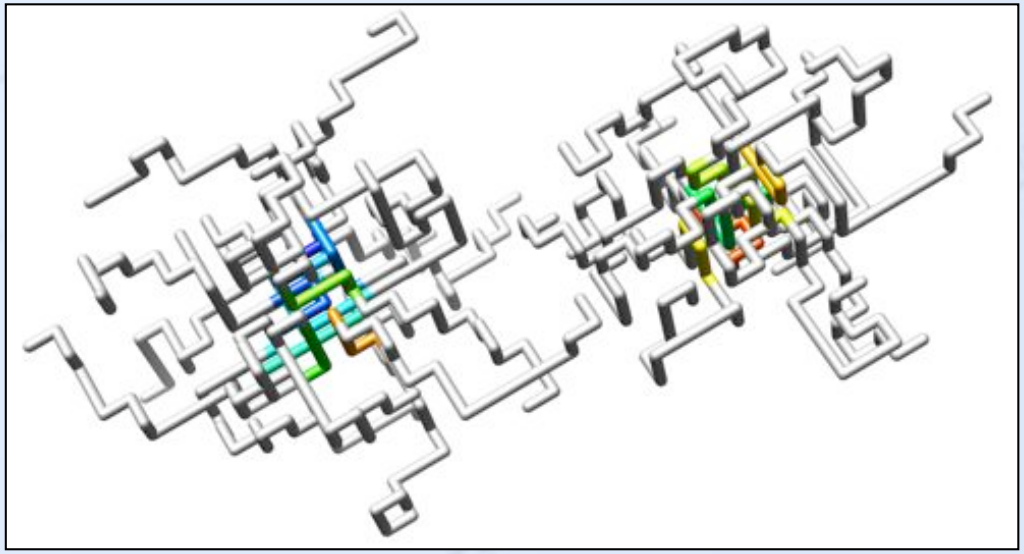
Binding motifs

300 x 

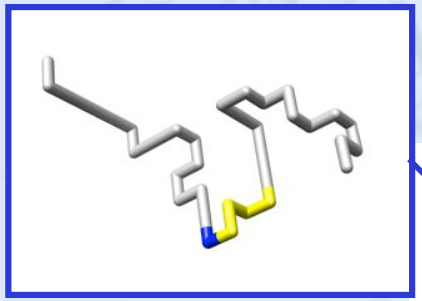


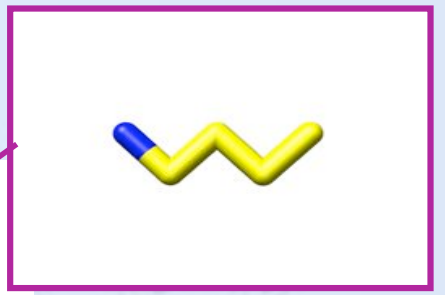
Binding motifs embedded in disordered flanks

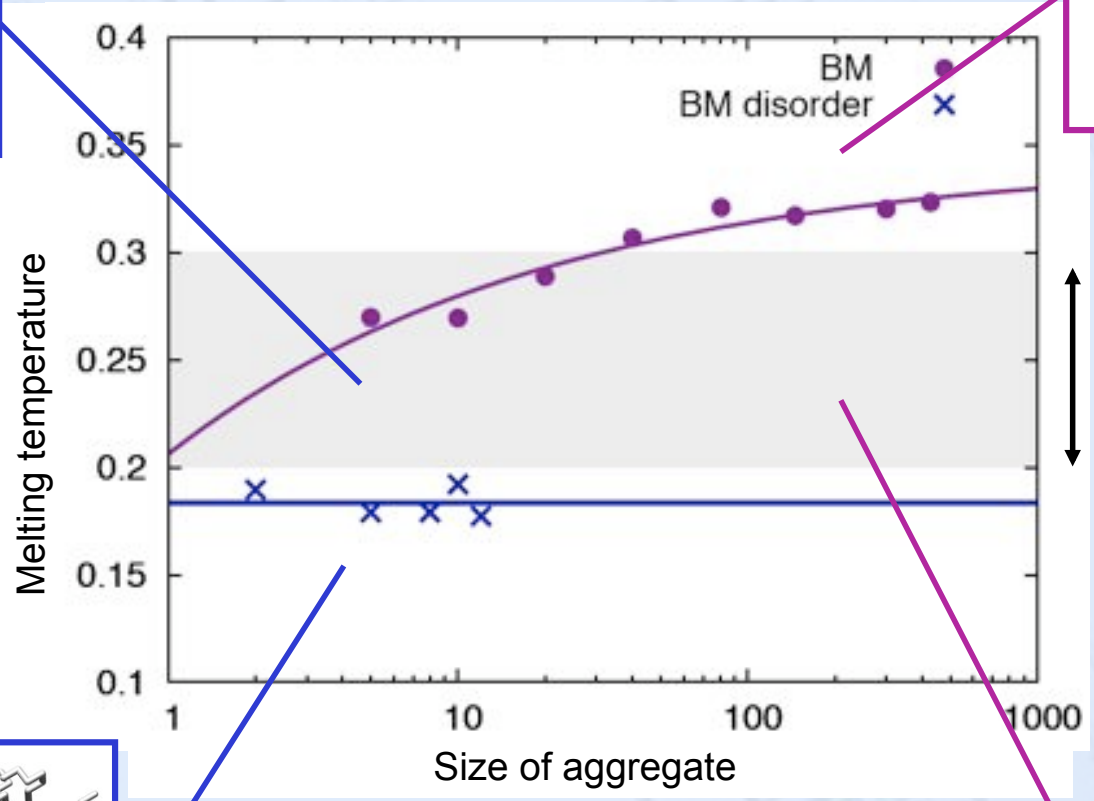
12 x 



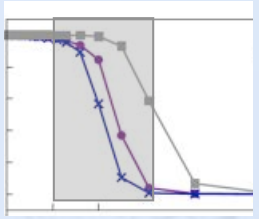
Disordered flanks prevent aggregation

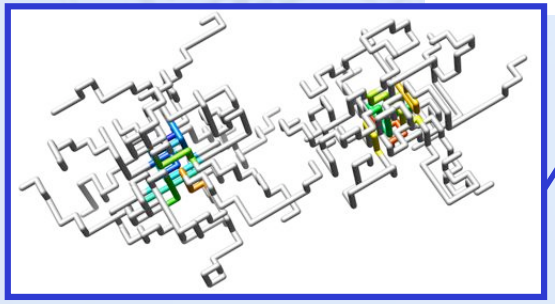


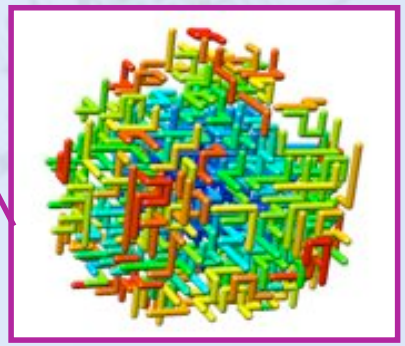




reversible
bindin g possible





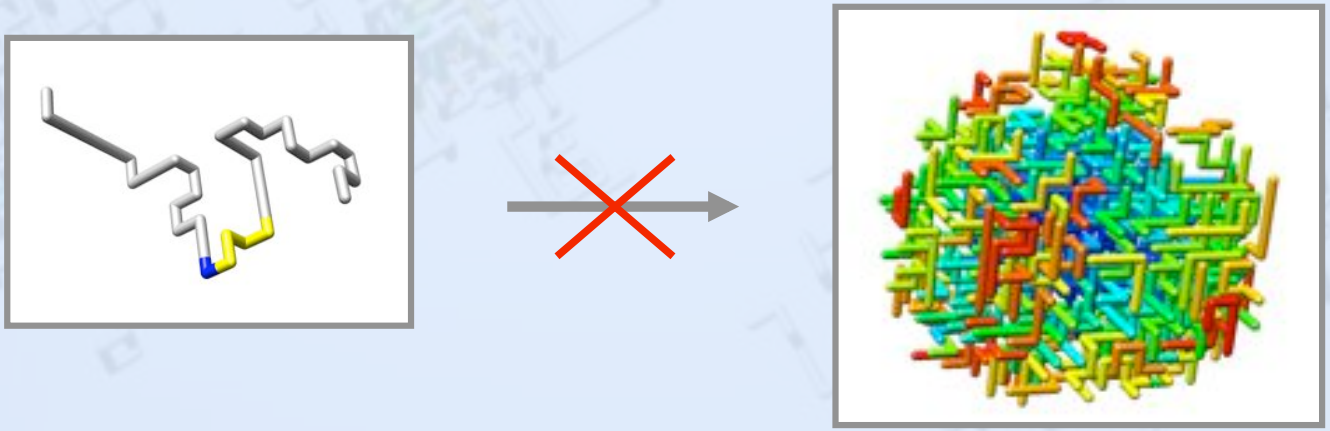


Conclusion

Disordered flanks:

- have little effect on binding of a flexible motif

- Prevent aggregation of multiple binding motifs

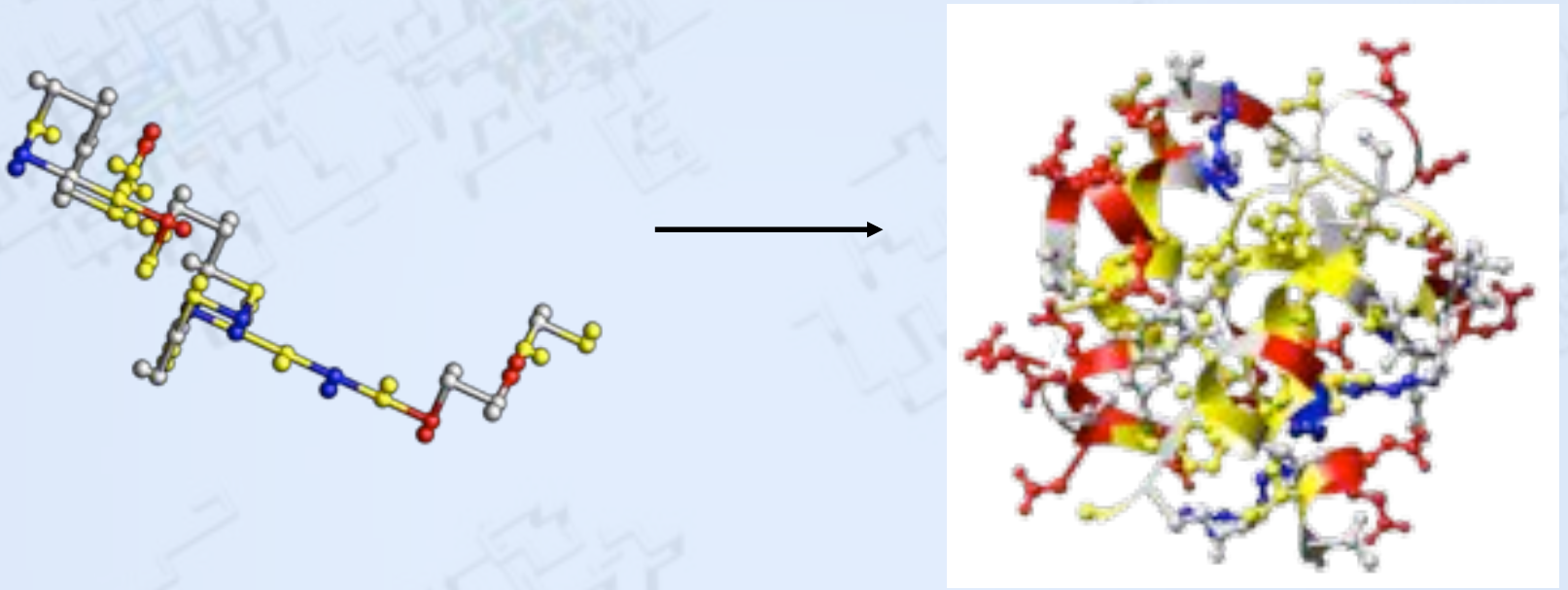


Abeln & Frenkel, PLoS Comp. Biol. 2008

Alternative mechanism to prevent aggregation

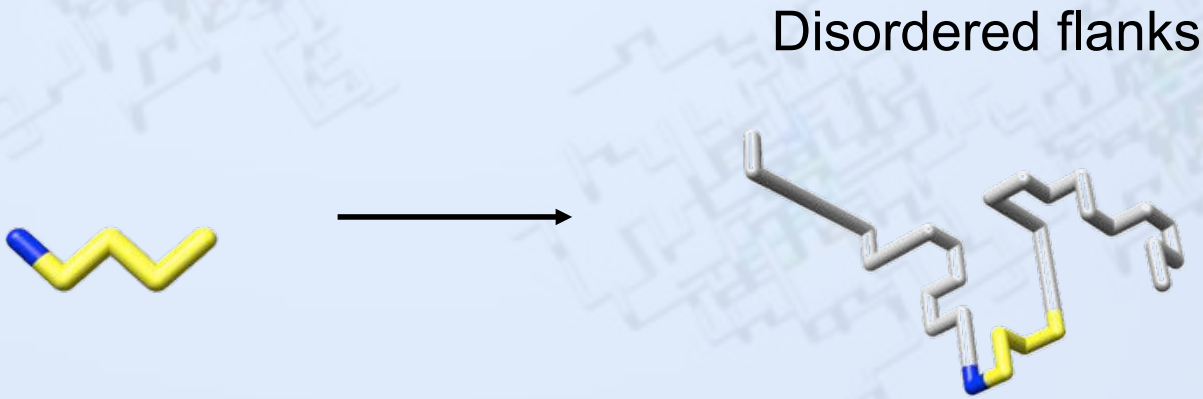
General mechanism to avoid protein aggregation:

Folding

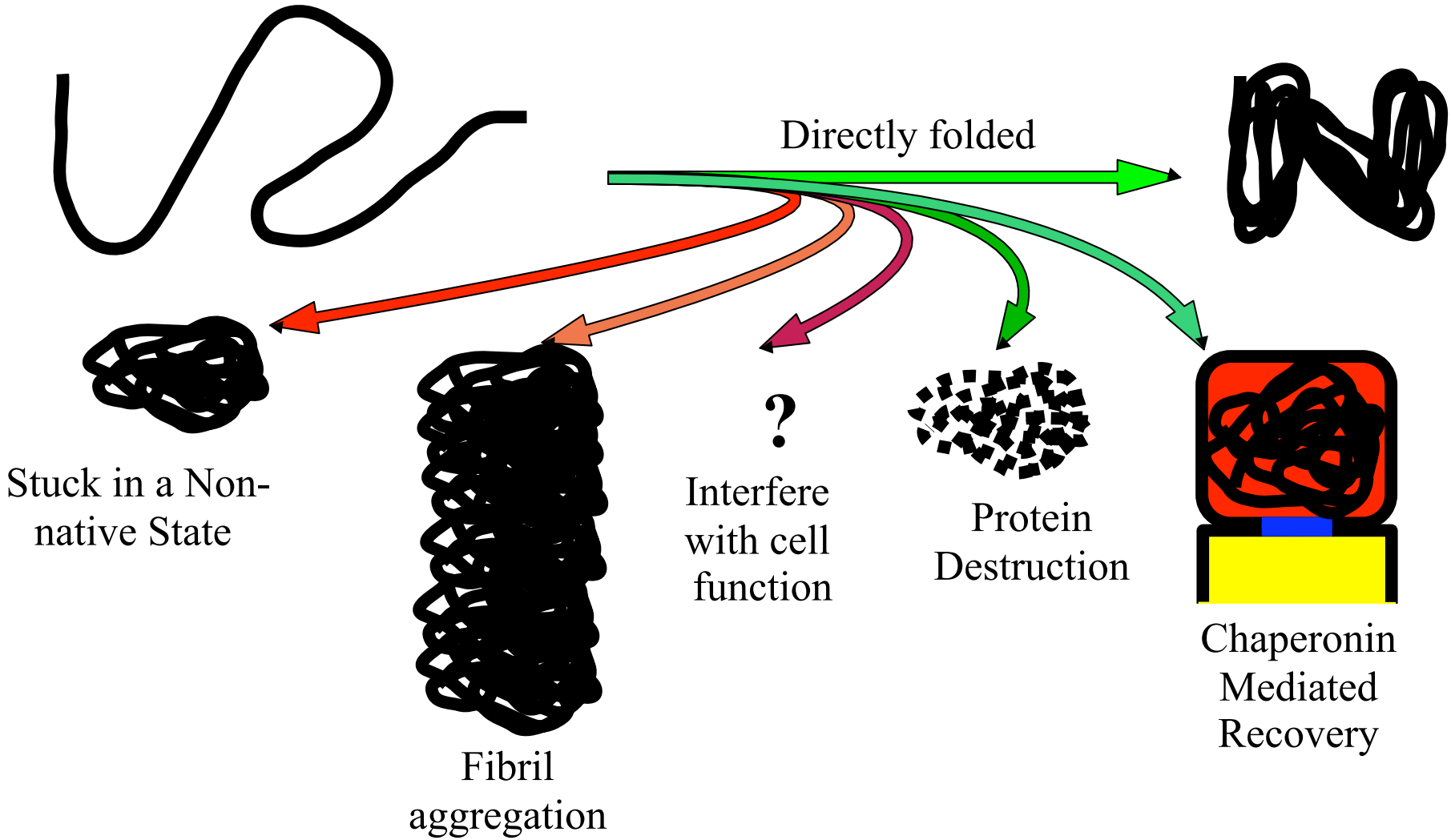


Alternative mechanism to prevent hydrophobic aggregation:

Disordered flanks



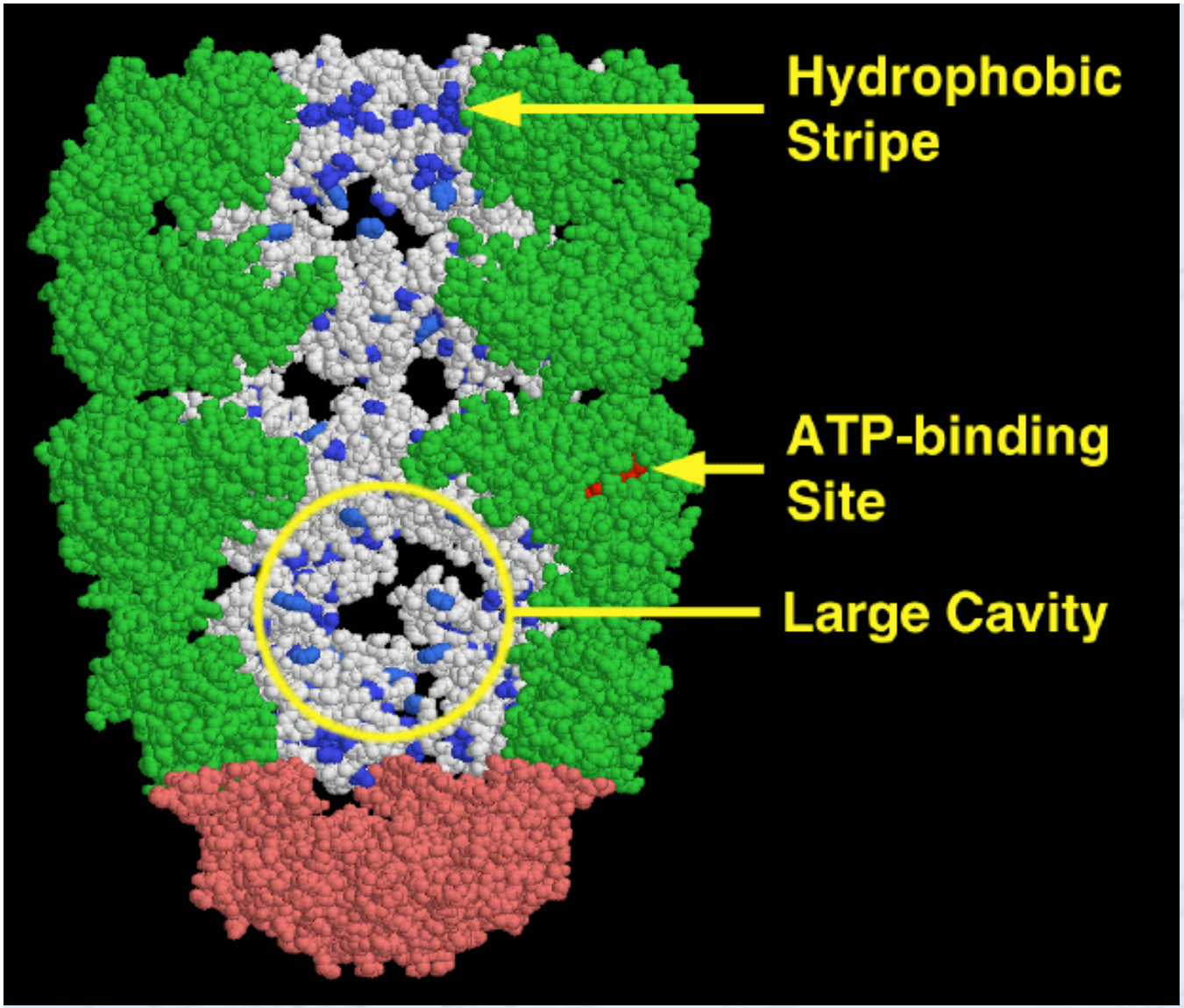
Chaperonins



Chaperones

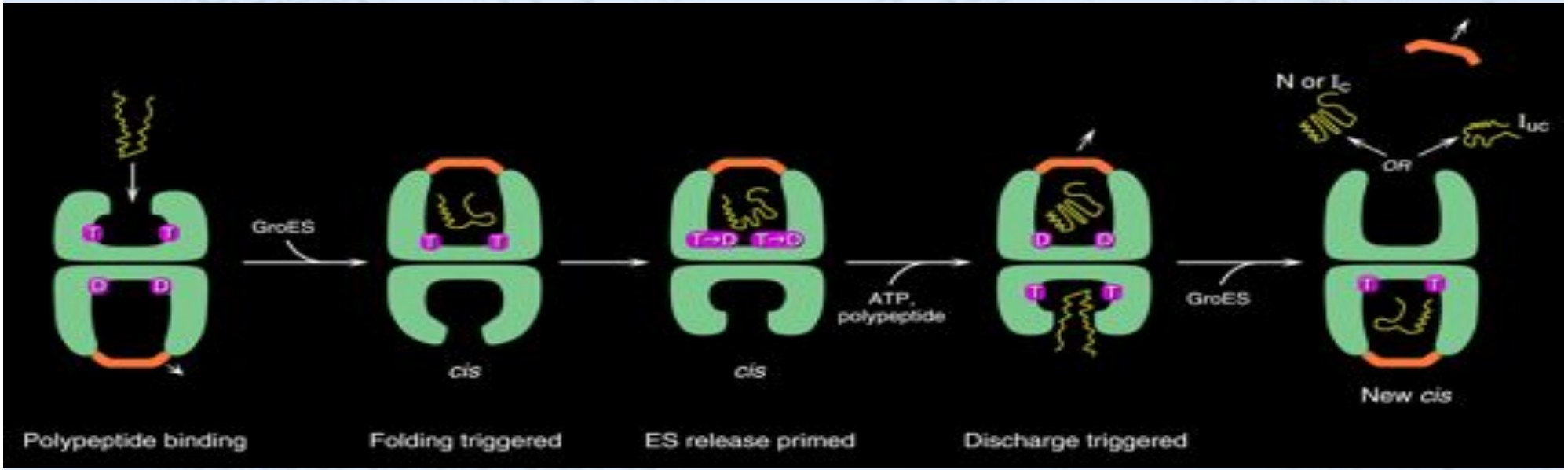
or

How Physicists may upset a biological community...



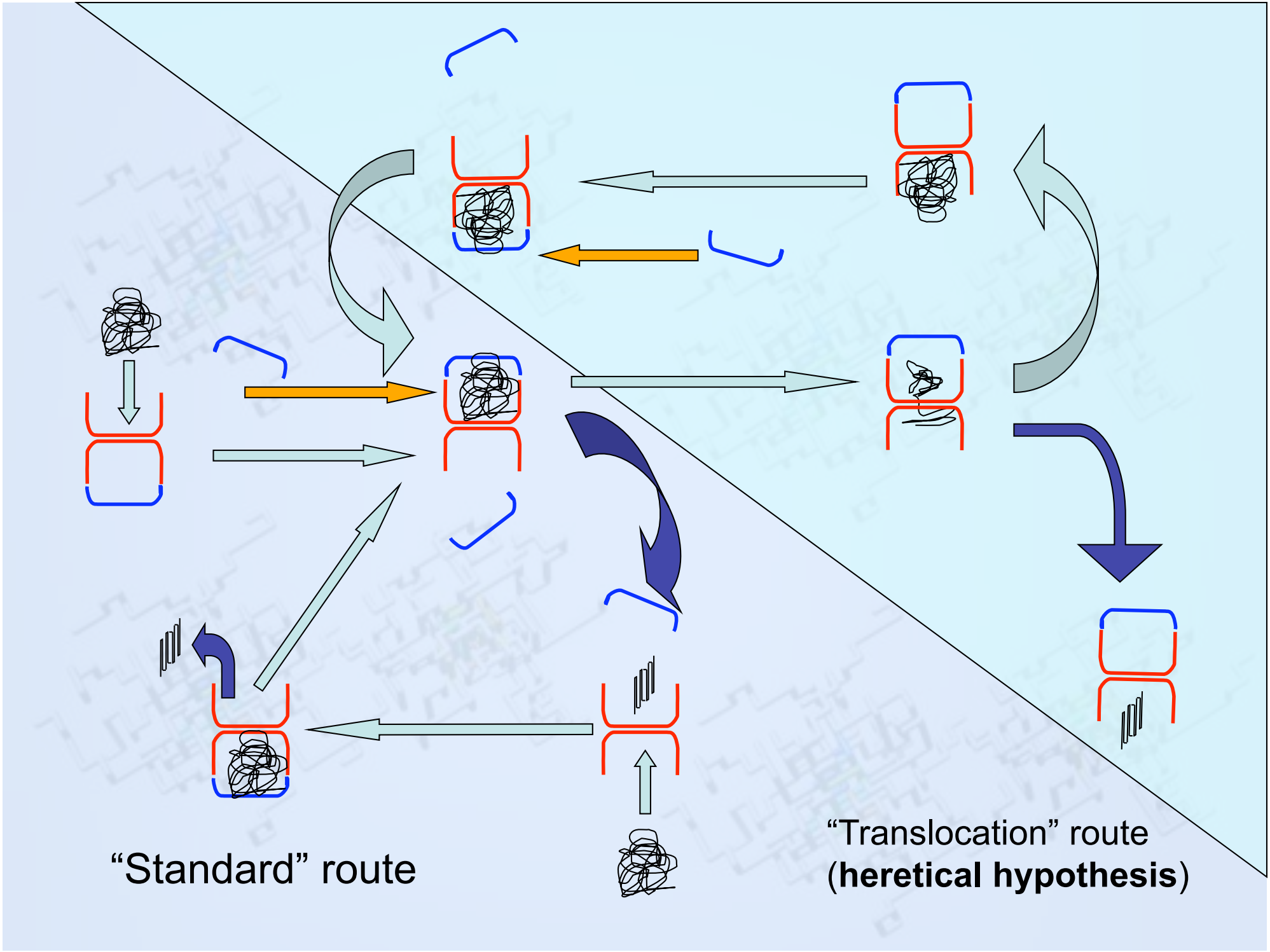
GroEL chaperonin complex folds medium-size proteins

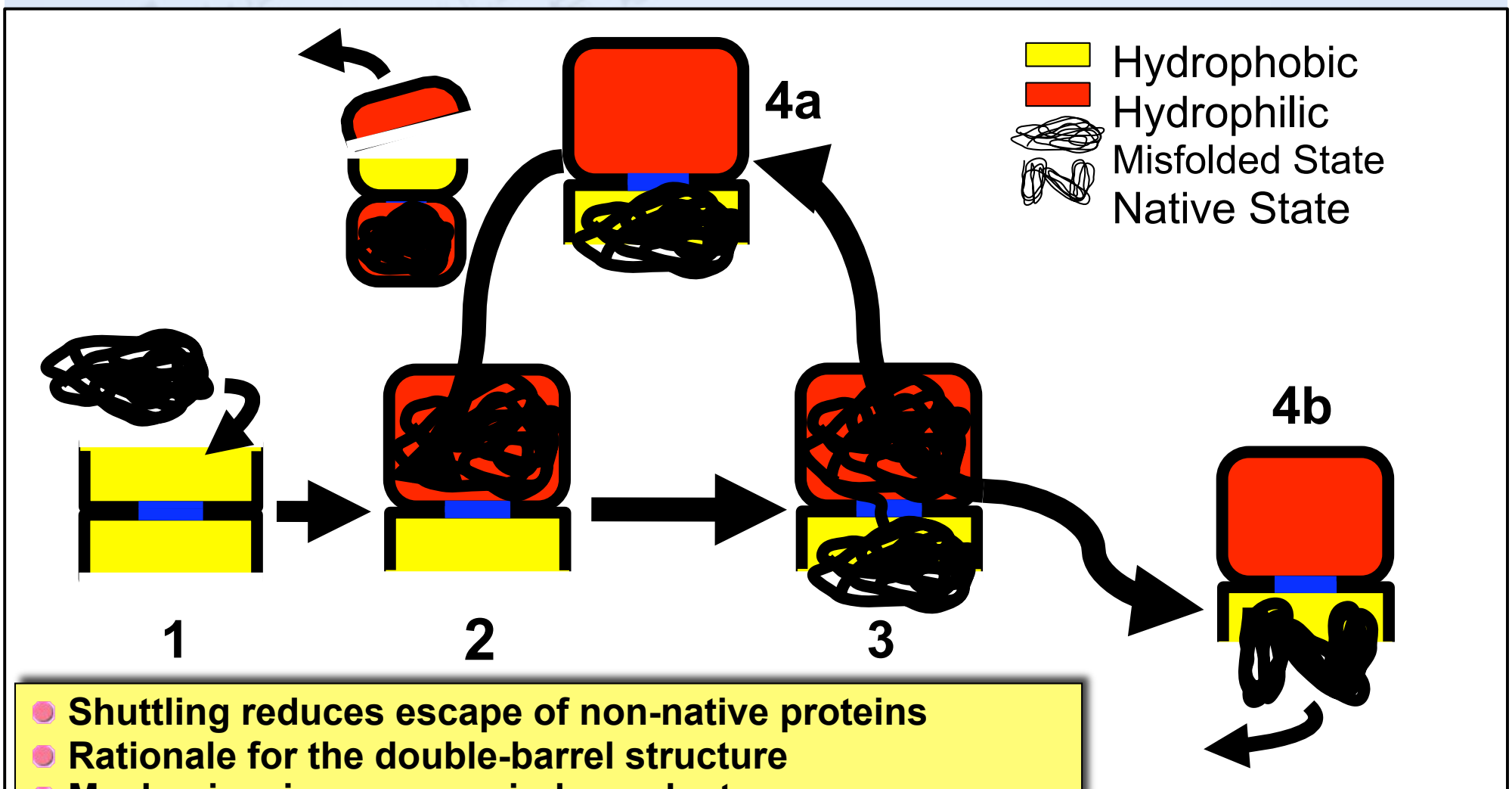
STANDARD MODEL



Mysteries:

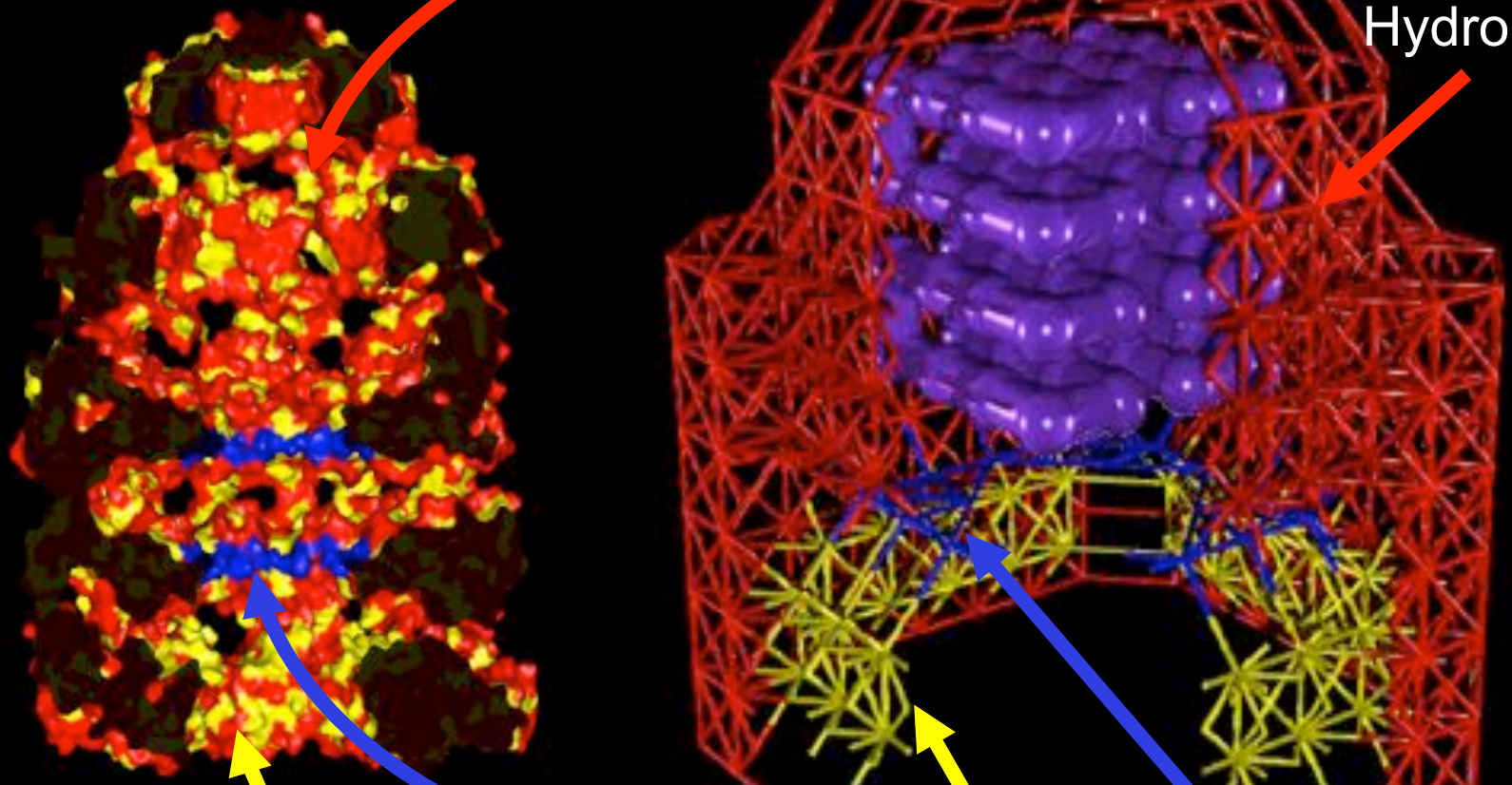
- Why two chambers,when a single one also works?
- Why are poorly-folded (“dangerous”) proteins allowed to escape before recapture (recapture probability ~30%)?


“Standard” route
“Translocation” route
(**heretical hypothesis**)



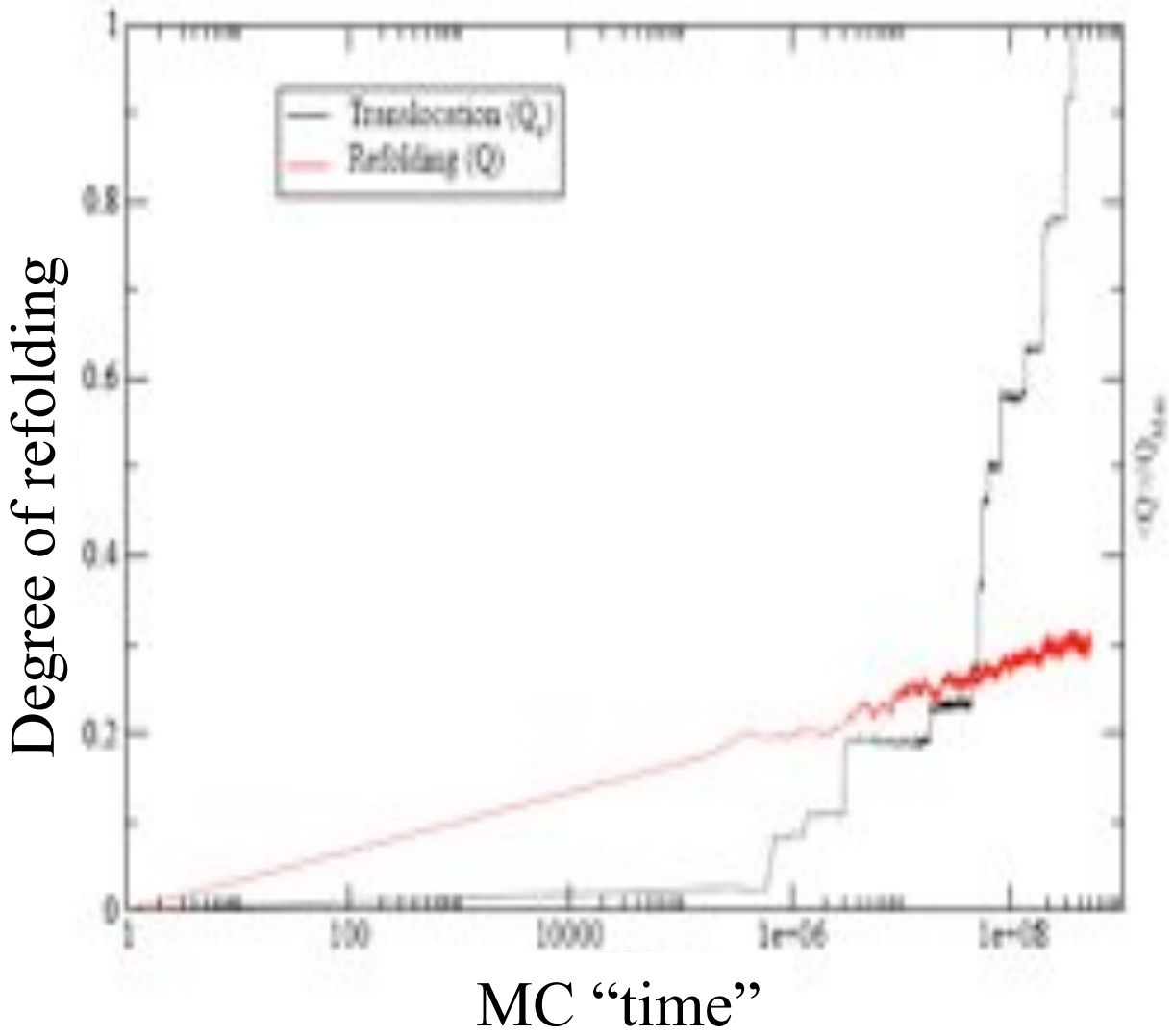
- Shuttling reduces escape of non-native proteins
- Rationale for the double-barrel structure
- Mechanism is sequence independent

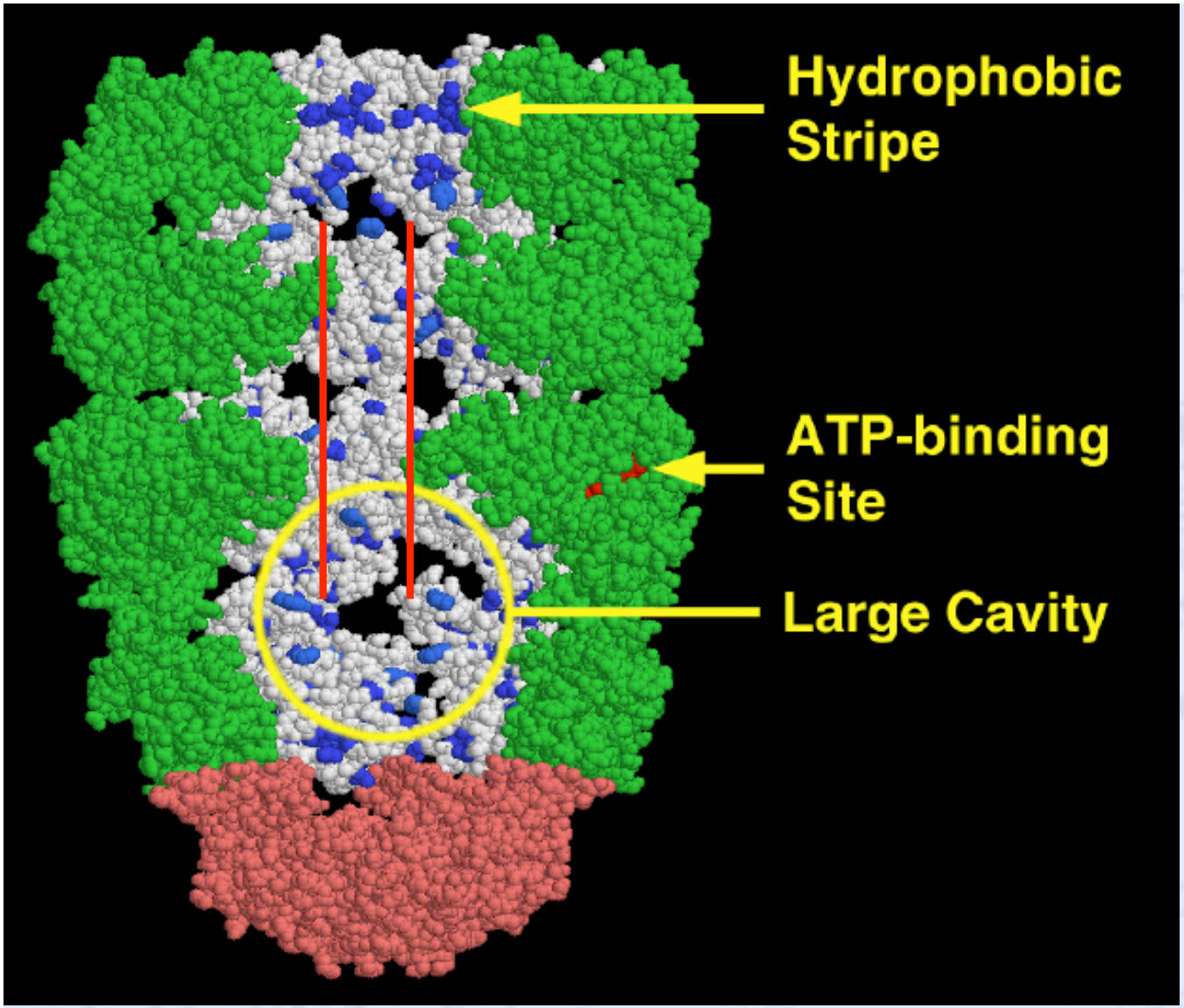
Miyazawa-Jernigan Lattice Model



- Simulation of this very simple model suggests:
  - Translocation is possible
  - It would suppress escape of seriously misfolded proteins.
  - … and, of course, in reality, things are not quite that simple…

Translocation is usually faster than refolding



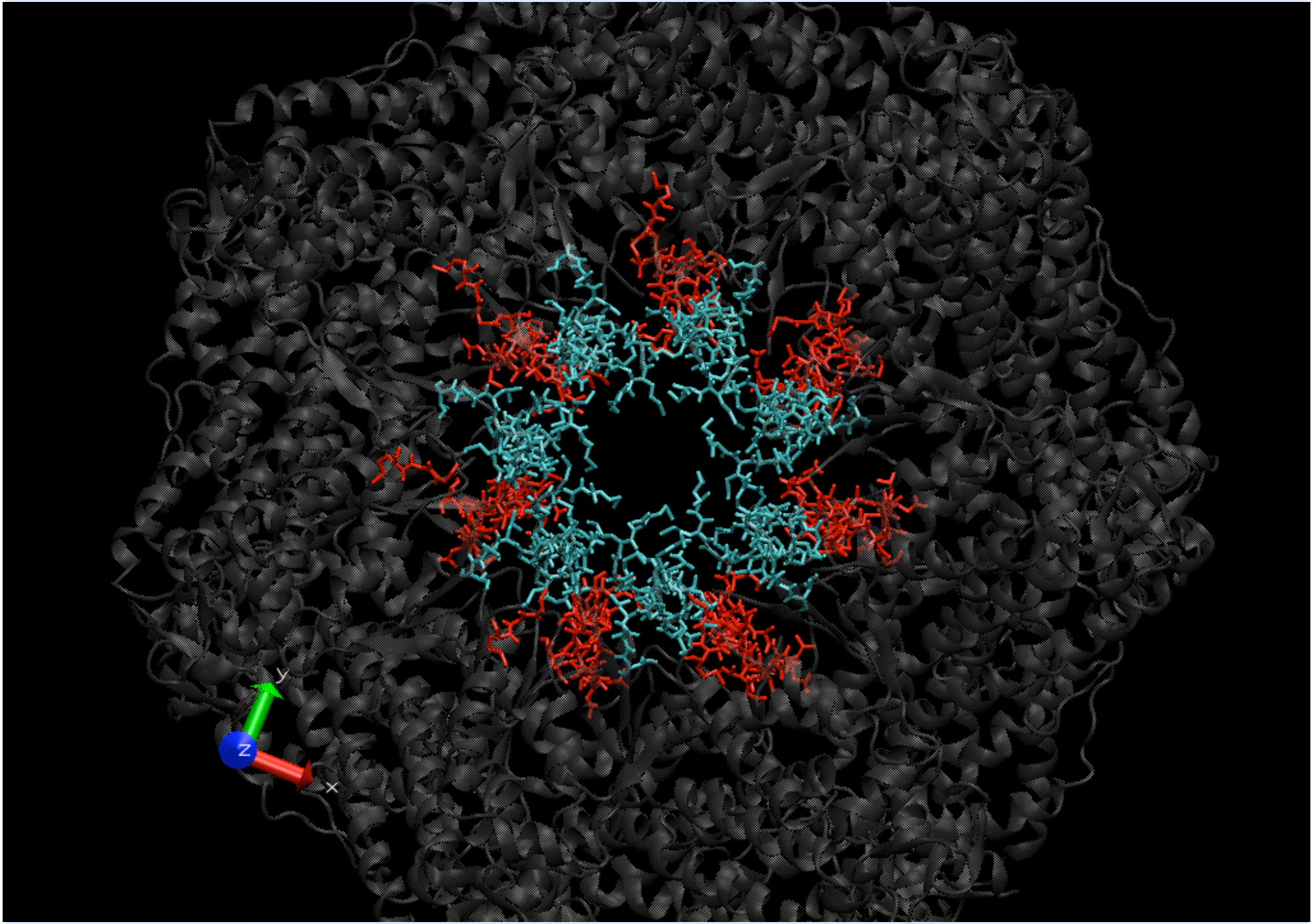

Hydrophobic Stripe
ATP-binding Site
Large Cavity

The X-ray scattering only sees ordered peptide chains.

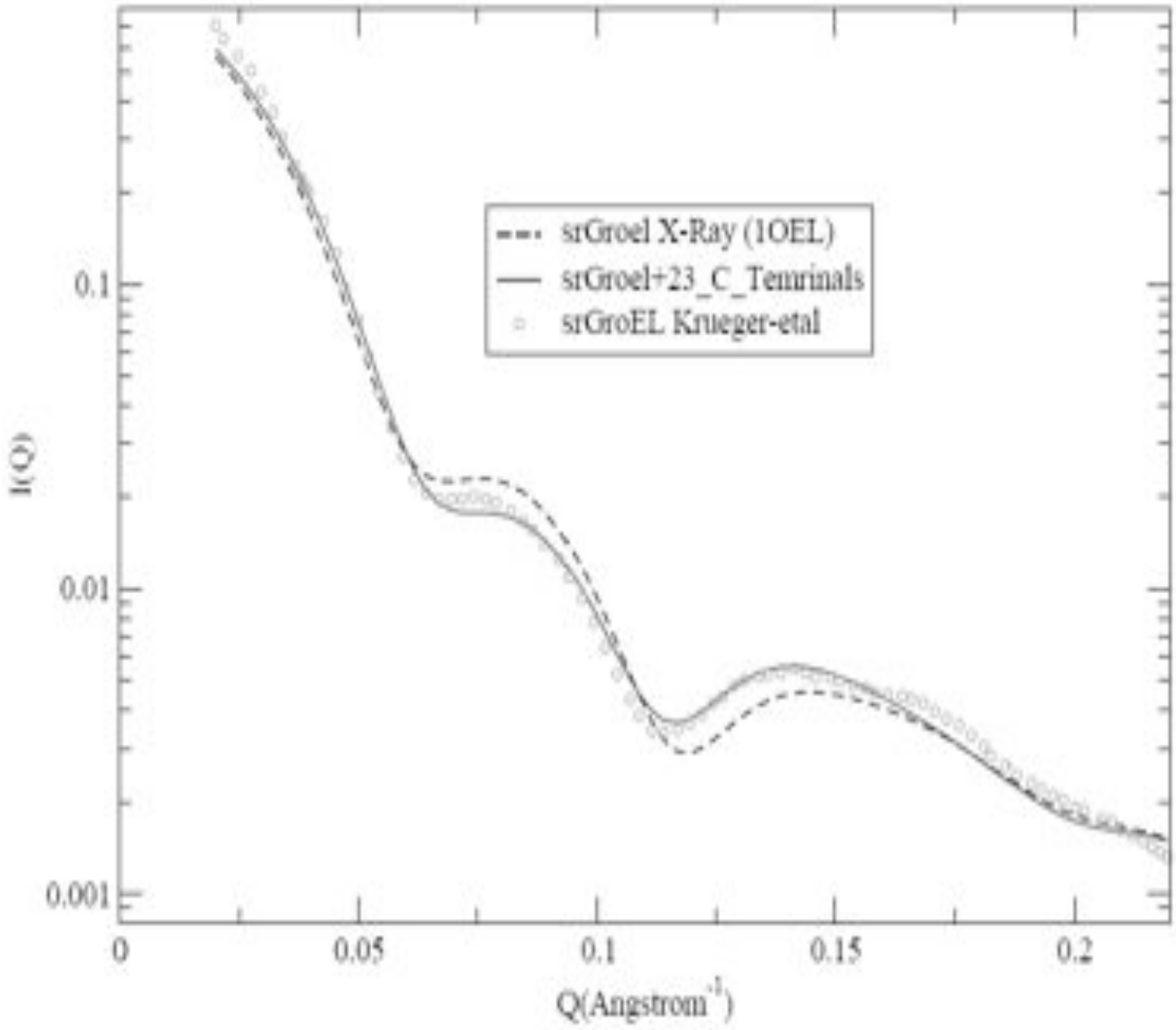
Evidence that the hole is filled comes from small-angle neutron scattering.

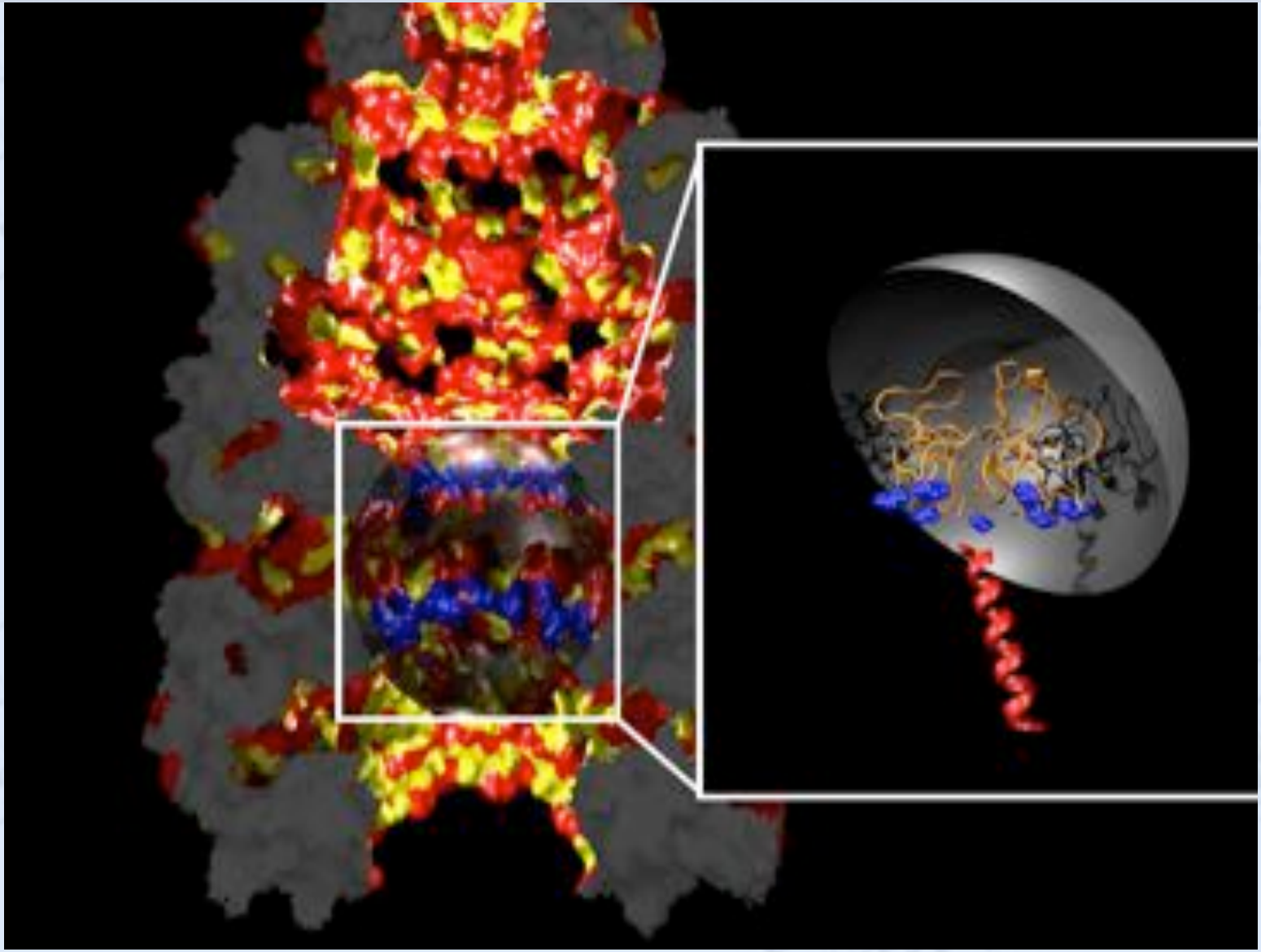
Is there enough room for proteins to translocate?

More precisely: Is the translocation picture compatible with X-ray scattering data?

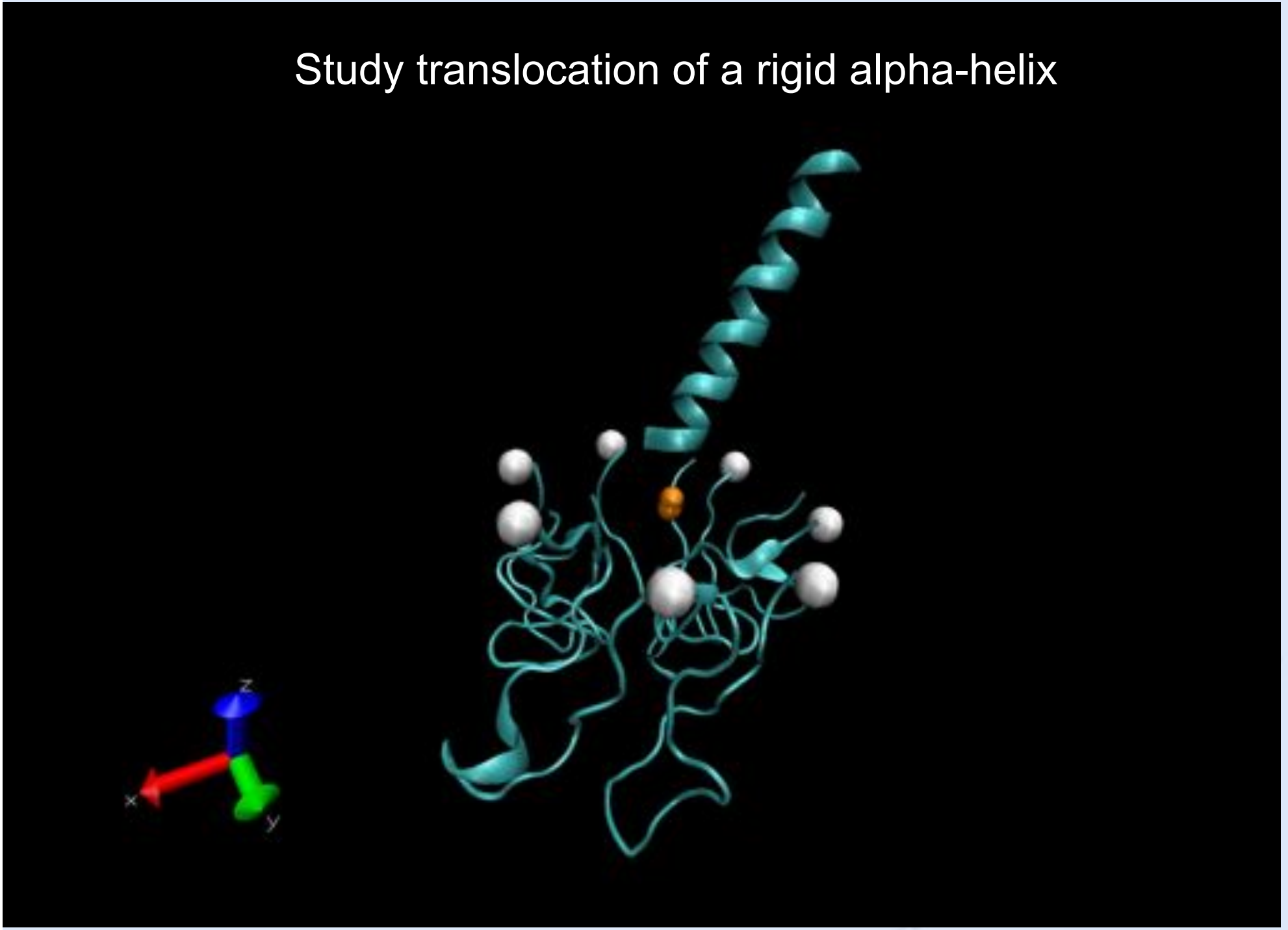

y
z
x

…and we DO reproduce the SAXS data.

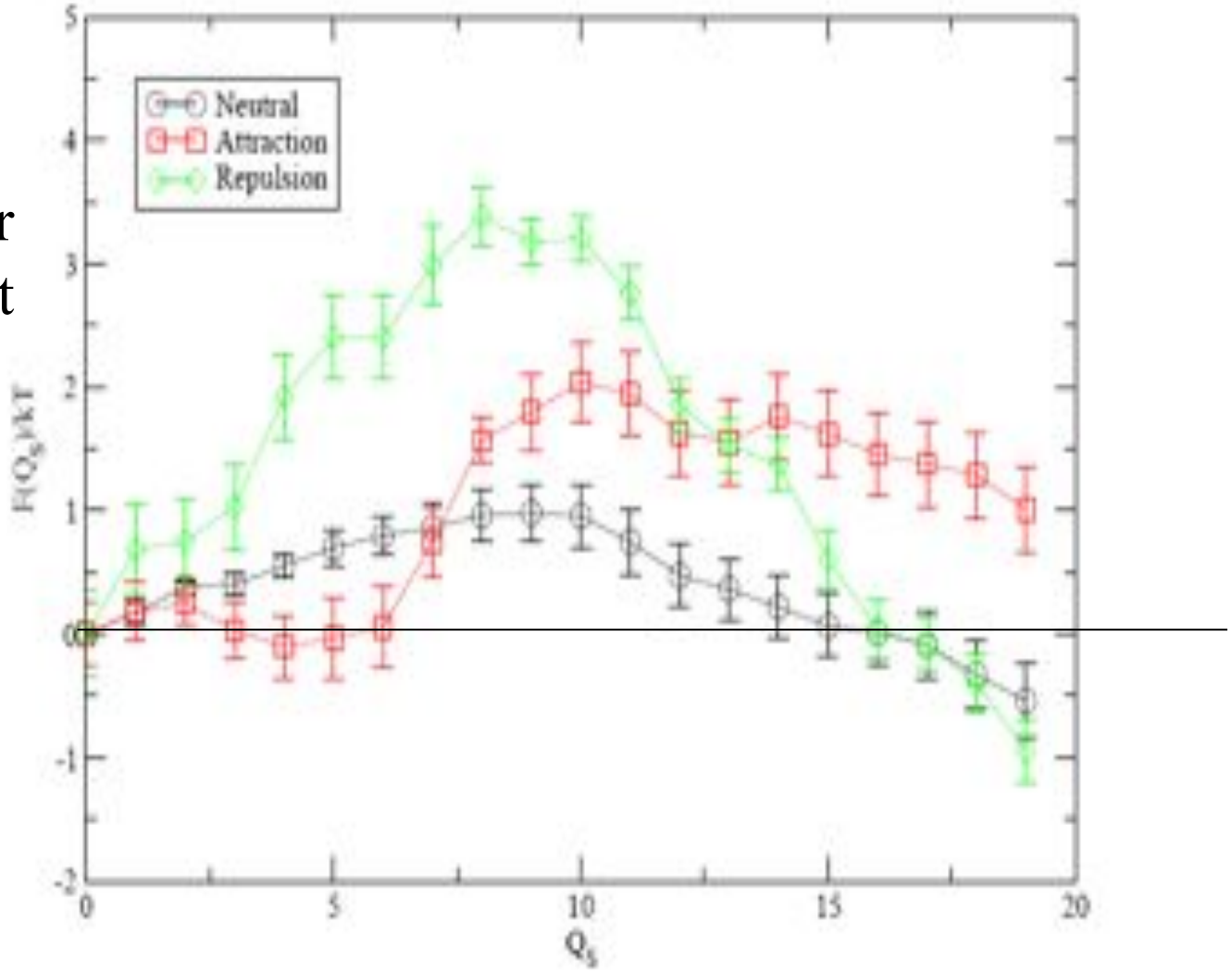




Study translocation of a rigid alpha-helix



Barrier height



Transported residues

The free-energy barrier for translocation can easily be overcome by thermal fluctuations…

References

- Disordered & binding

1: Abeln S, Frenkel D. Disordered flanks prevent peptide aggregation. PLoS Comput Biol. 2008 Dec;4(12):e1000241. Epub 2008 Dec 19.

- Chaperones

1: Coluzza I, De Simone A, Fraternali F, Frenkel D. Multi-scale simulations provide supporting evidence for the hypothesis of intramolecular protein translocation in GroEL/GroES complexes. PLoS Comput Biol. 2008 Feb 29;4(2):e1000006.

2: Coluzza I, van der Vies SM, Frenkel D. Translocation boost protein-folding efficiency of double-barreled chaperonins. Biophys J. 2006 May 15;90(10):3375-81. Epub 2006 Feb 10. PubMed

Lessons for full atomistic simulations - do not forget about evolution

- Try your simulation on a homologue (closely related sequence, with same function) - do your results hold?
- Make a sequence profile from homologues - are there any conserved residues? They may be important!
- Do not overtrust your potentials - most atomistic potentials are still “knowledge based”

Design Temperature