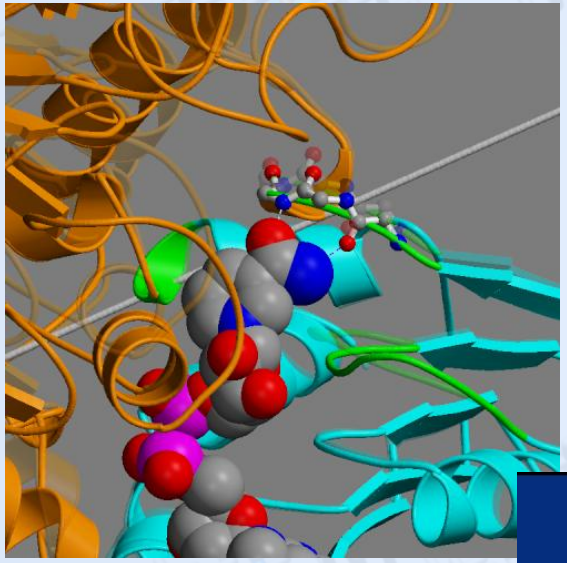
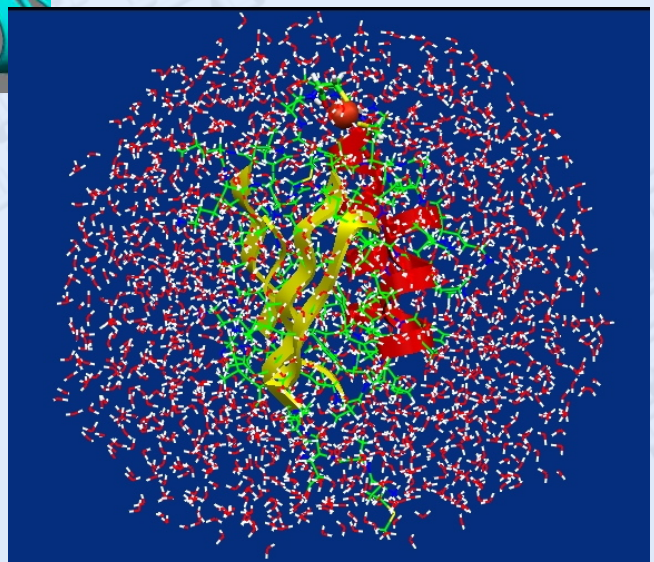
Biomolecular Simulation

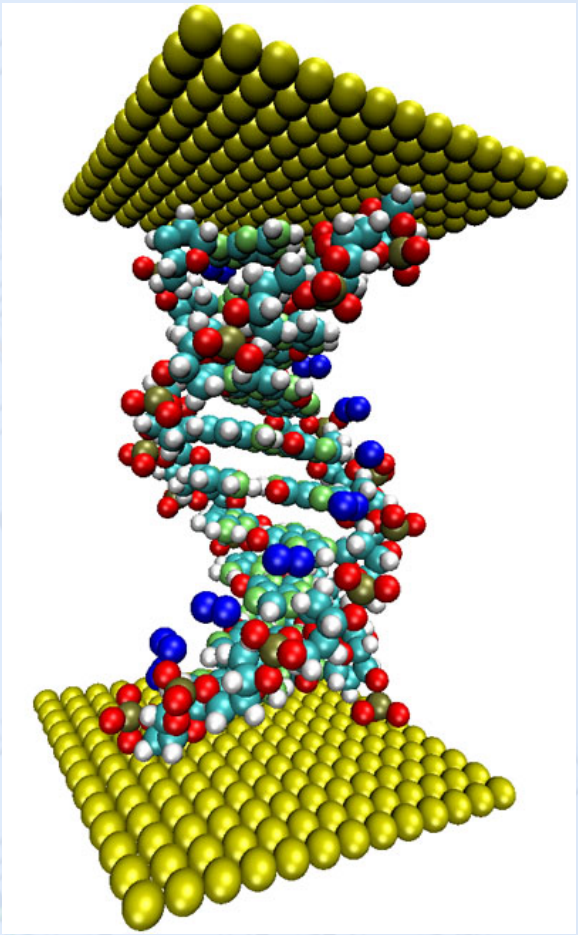
Where physics and biology meet

Sanne Abeln VU University s.abeln@vu.nl

Biomolecular simulation, some examples:







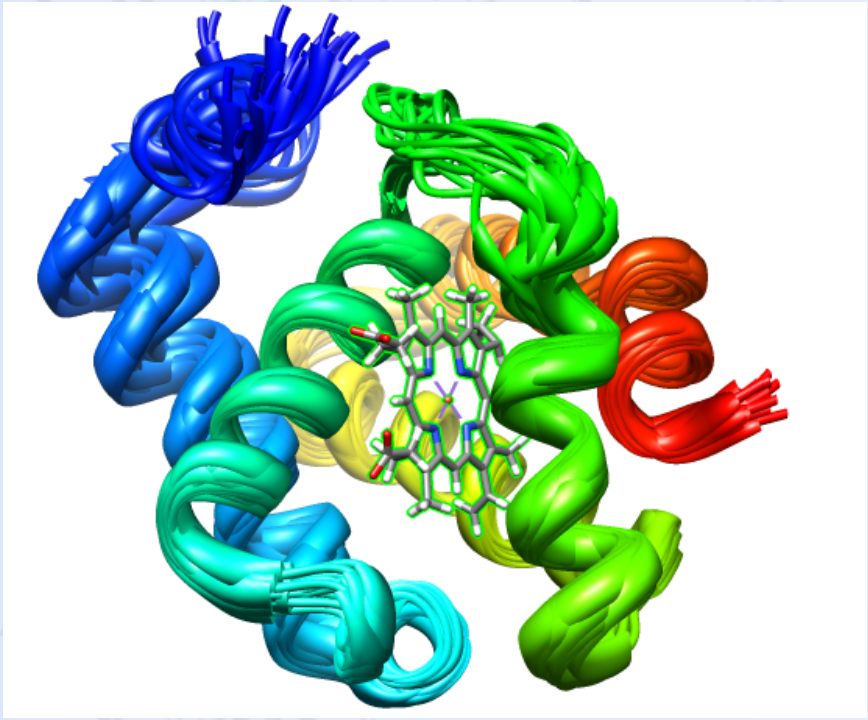
Why a lecture on biomolecular simulation?

- Who has worked with a simulation on biomolecules or is planning to do so?

My Background

- First Degree: Computer Science / Mathematics (University of Manchester)
- PhD: Structural Bioinformatics / Biochemistry (University of Oxford)
- PostDoc: Computational (Bio)-Physics (Amolf, Amsterdam)
- Currently, Assistant Professor Bioinformatics (VU University), use all of the above

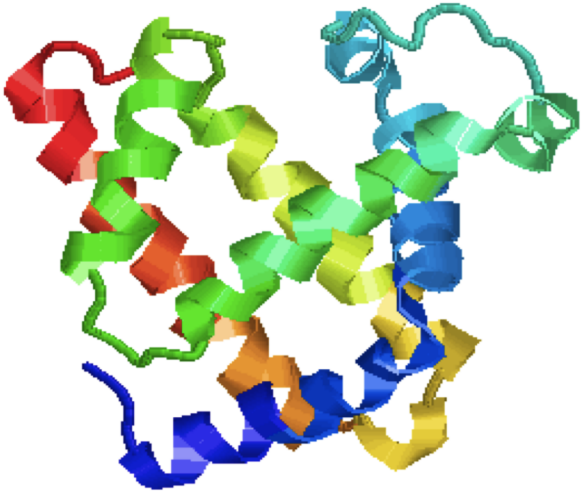
Why “Biomolecular Simulation”?



- 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 change the sequence of the protein?
  - Where does the substrate bind?
  - Do evolutionary related proteins bind the same substrate?

The biological approach or the physical approach?

?



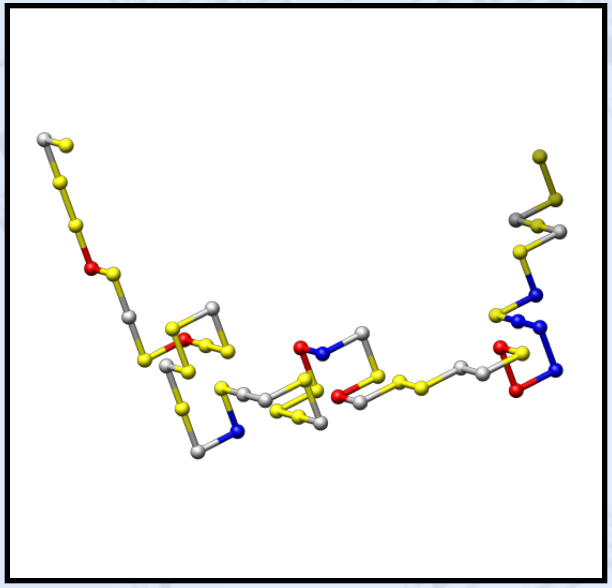
TASK:

Given a protein sequence predict in which 3D structure the protein will fold.

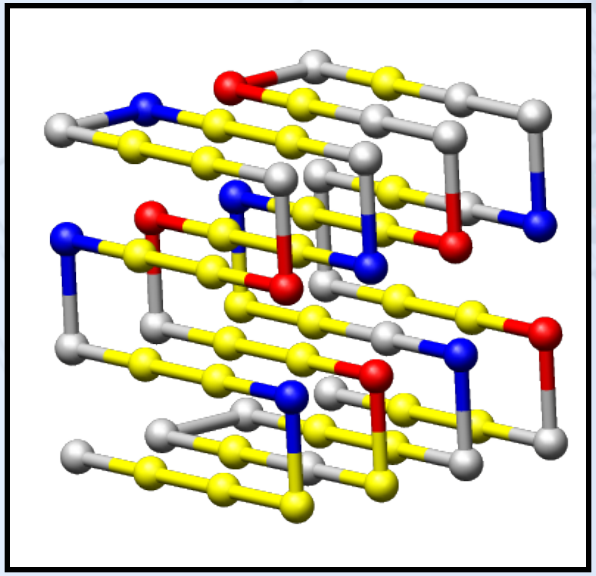
Who wins, the physical or biological approach?

Biomolecular simulation (this afternoon)

Sequence



Structure



hydrophobic

polar (hydrophilic)

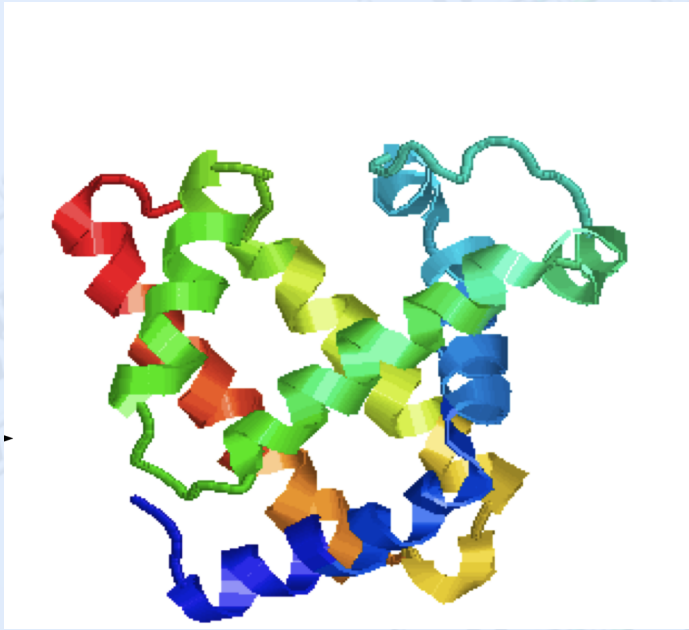
negative charge

positive charge

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

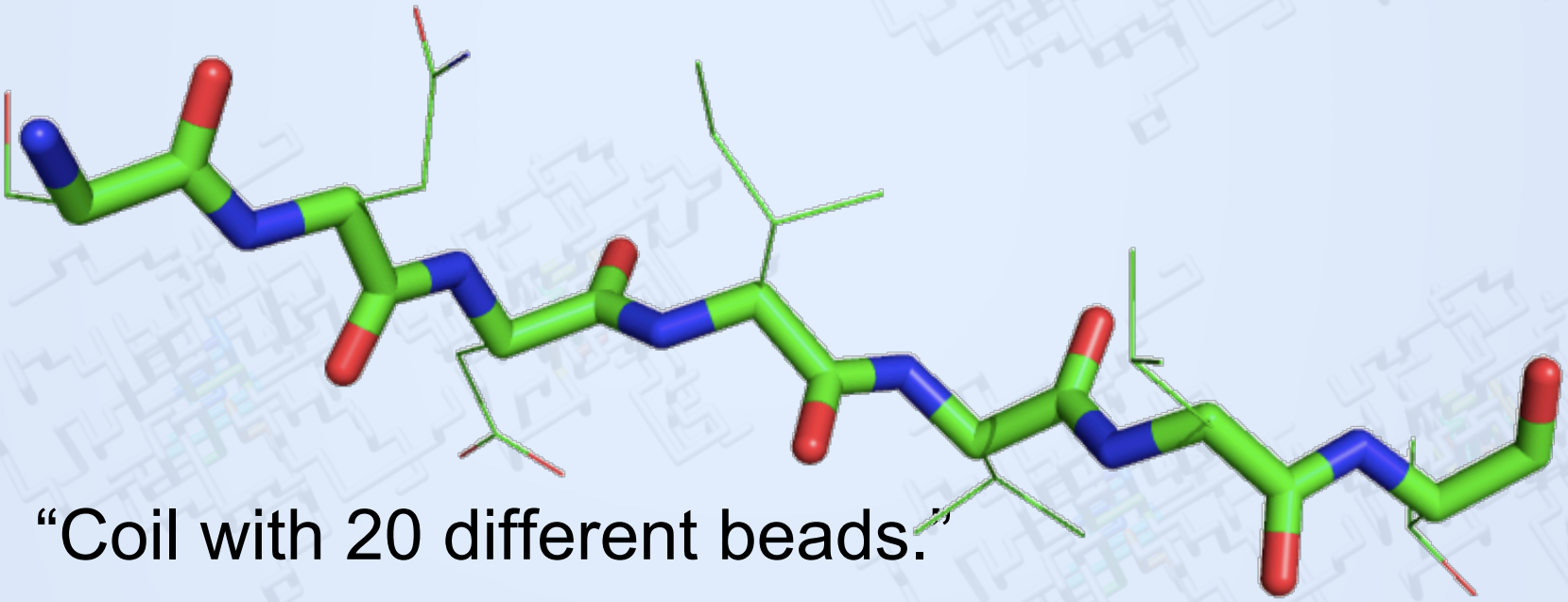
Let’s have a look at the physical forces first

- What are the most important forces that act on a protein?
- We stay away from quantum mechanics for now.



Lets consider the basic protein chemical structure

- Identical backbone for each residue (peptide)
- Amino acid side chains with 20 different chemical structures



“Coil with 20 different beads.”

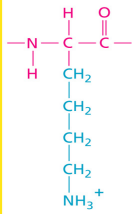
What type of forces and effects would be relevant?

- Van der Waals
- Electrostatics
- Hydrogen bonding
- Entropic effects
- Hydrophobic effect

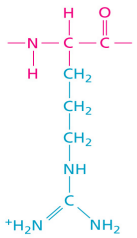
Primary Protein Structure

+ve charge

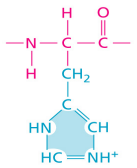
lysine (Lys, or K)



arginine (Arg, or R)

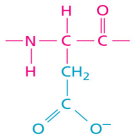


histidine (His, or H)

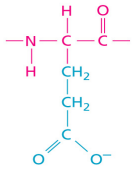


-ve charge

aspartic acid (Asp, or D)

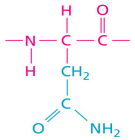


glutamic acid (Glu, or E)

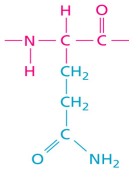


other polar

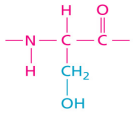
asparagine (Asn, or N)



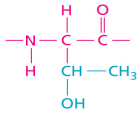
glutamine (Gln, or Q)



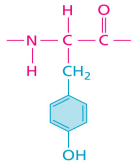
serine (Ser, or S)



threonine (Thr, or T)

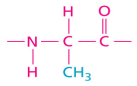


tyrosine (Tyr, or Y)

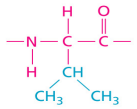


hydrophobic

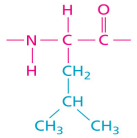
alanine (Ala, or A)



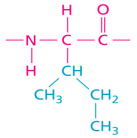
valine (Val, or V)



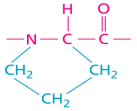
leucine (Leu, or L)



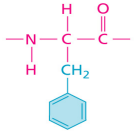
isoleucine (Ile, or I)



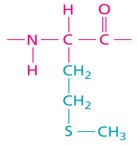
proline (Pro, or P)



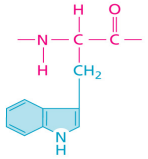
phenylalanine (Phe, or F)



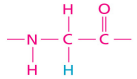
methionine (Met, or M)



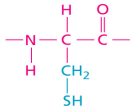
tryptophan (Trp, or W)



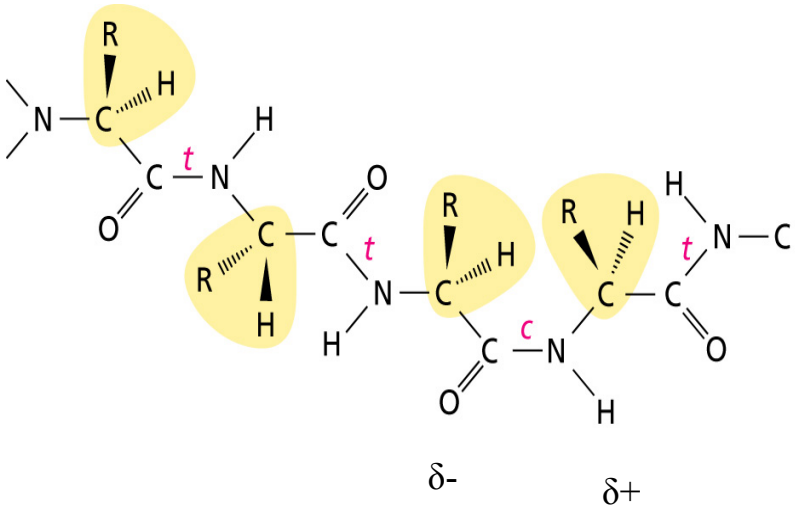
glycine (Gly, or G)



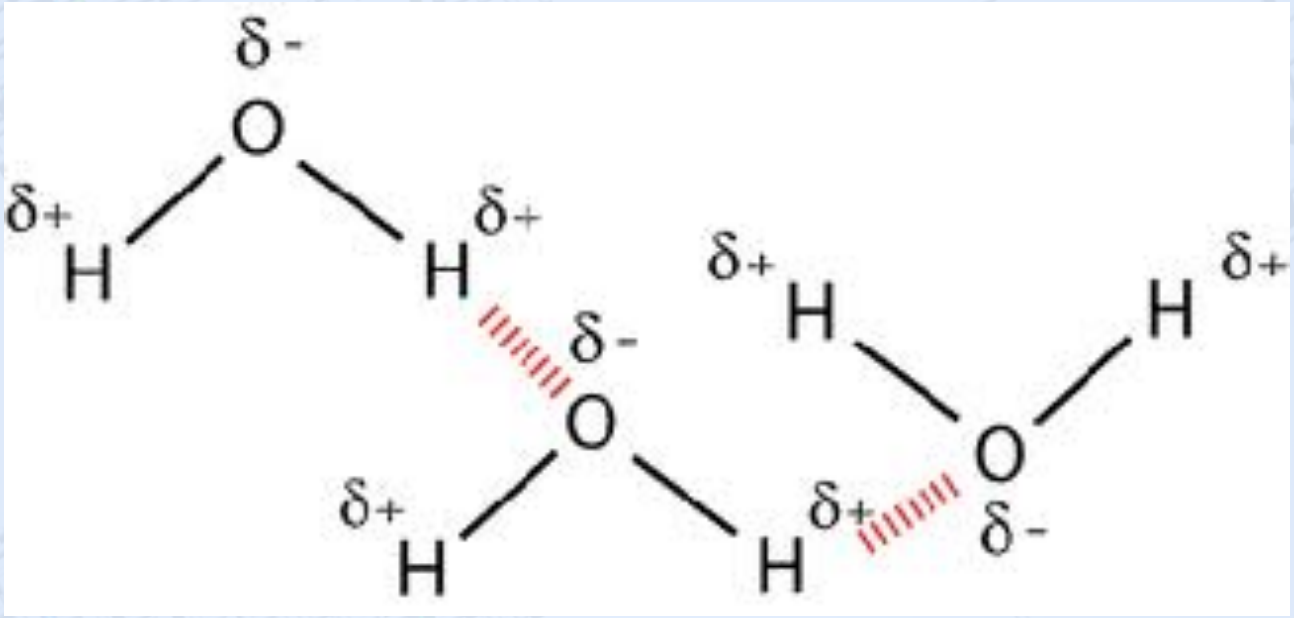
cysteine (Cys, or C)



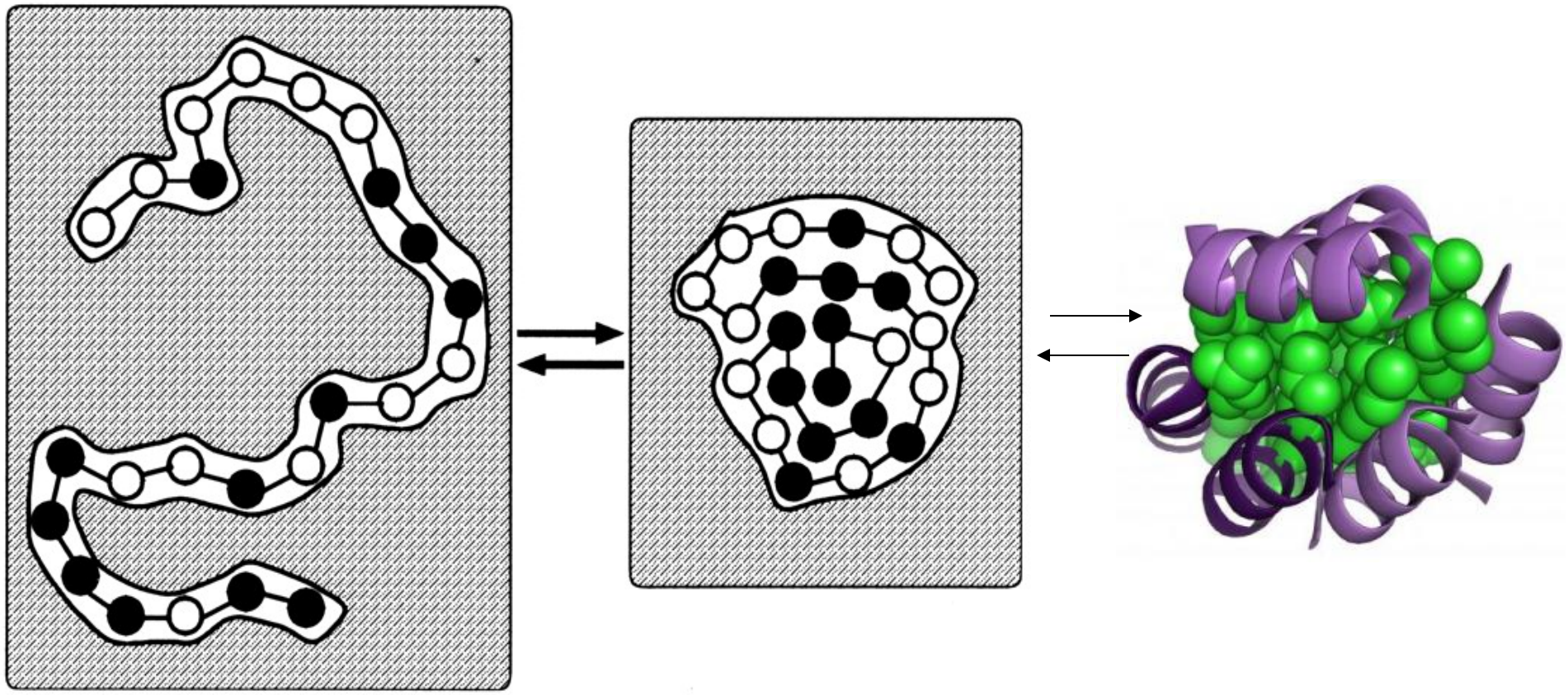
Backbone has a hydrogen bond donor and acceptor per residue



Proteins live in water

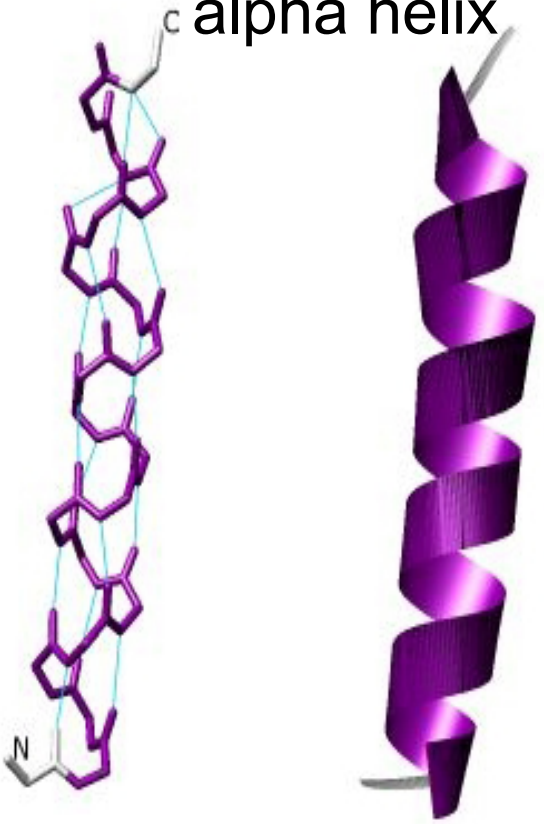


Hydrophobic Collapse

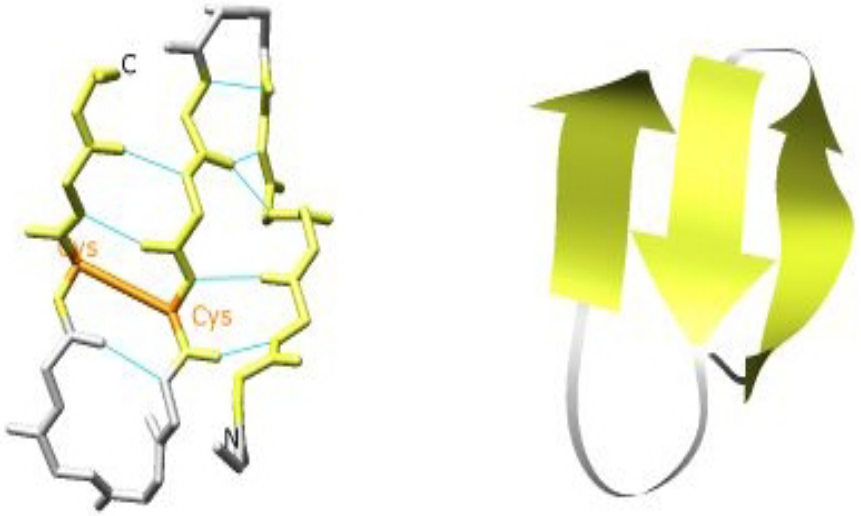


Hydrogen bonds & secondary structure of backbone

alpha helix

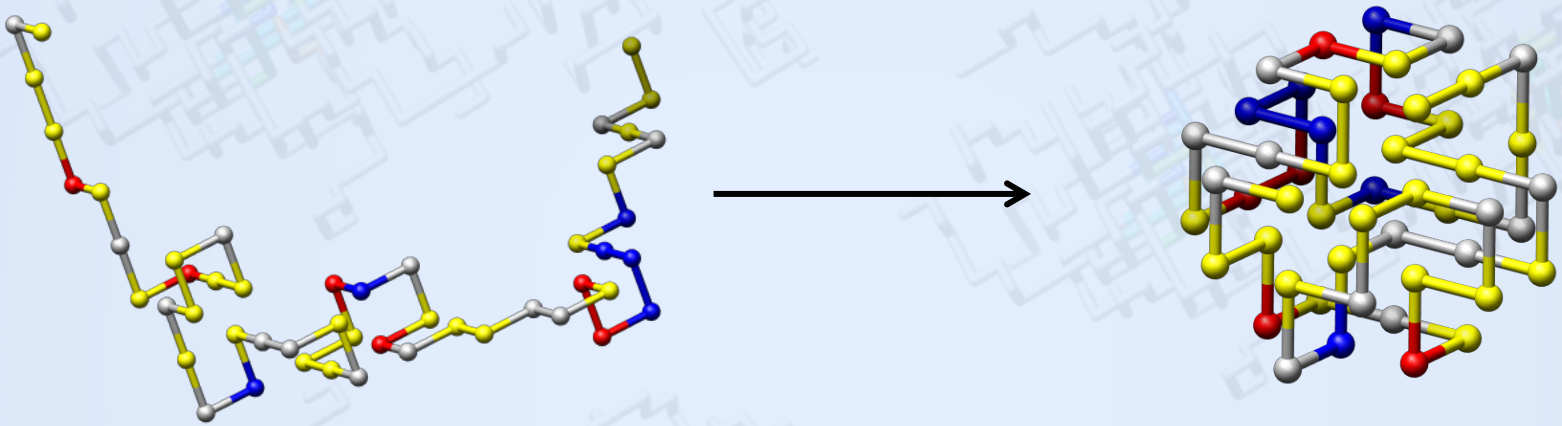


beta sheet



We will ignore this in the practical exercises!

What are the effects that contribute to a stably folded protein?

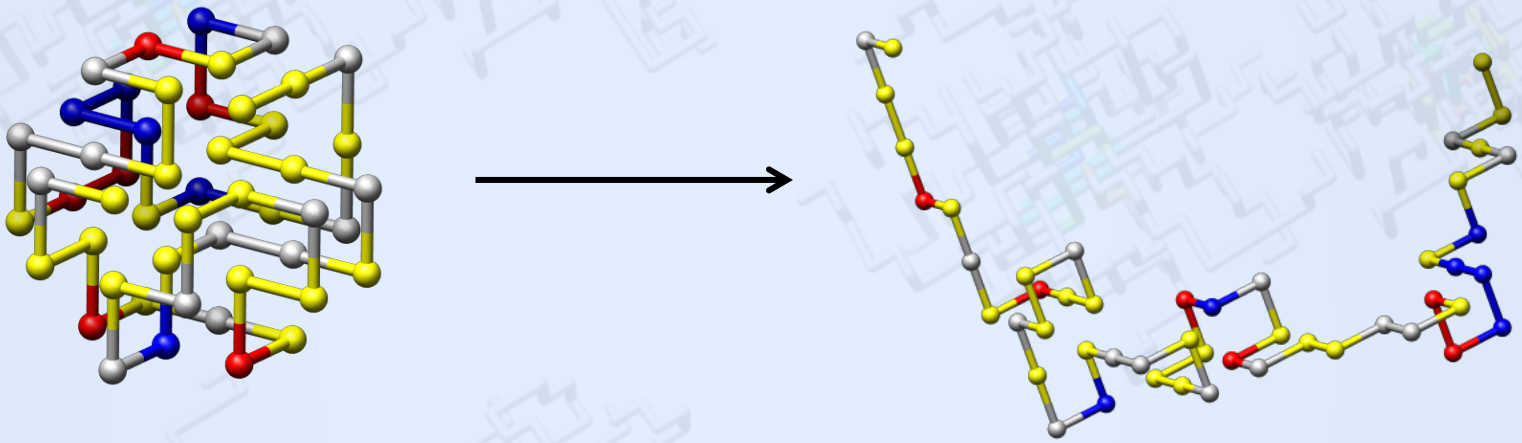


1) Hydrophobicity (oil in water)

   - note this is an effective force that contains enthalpic and entropic components

2) Hydrogen bonds form secondary structure

What forces / effects destabilize a folded protein?



1) Chain entropy!

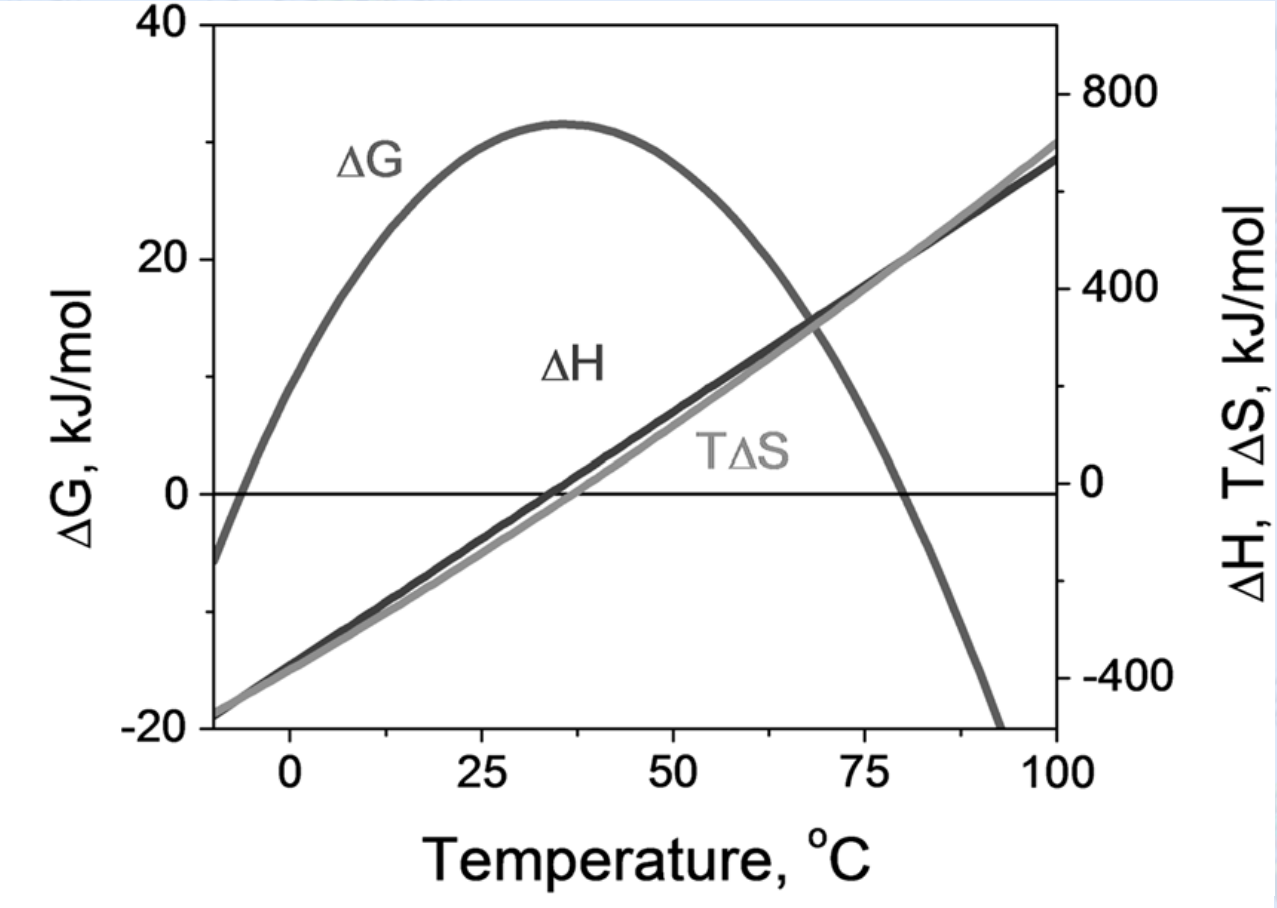
Questions for you:

Would secondary structures form / be stable in vacuum?

What is the influence of water molecules on secondary structure formation (does it help, does it hinder formation)?

How can you explain secondary structure formation in an aqueous environment?

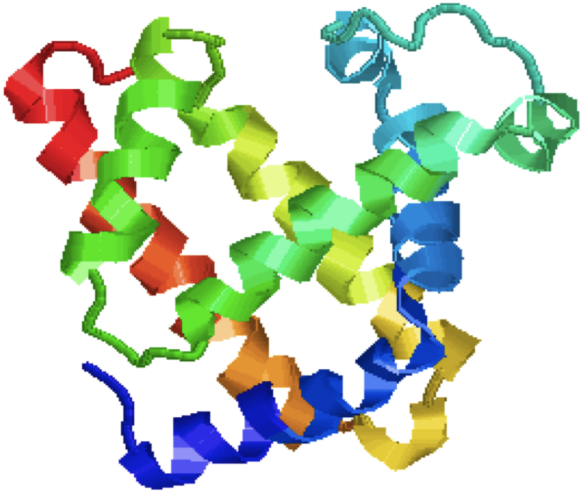
Entropic and enthalpic contributions compensate (experimental)



Look at the scale of the axis!

The biological approach or the physical approach?

?

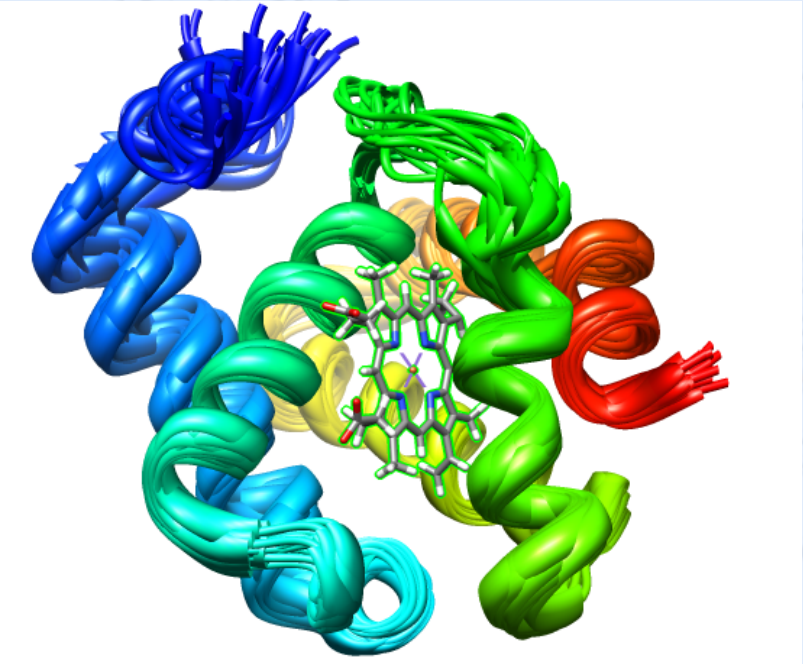


TASK:

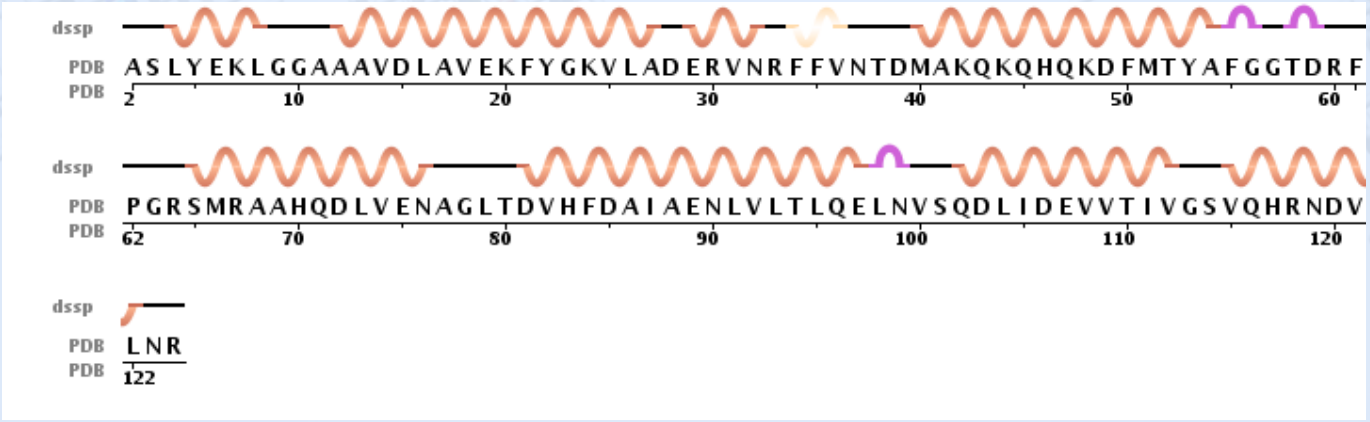
Given a protein sequence predict in which 3D structure the protein will fold.

Who wins, the physical or biological approach?

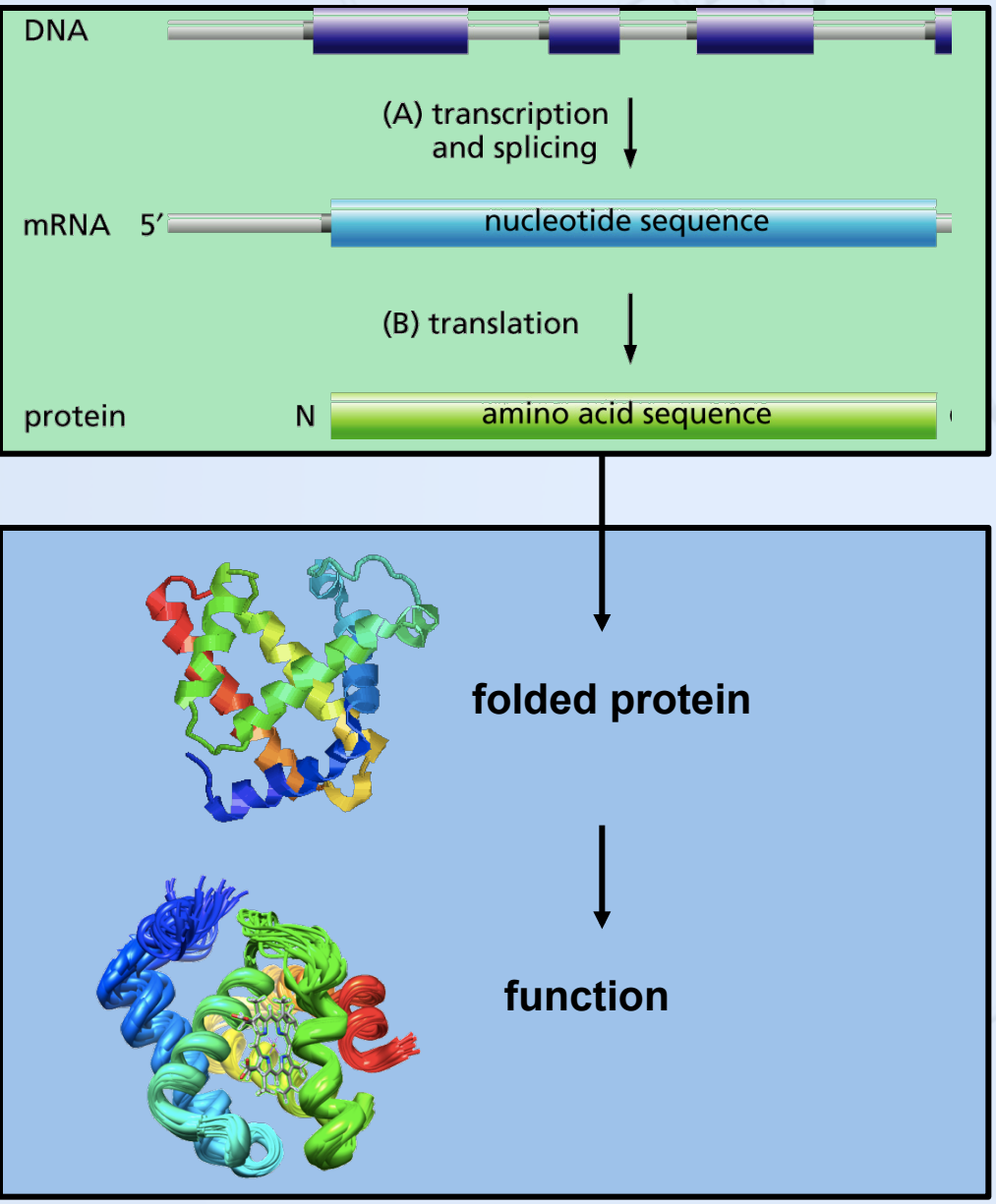
What about the biology?



The sequence is key to biological understanding



DNA - RNA – Proteins- function



The DNA is the source of information for a living species

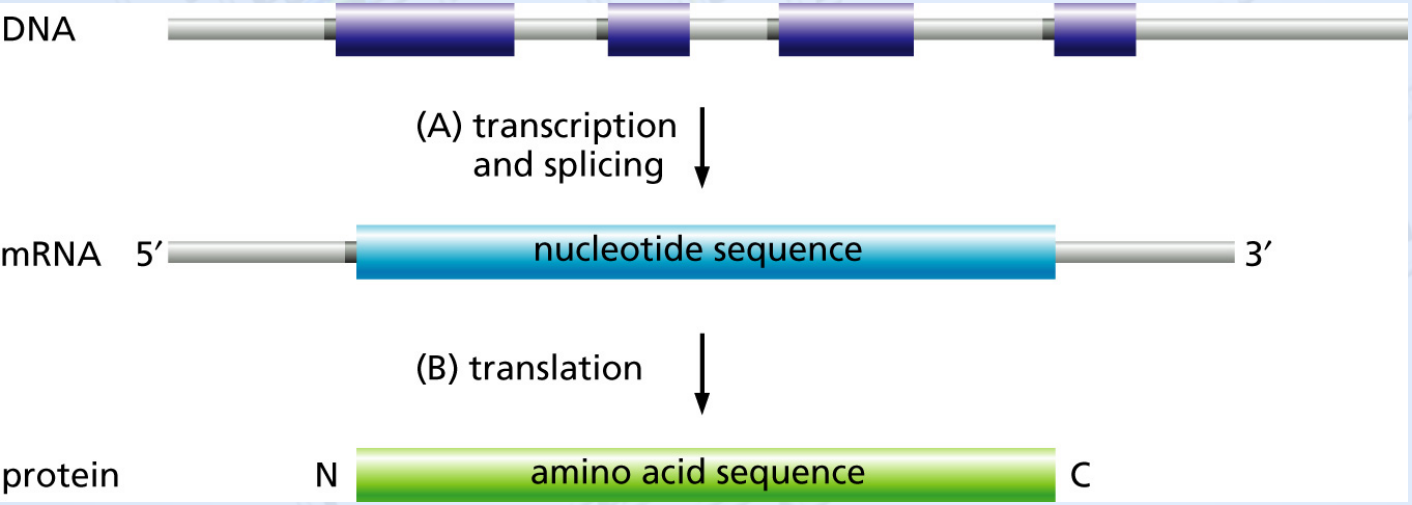
The sequence provides a specification

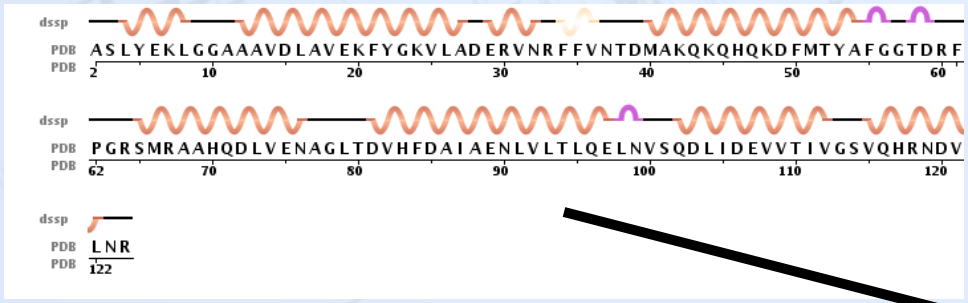
The sequence encodes biological function of the protein

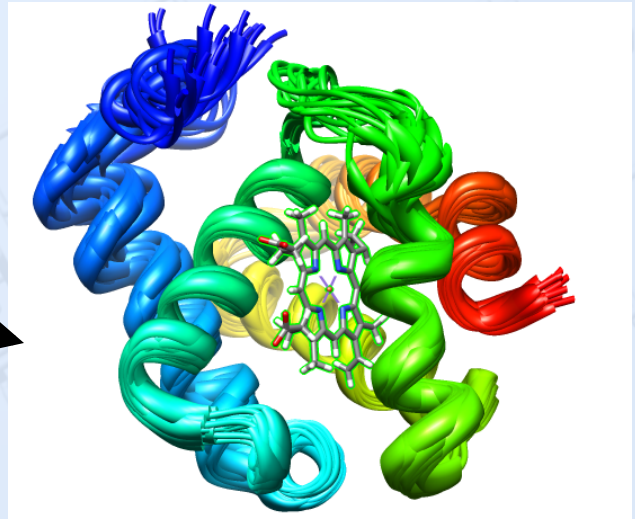
Physics & Biology: folding specificity

Unique characteristic:

The sequence of a protein determines and specifies its structure

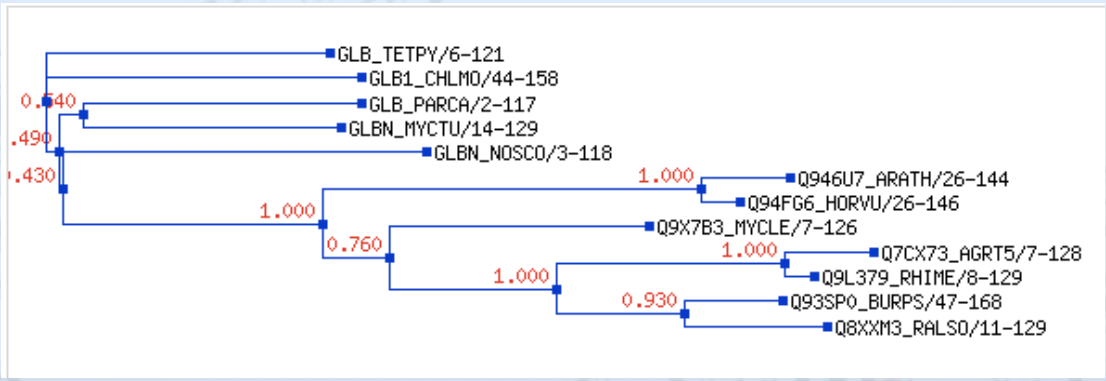


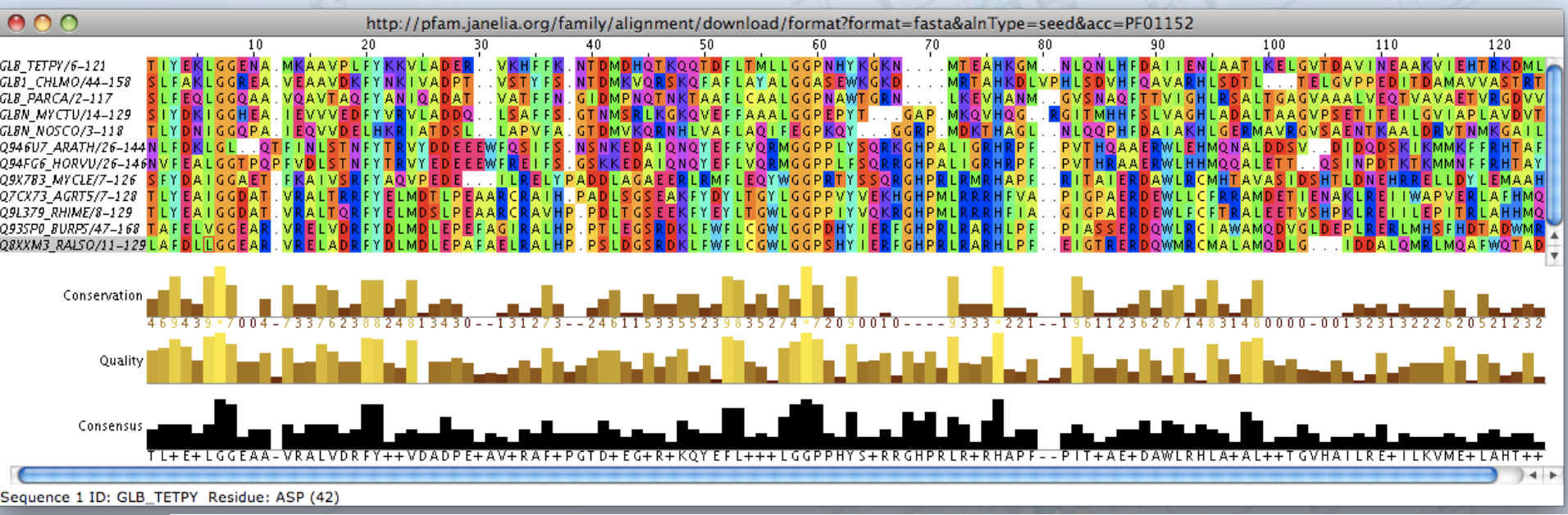


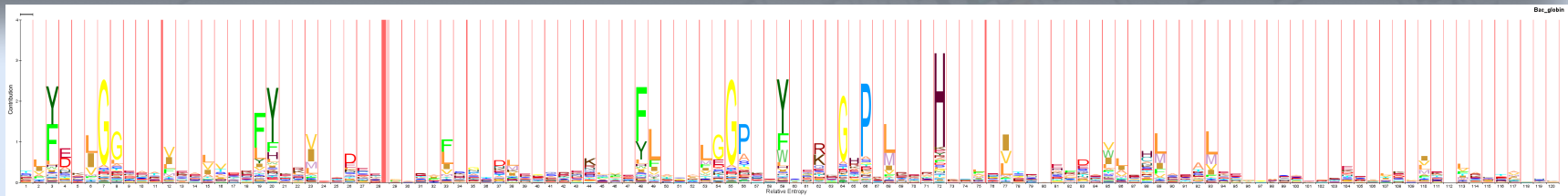


Sequences evolve!

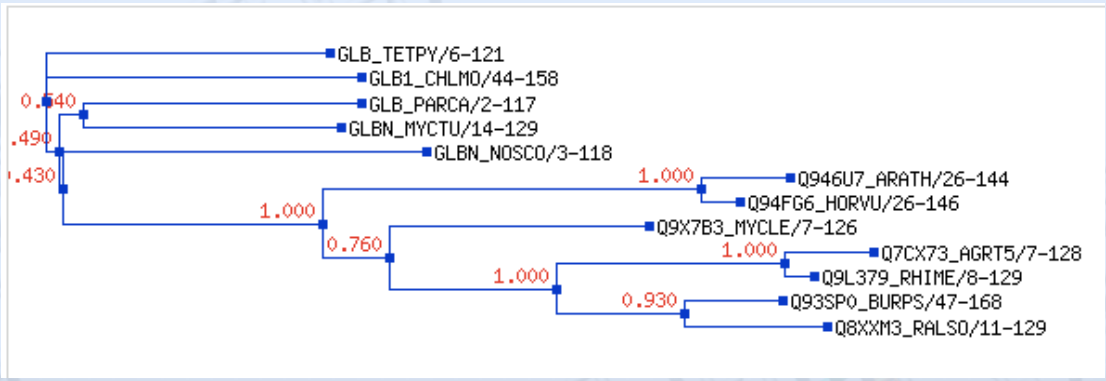
The biology: an evolutionary tree

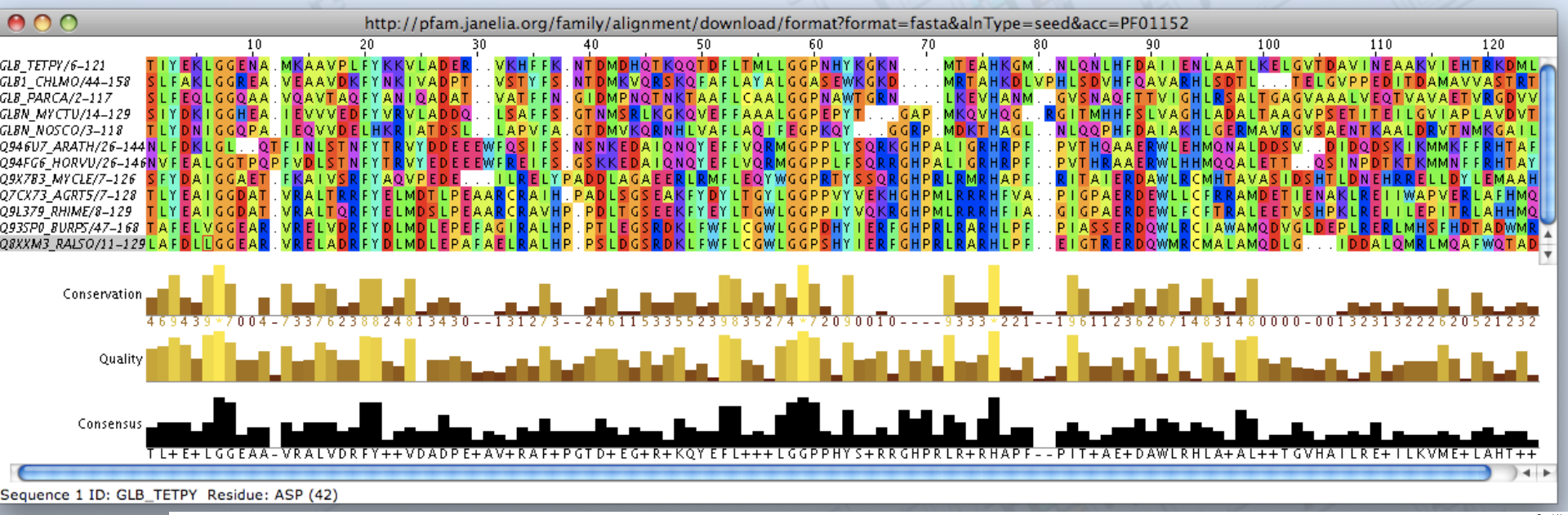


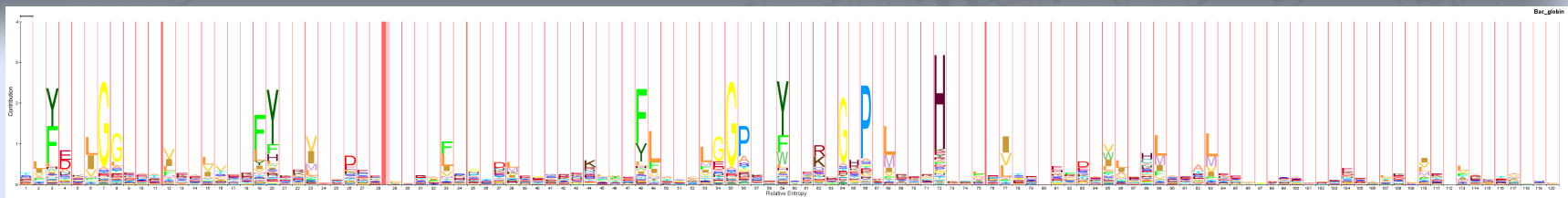




What can evolutionary history tell us?







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

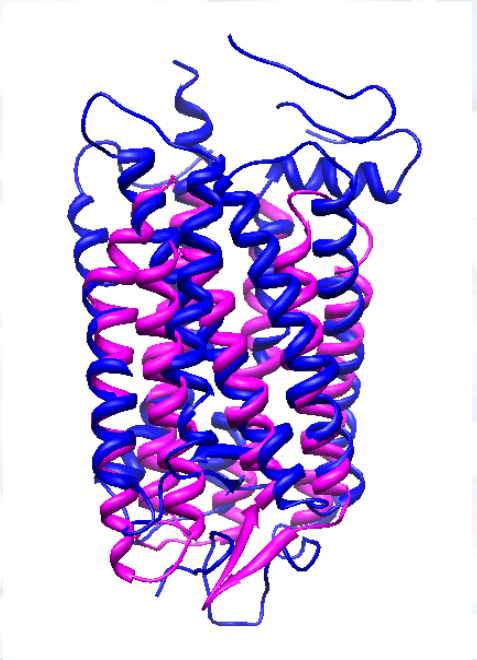
Sequence alignment based on structure alignment.

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

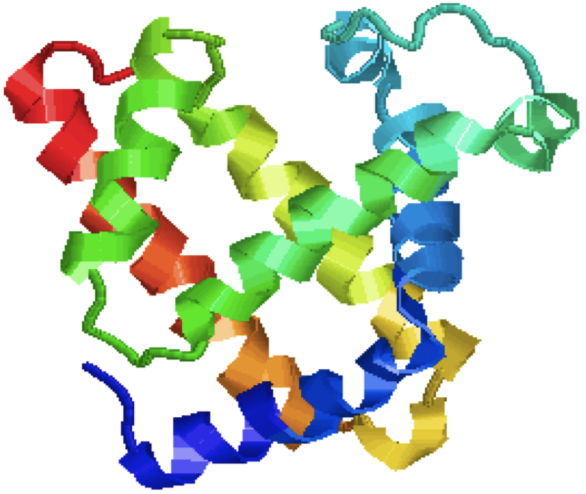
```
1L9H:A    33/33    AEPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGF
1GUE:A     3/4     GLTTLFWLGAIGMLVGTLAFAWAGRDA--GSG-----------ERRYYVTLVGISGIAA

1L9H:A    93/93    TTTLYTSLHGYFVFG--------PTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPM
1GUE:A    49/50    VAYVVMALG---VGWVPVAERTVFAPRYIDWILTTPLIVYFLGLLAG------------

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-LWAIYPFIWLLGPPGVALLTPTVDVALIVYLDLVTKVGFGFIALDAAATL
```



So if we have a protein with a known structure that has a similar sequence – we have solved our problem.

The biological approach or the physical approach?

?

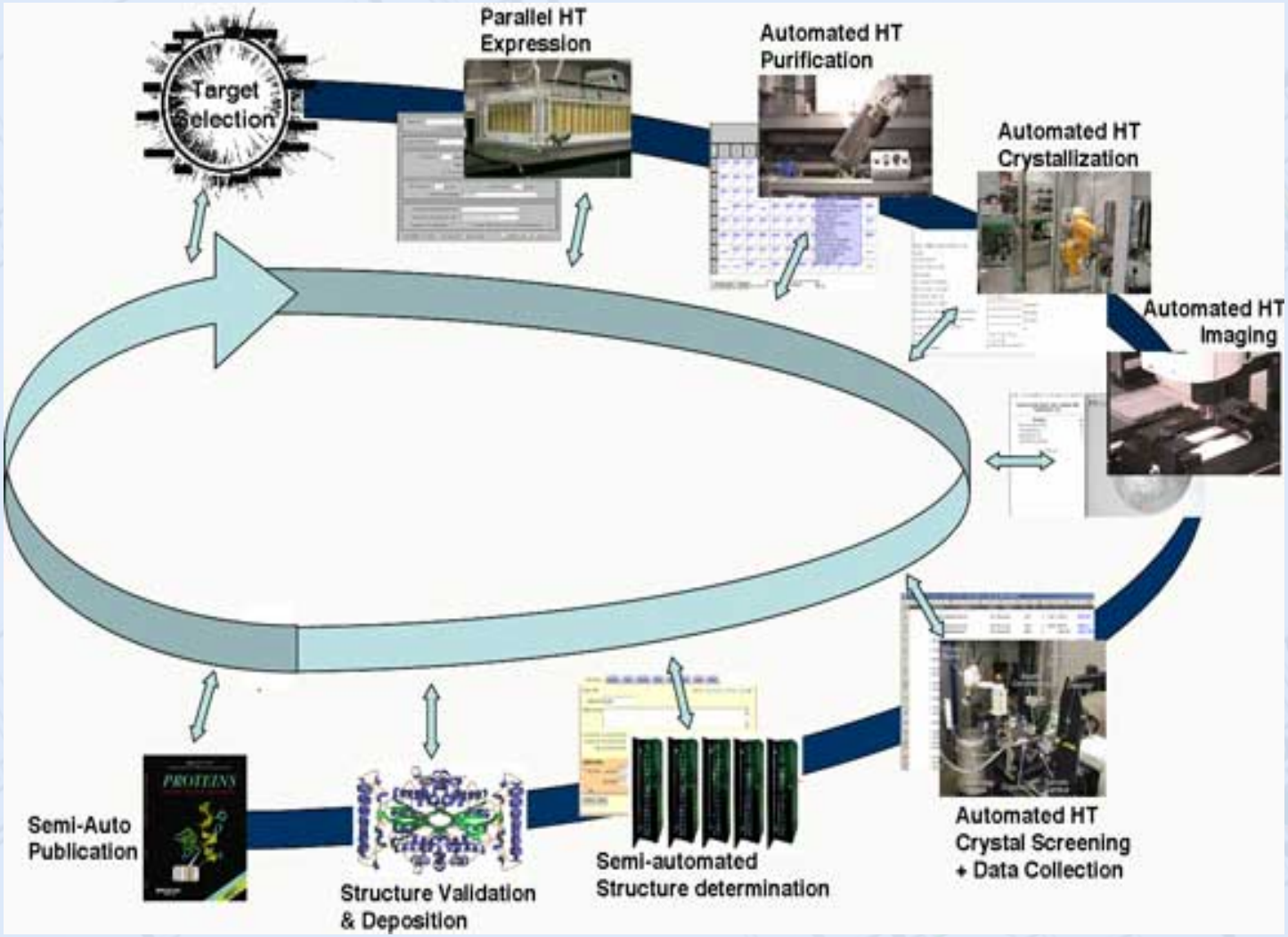


TASK:

Given a protein sequence predict in which 3D structure the protein will fold.

Who wins, the physical or biological approach?

Structural Genomics



CASP

Critical Assessment of protein Structure Prediction

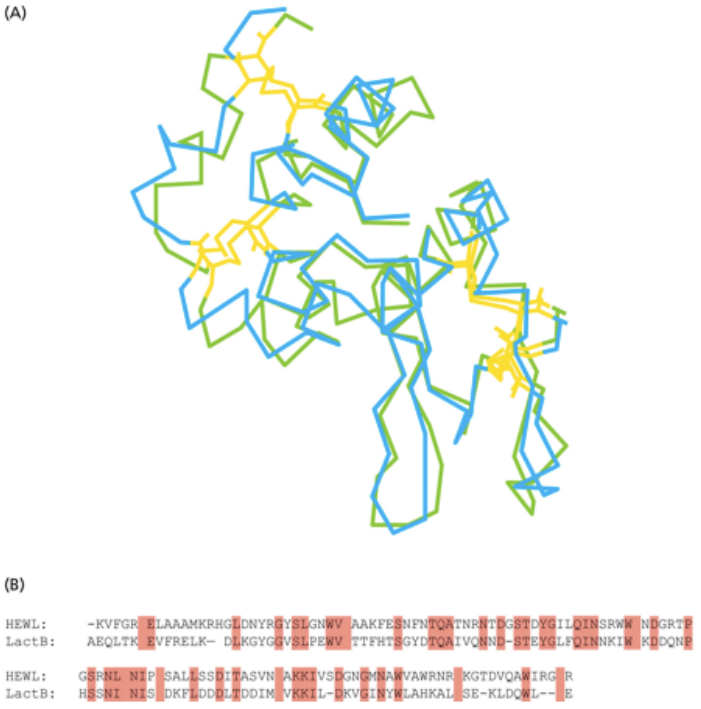
AIM: assess progress in structure prediction

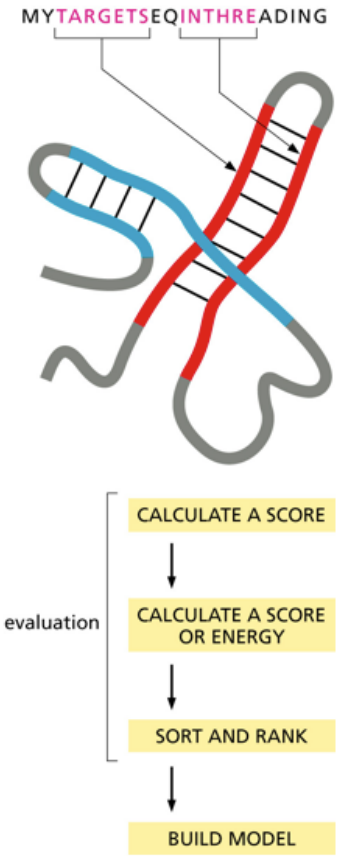
- Blind test
- Ongoing for three months
- Experimentalist submit targets
- Bioinformatics groups predict structures
  - groups

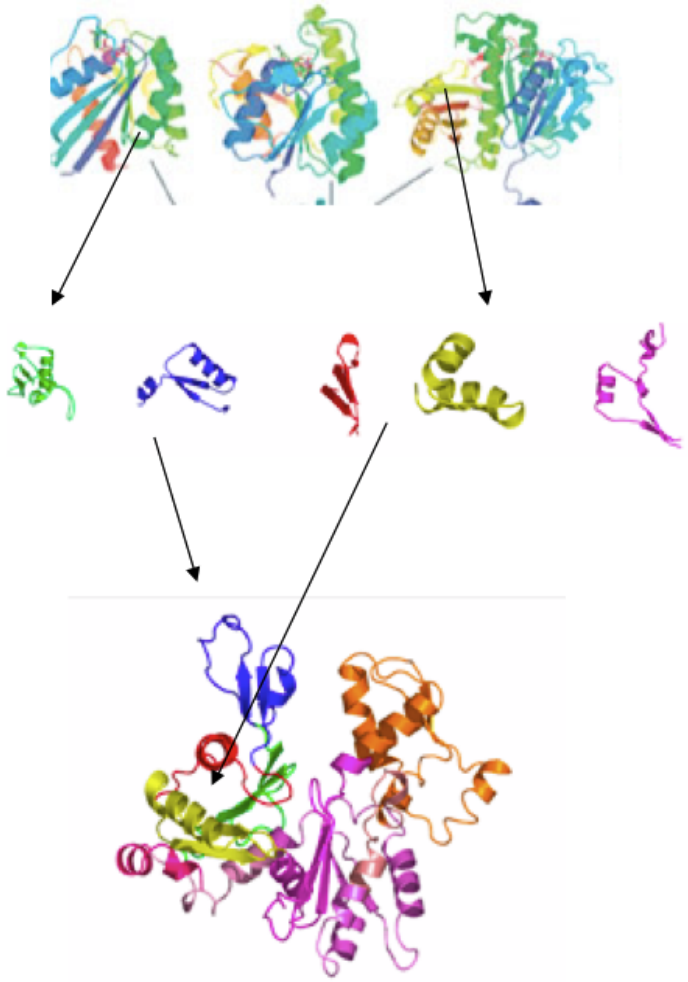
Who wins BIOLOGY or PHYSICS?

Structure Prediction

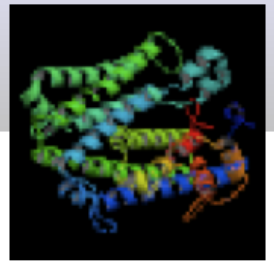
Homology – Fold Recognition - “Ab initio”

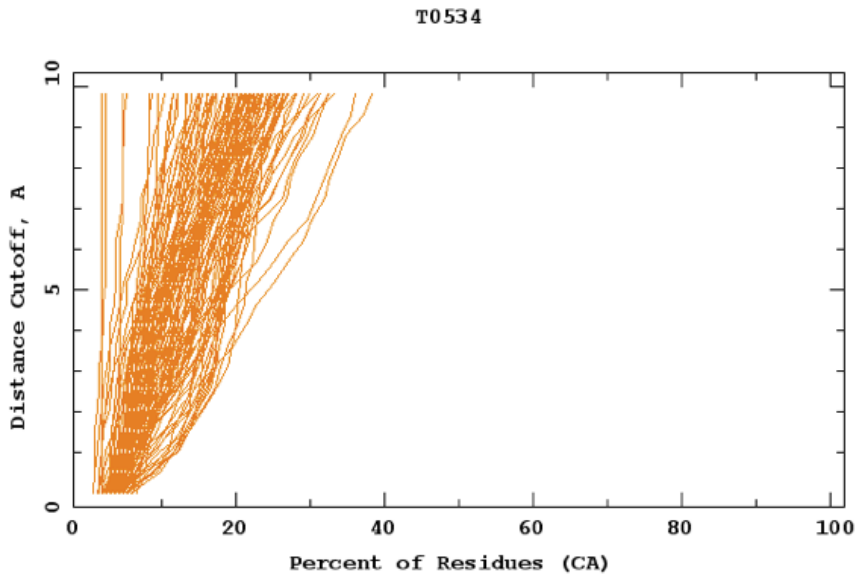


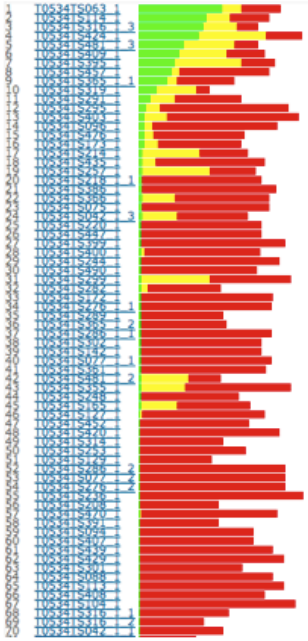




“difficult“ target







GREEN Correctly aligned residues (0 shift) according to AL0_P

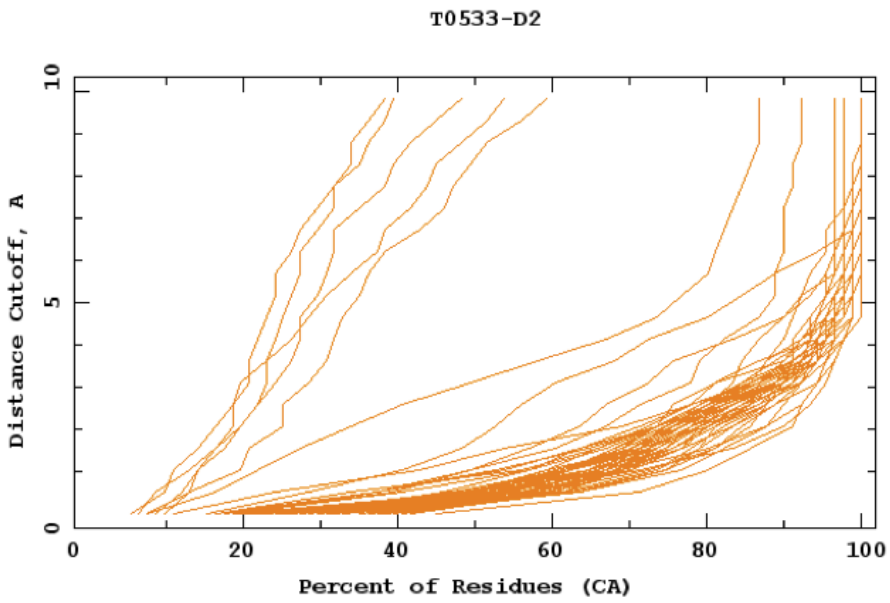
YELLOW Residues aligned within " -4 , +4 " window (4 shift) according to AL4_P

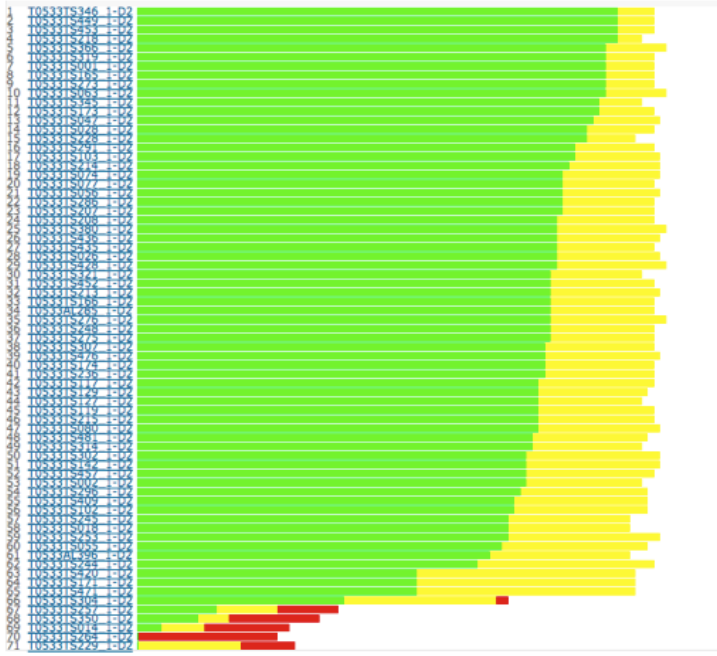
RED Residues aligned outside " -4 , +4 " window (4+ shift) according to ALI_P

WHITE Residues not aligned or not predicted

RMSD calculated on all N residues superimposed under 5.0 Angstrom distance cutoff

“easy” target





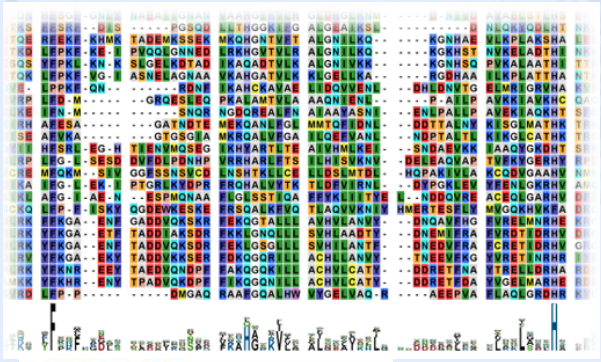
GREEN	Correctly aligned residues (0 shift) according to AL0_P
YELLOW	Residues aligned within " -4 , +4 " window (4 shift) according to AL4_P
RED	Residues aligned outside " -4 , +4 " window (4+ shift) according to ALI_P
WHITE	Residues not aligned or not predicted

RMSD calculated on all N residues superimposed under 5.0 Angstrom distance cutoff

What kind of biological data is available?

Structures

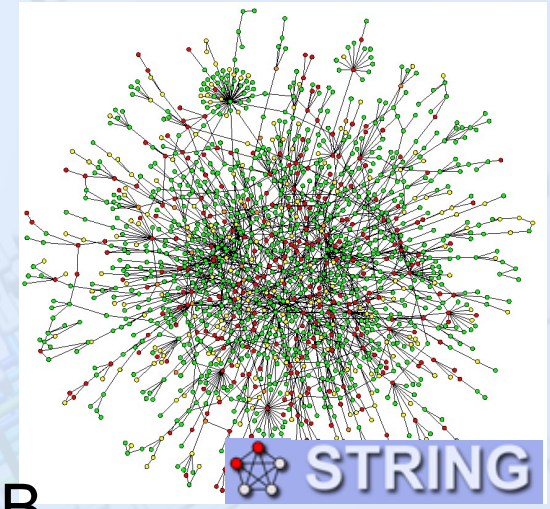
Sequences



Pfam

NCBI

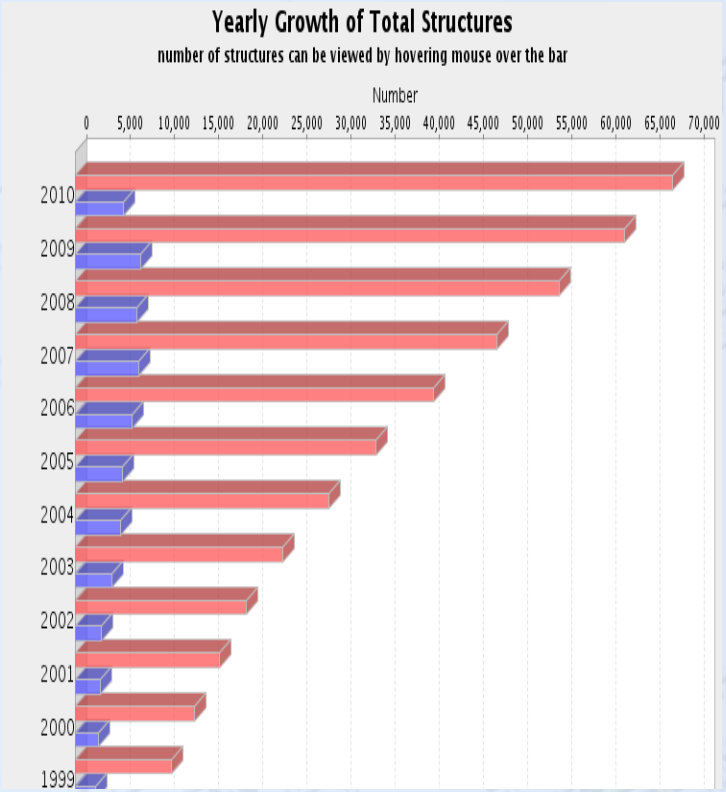
Function



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

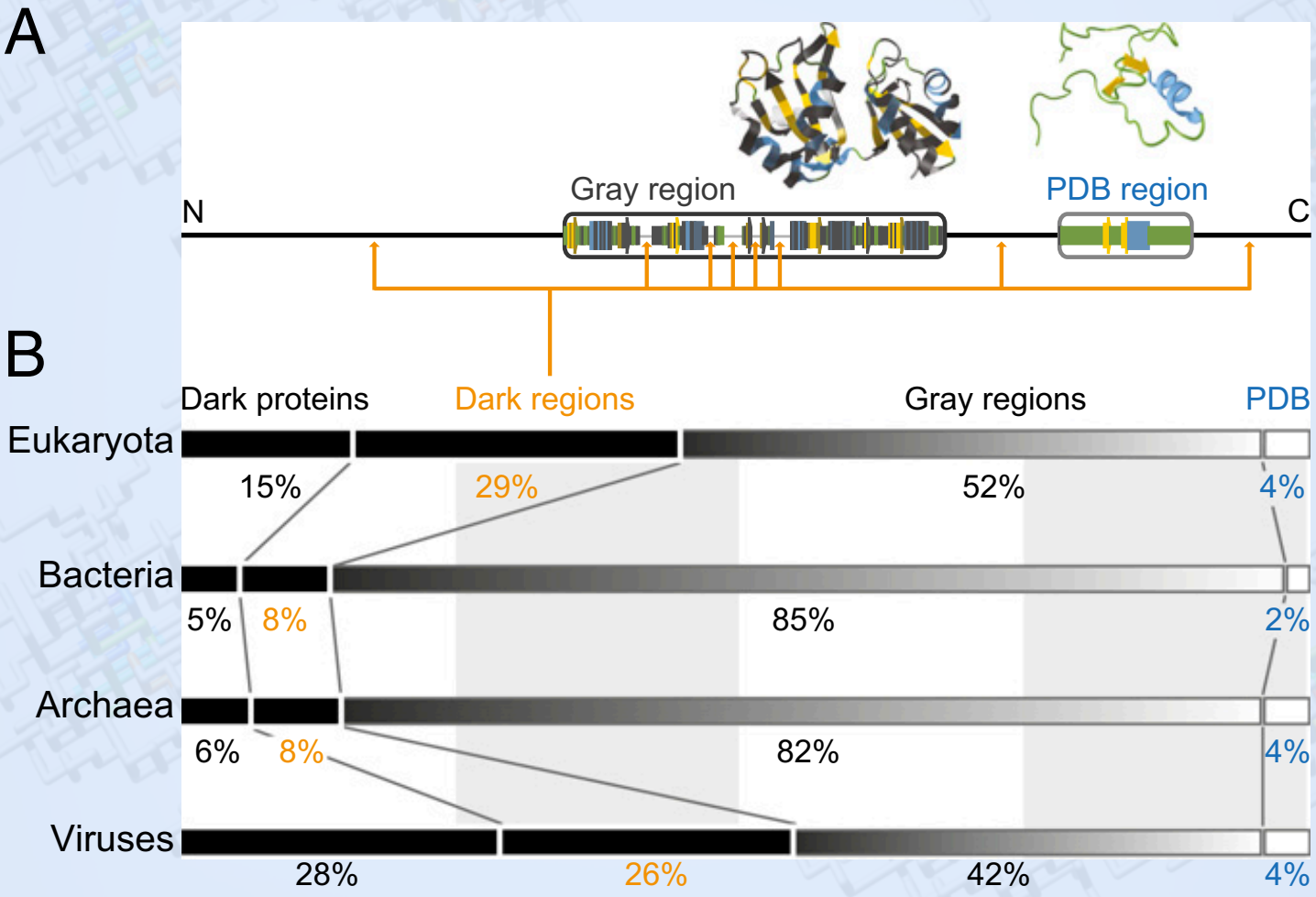
PDB

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

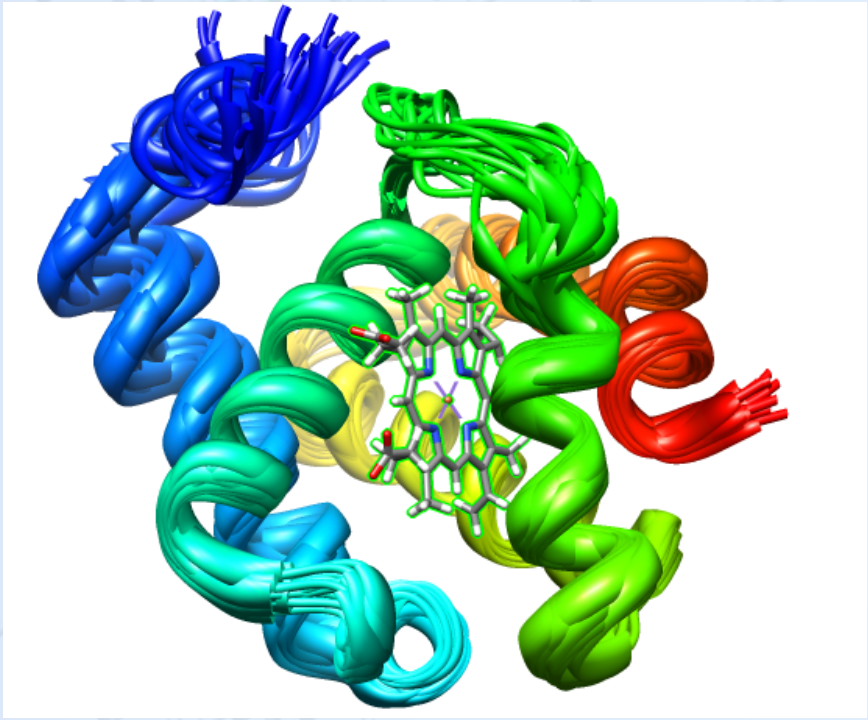


no filter		**65.000**
blast	**100% identity**	**38.974**
blast	**95% identity**	**27.892**
blast	**90% identity**	**26.831**
blast	**70% identity**	**24.204**
blast	**50% identity**	**21.249**
blast	**40% identity**	**19.002**
blast	**30% identity**	**16.388**

Dark proteins: no structures available with similar sequences



So is there any role for physics based approaches?

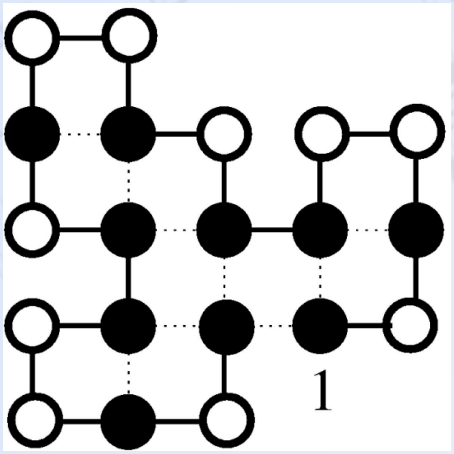


- Physical questions:
  - How stable is this protein
  - Under which conditions will this protein fold?
  - How strong is the binding to a substrate?

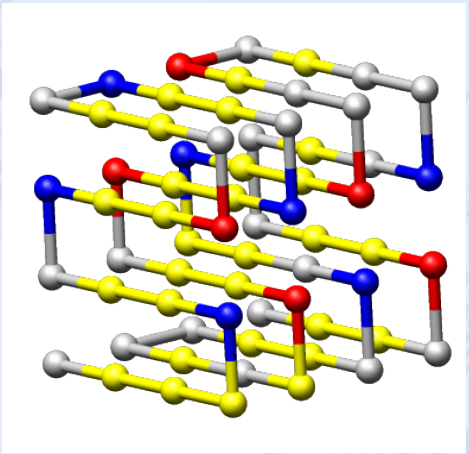
  - Fundamental understanding of mechanisms and forces involved in folding
  - Detailed simulation under (experimental) constraints.

Why use simple models?

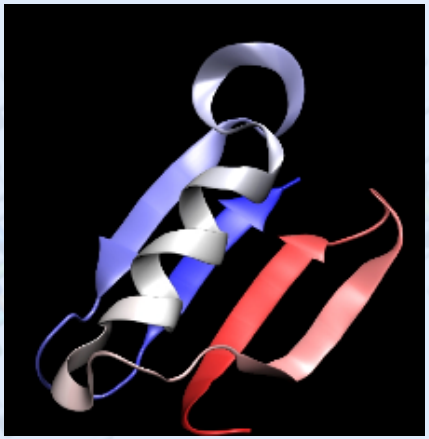
HP model - minute



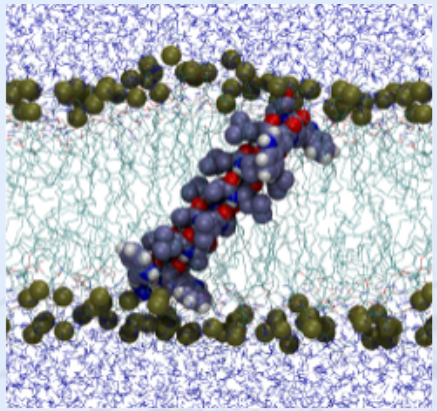
cubic lattice model - hour



backbone model - week



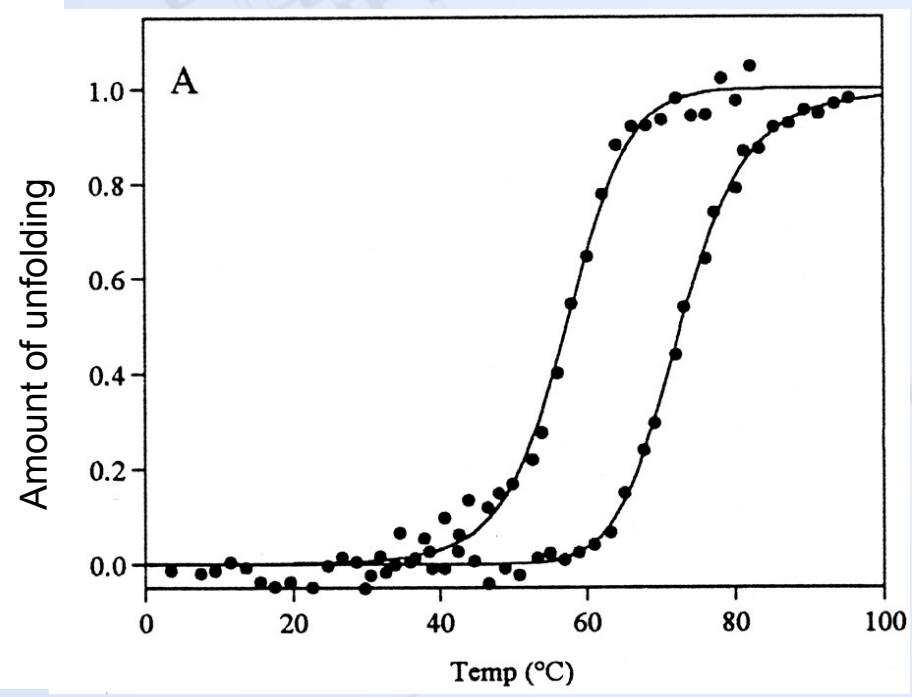
full atomistic model - year(s)

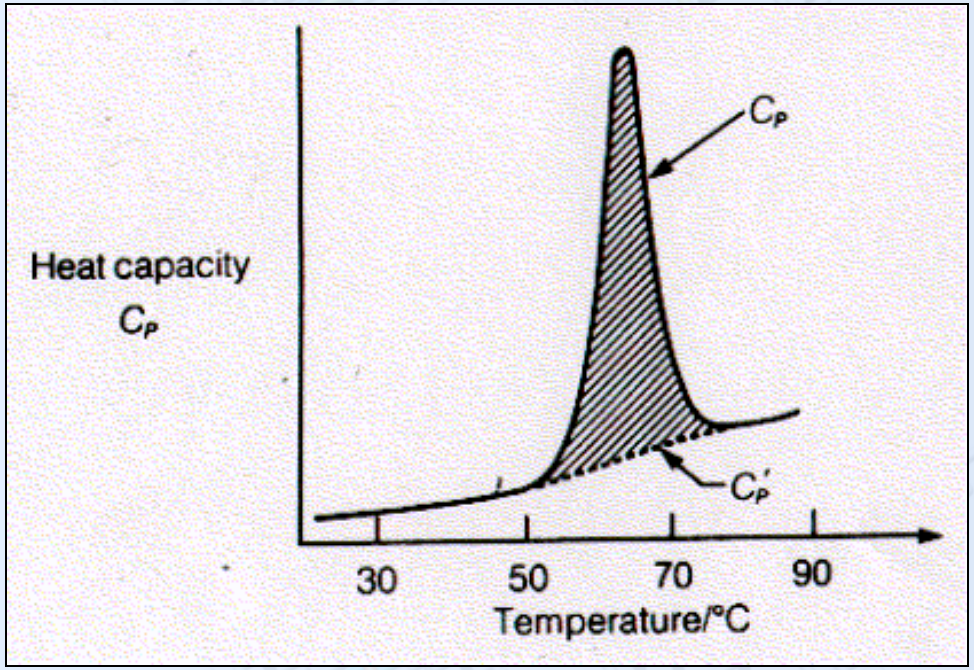


- Sampling lowest free energy state
- Different conditions
- Larger systems

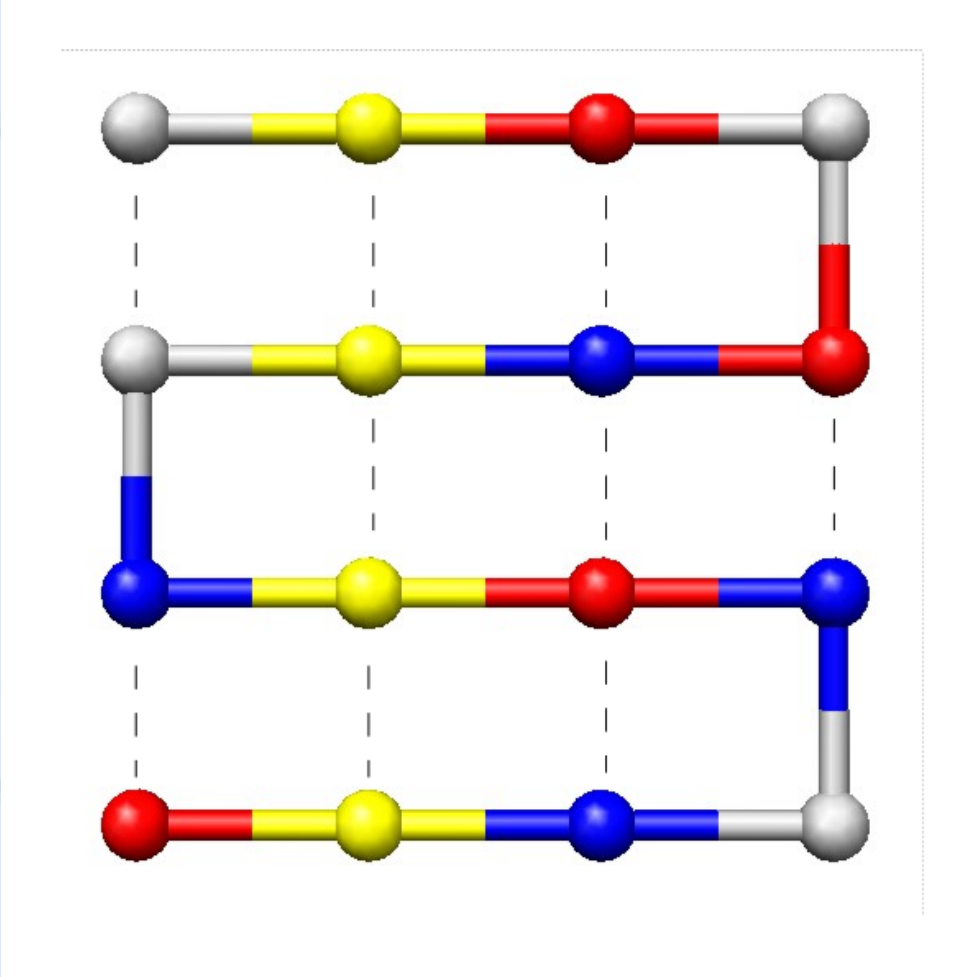
Physics: folding specificity - perfect self assembly

Experimental curves – can we understand these?





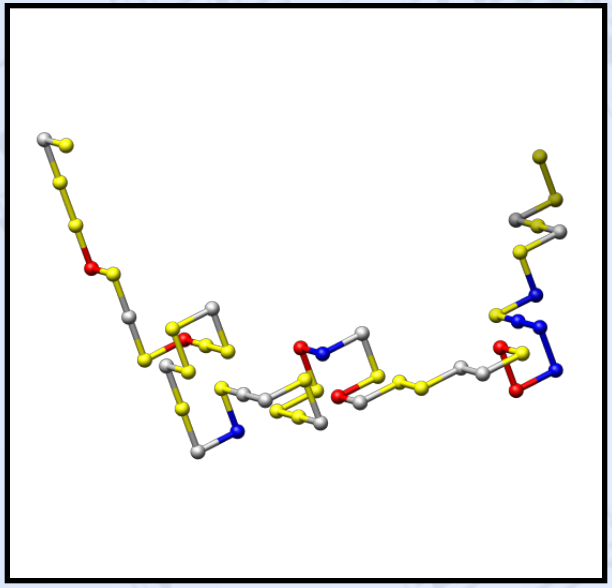
A very simple model



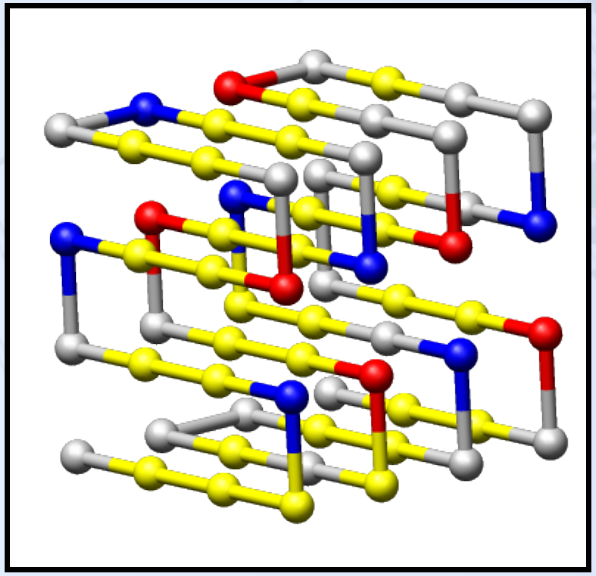
- 3D for research
- 2D in practical

Lattice Model

Sequence



Structure



hydrophobic

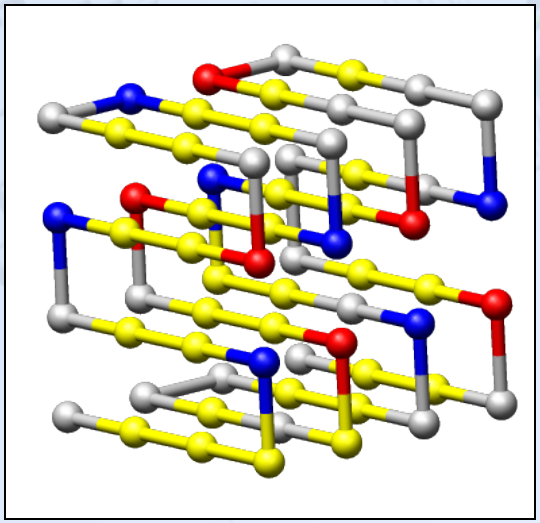
polar (hydrophilic)

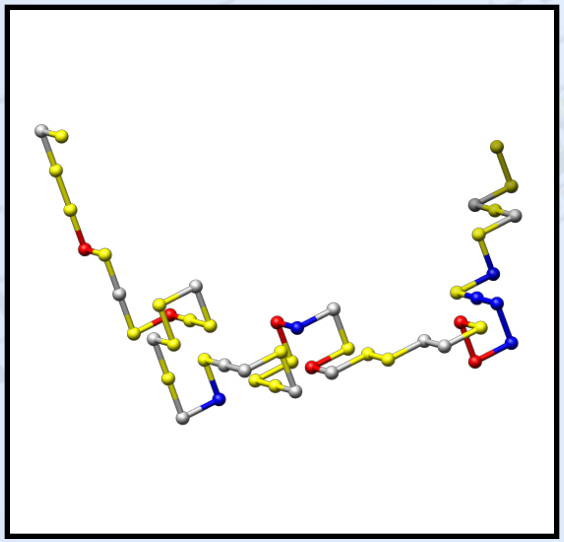
negative charge

positive charge

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

Cubic Lattice Model





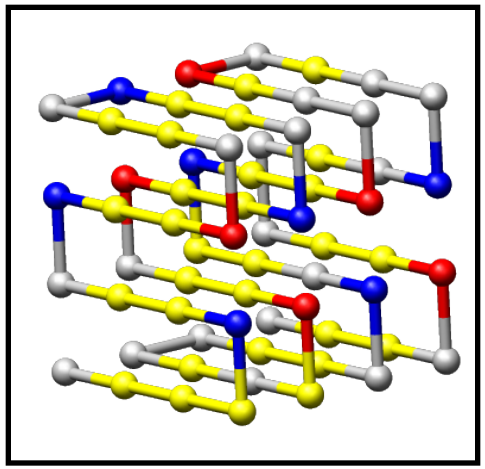
- Cheap & simple
  - Use for right purpose
- Can model:
  - General trends
  - Folding specificity
  - Heat capacity
  - Binding and unbinding
- Not captured:
  - Secondary structure
  - Hydrophobic effect (cold denaturation)
  - Specific proteins

Shakhnovich & Gutin 1993 PNAS 90

Coluzza et al 2003 Phys Rev E 68

Lattice Model, Potential, Design & Simulation

Sequence Design



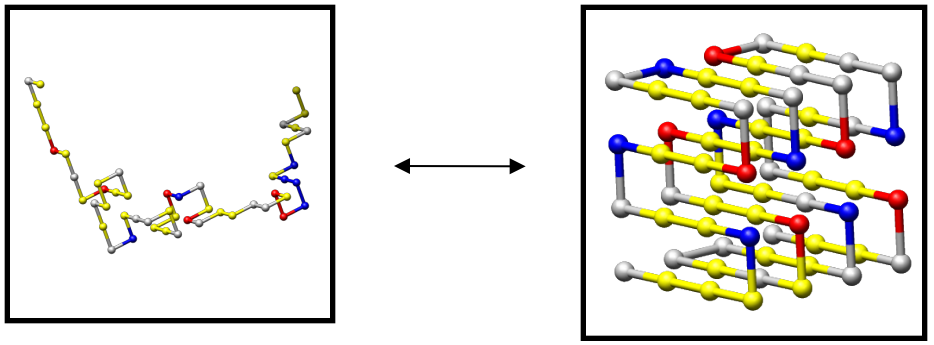
hydrophobic

polar (hydrophilic)

negative charge

positive charge

Folding Simulation



Interaction Potential

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

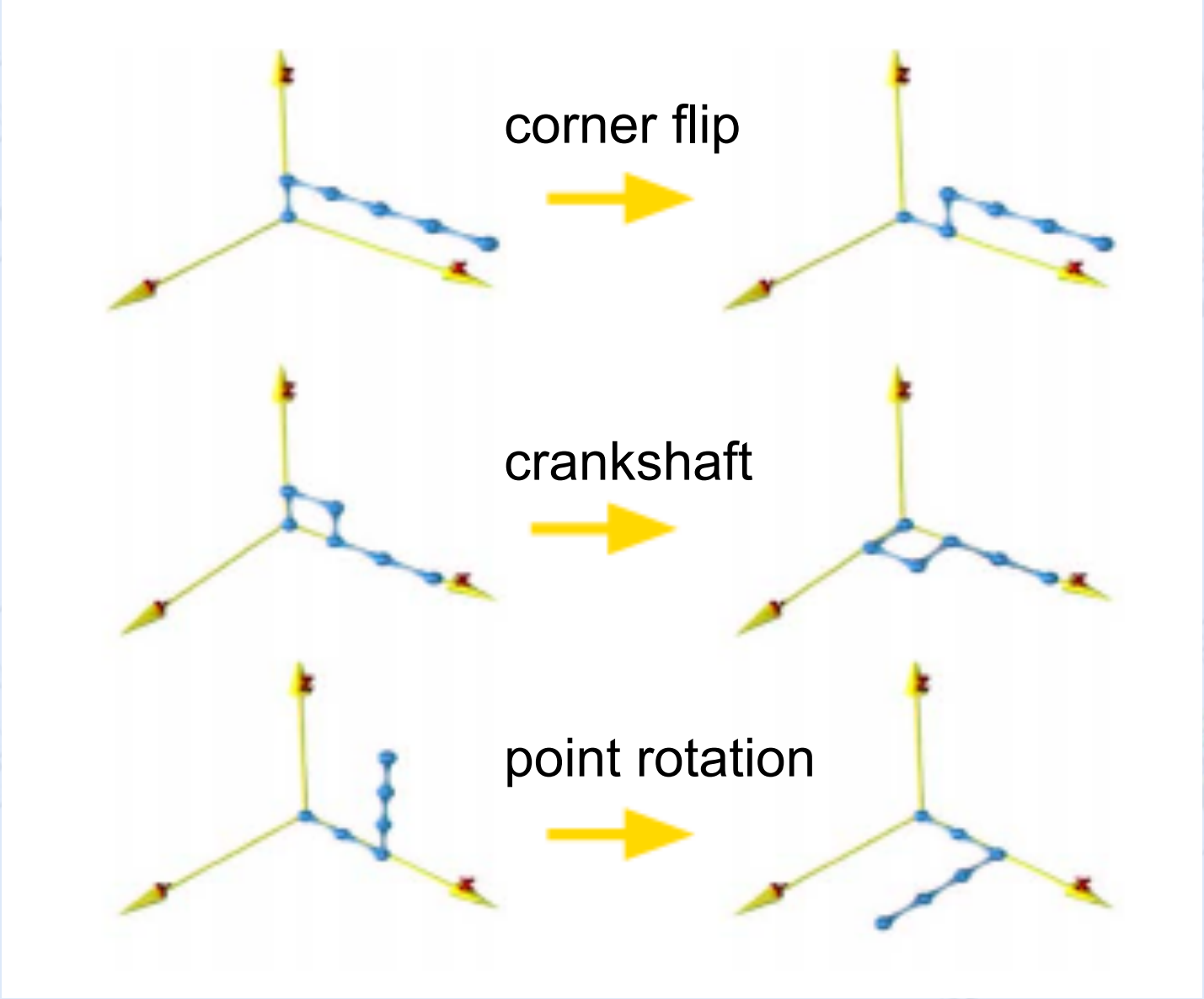
hydrophobic

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

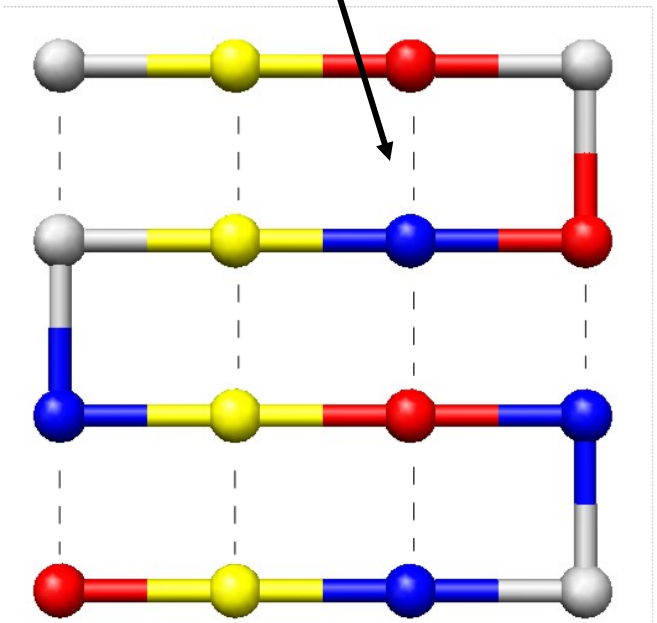
hydrophobic

polar (hydrophilic)

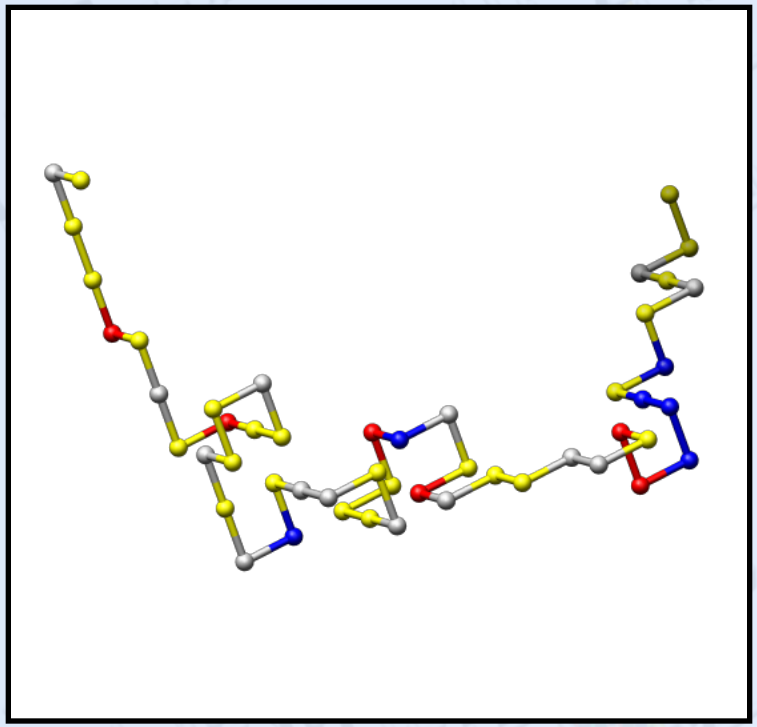
negative charge

positive charge

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



Simulation: Monte Carlo



Monte Carlo:

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

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

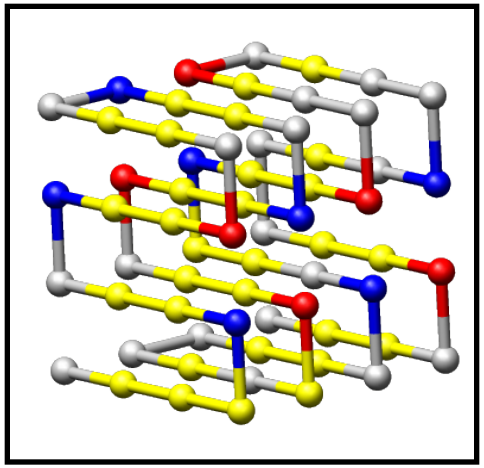
Shakhnovich & Gutin 1993 PNAS 90

Coluzza et al 2003 Phys Rev E 68

Betancourt & Thirumalai 1999 Protein Sci 8

Sequence Design

Design



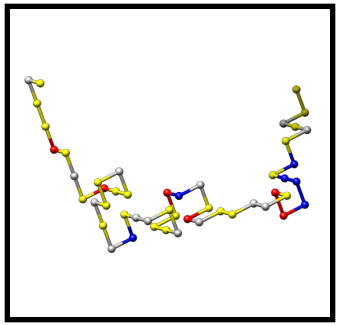
hydrophobic

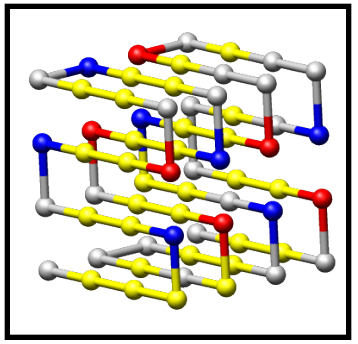
polar (hydrophilic)

negative charge

positive charge

Simulation





Potential

	hydrophobic	polar	negative	positive
hydrophobic	--	+	+	+
polar	+			
negative	+		+	-
positive	+		-	+

hydrophobic

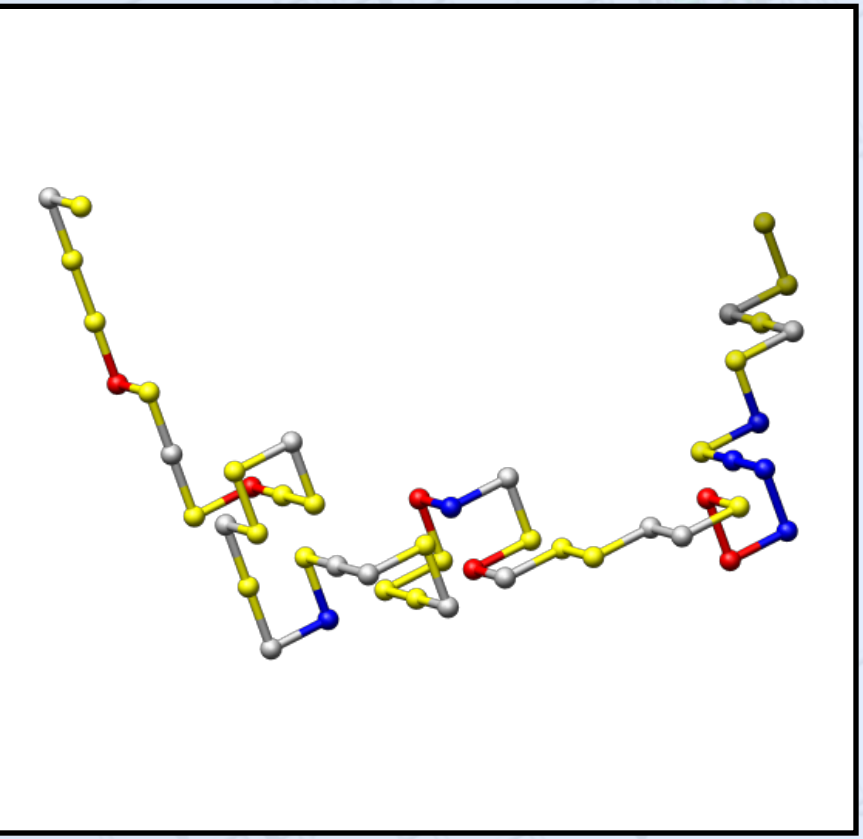
polar (hydrophilic)

negative charge

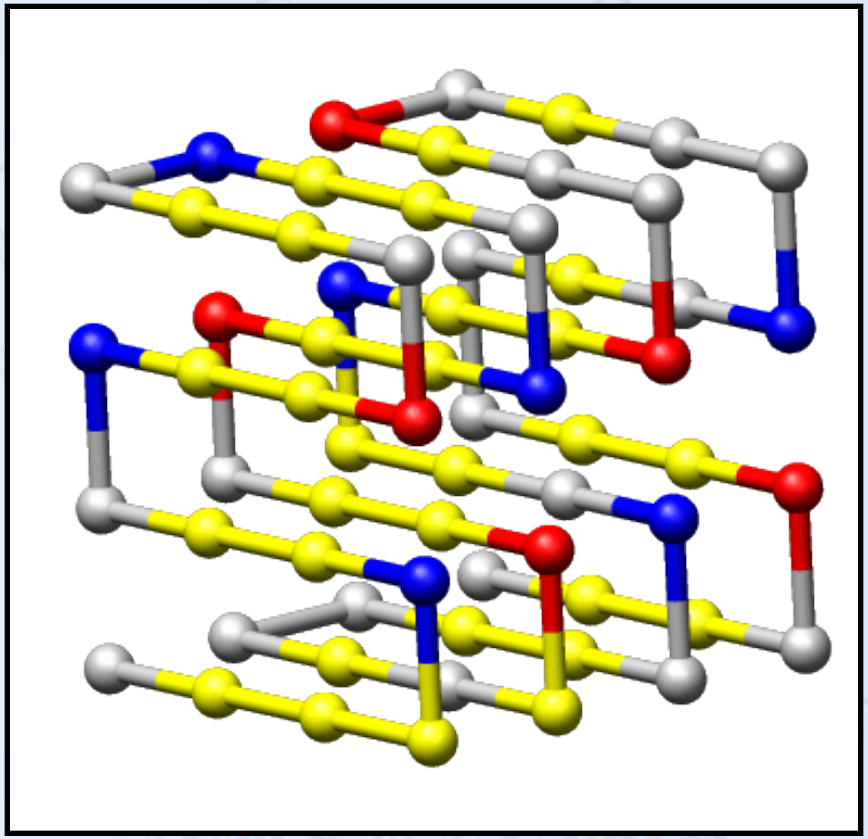
positive charge

Problem: how to create a folding sequence?

We cannot take real protein sequences (why not?)



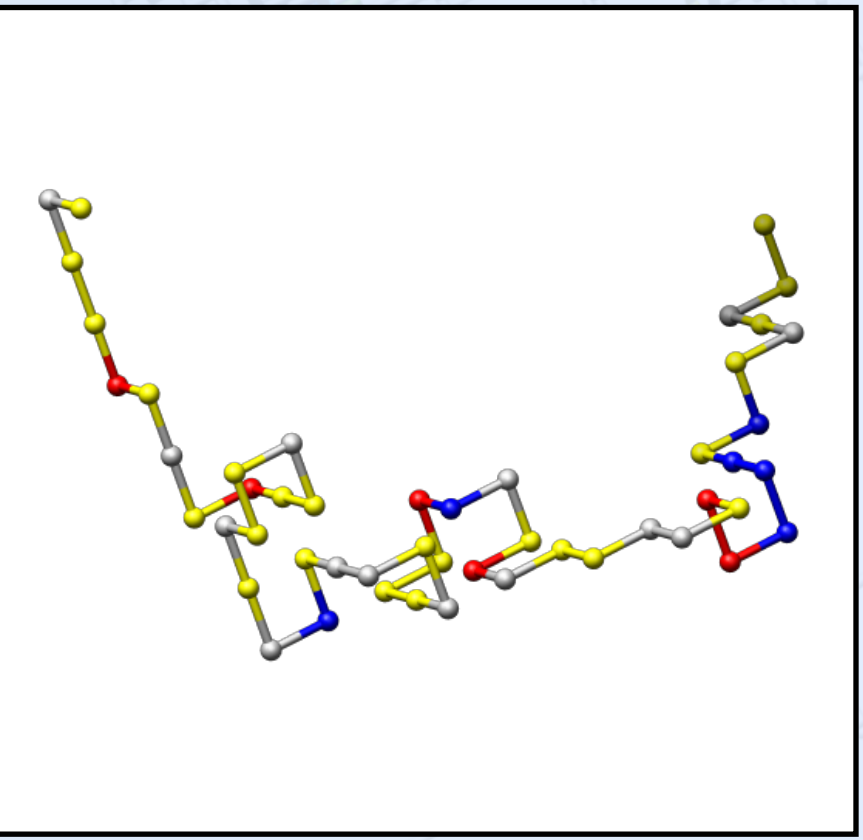
?



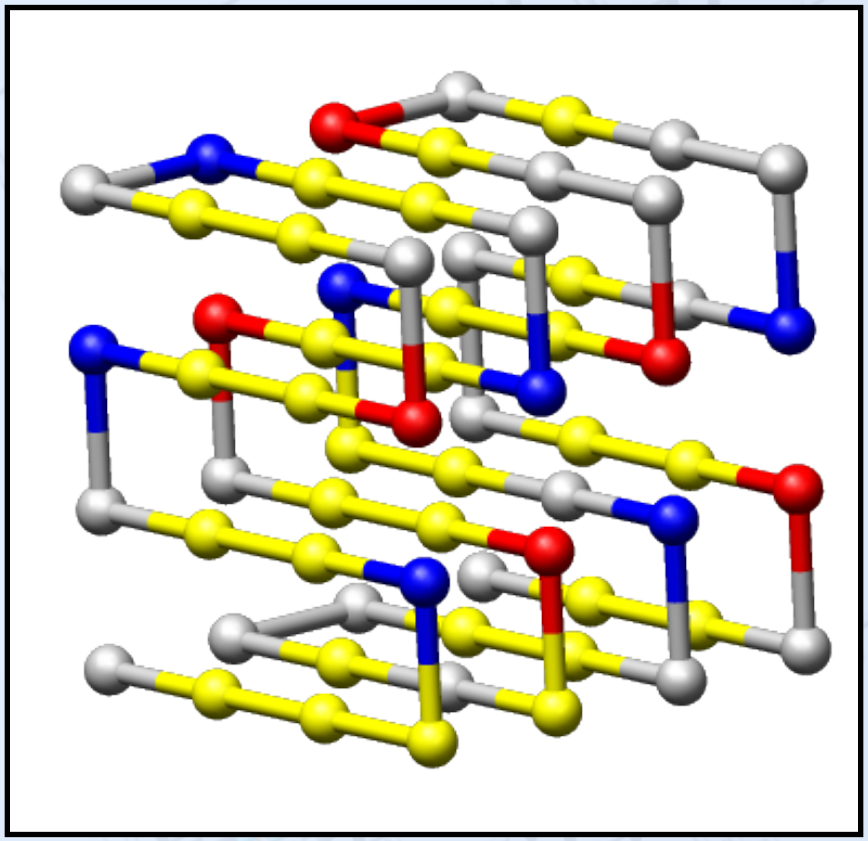
In nature evolution ensures folding...

Solution: energy minimization

Given a structure, what are sequences with a low (potential) energy?

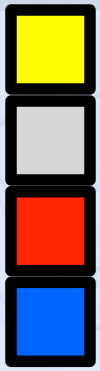


?



we can simulate evolution by changing the sequence with random substitutions

Lattice Model: design



hydrophobic

polar (hydrophilic)

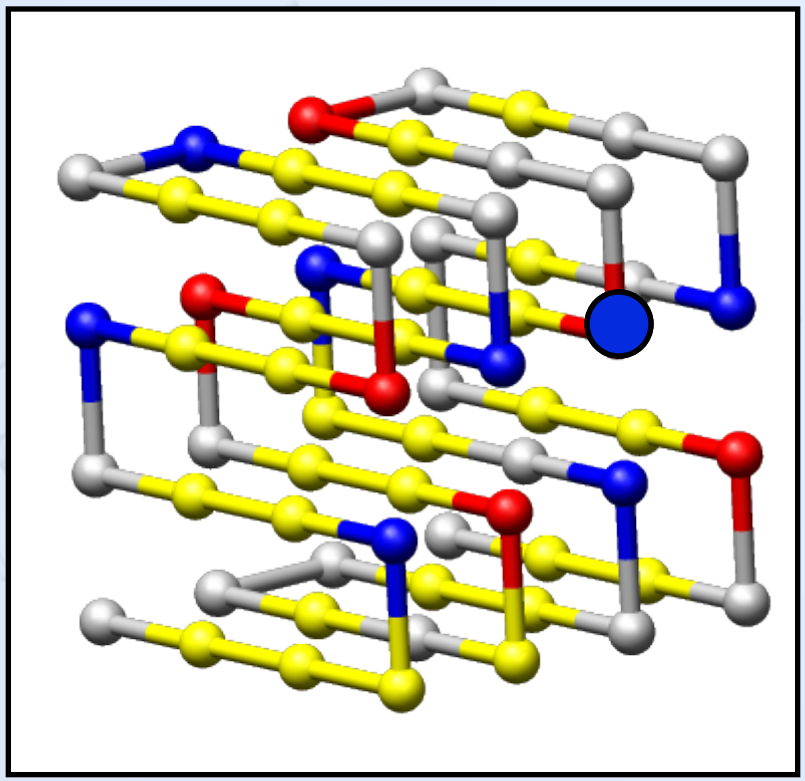
negative charge

positive charge

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

Miyazawa & Jernigan 1993 Protein Eng 6

Betancourt & Thirumalai 1999 Protein Sci 8



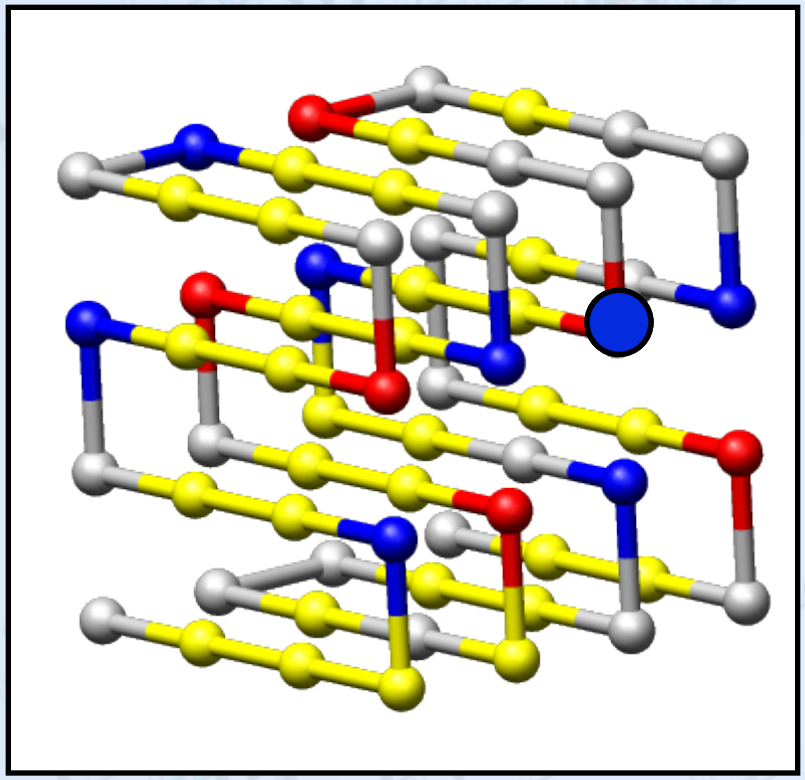
Design loop:

- 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

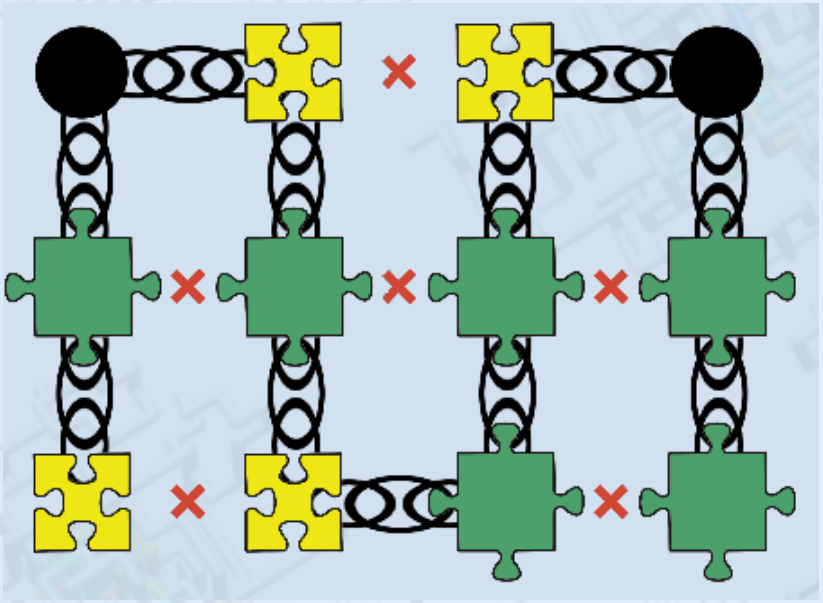
Coluzza et al 2003 Phys Rev E 68

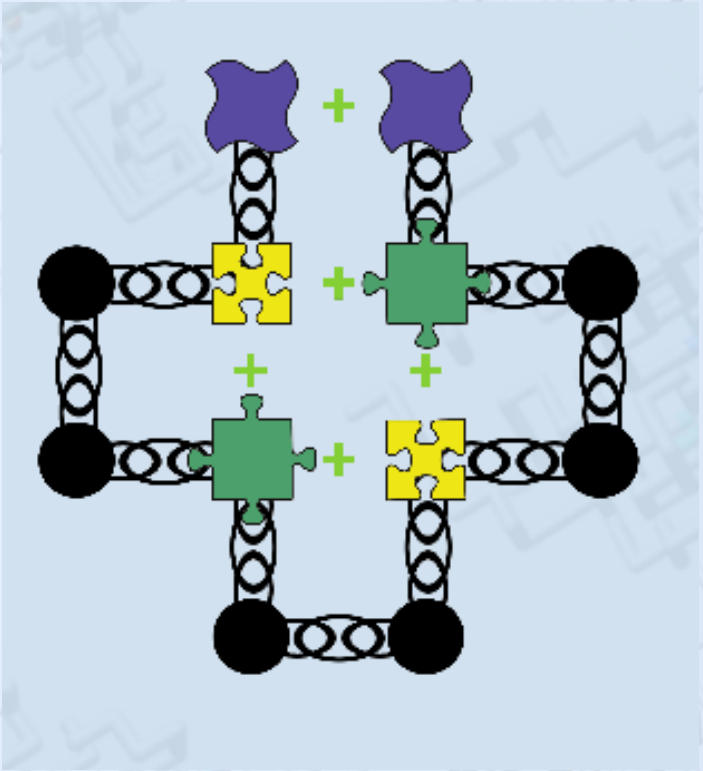
Sequence design: energy minimization



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

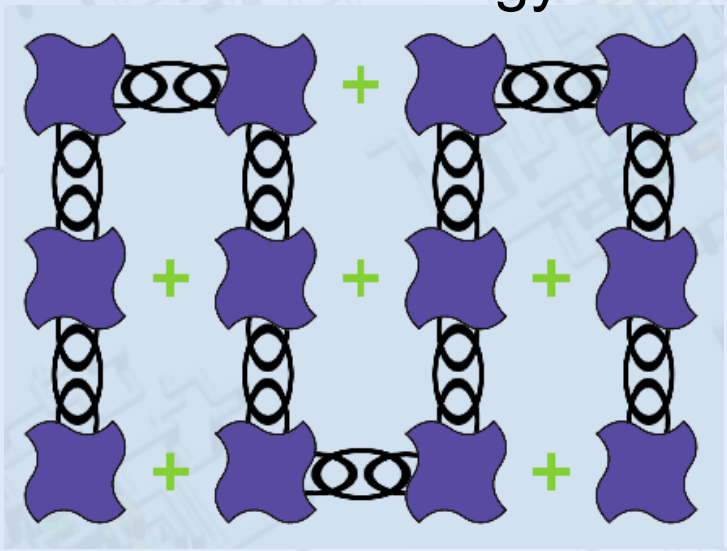
Interactions: toy example 2D



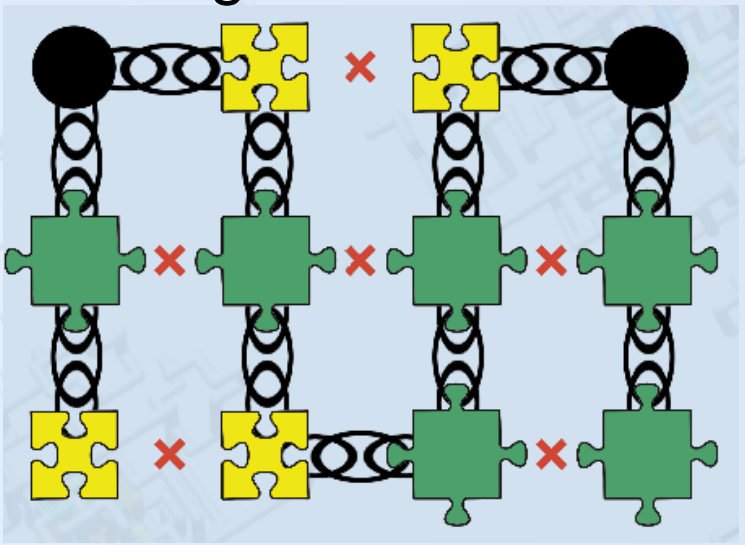


Sequence Design

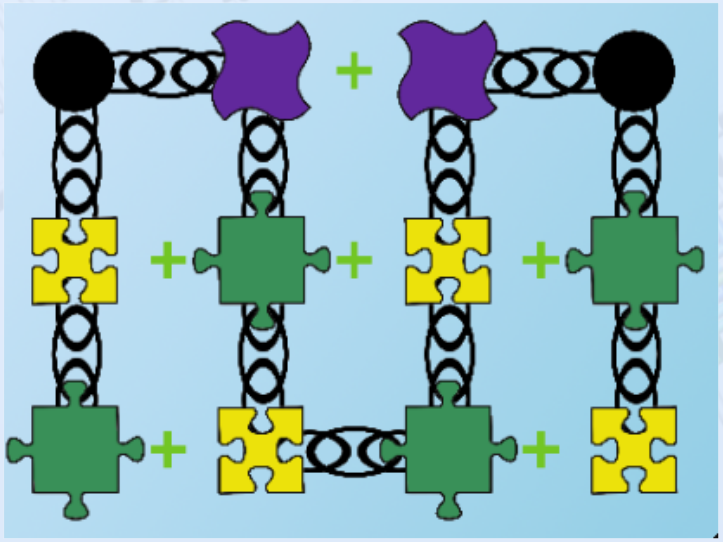
Low energy



High variance



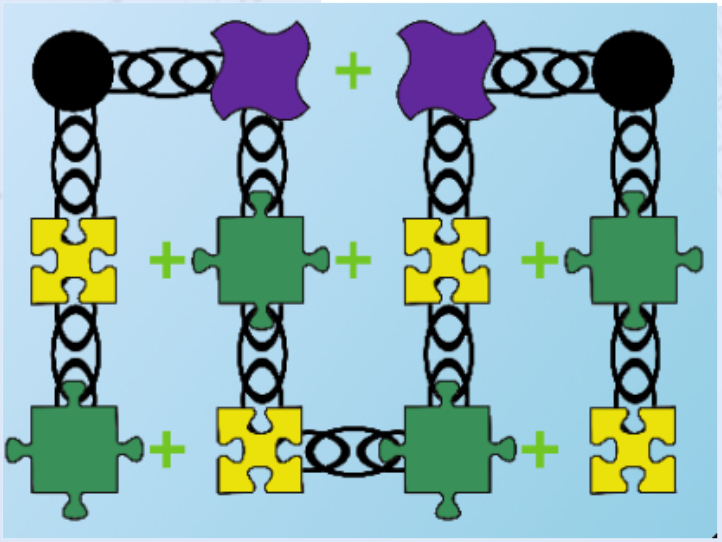
Good folder

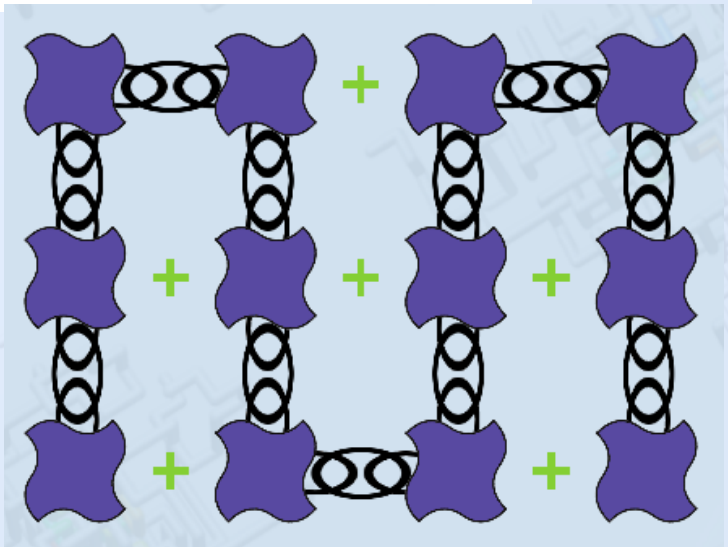


Sequence Variance

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

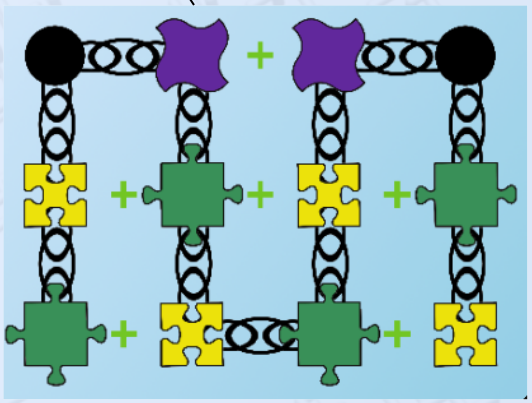
$$P_{\text{acc}} = \min\left\{1, \left(\frac{N_p^{\text{new}}}{N_p^{\text{old}}}\right)^{1/q}\right\}$$

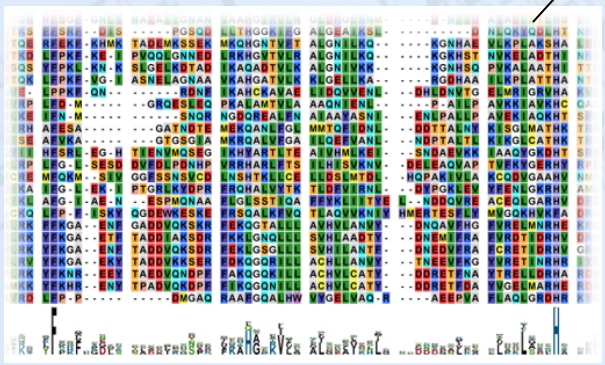




Sequence Variance & Biology

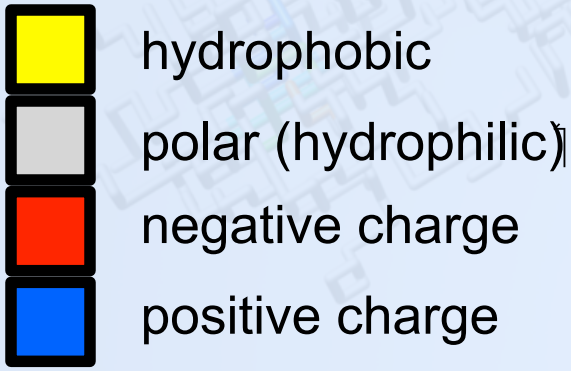
$$d = \sum_{a=0}^{a=20} (p_a - s_a)^2$$

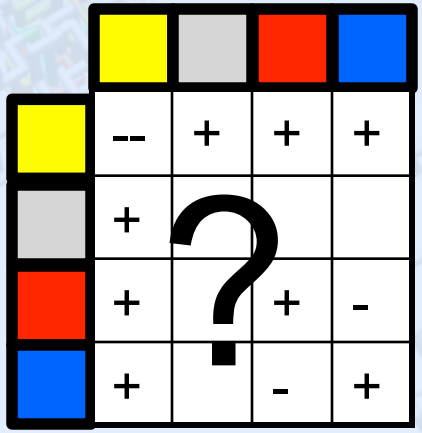


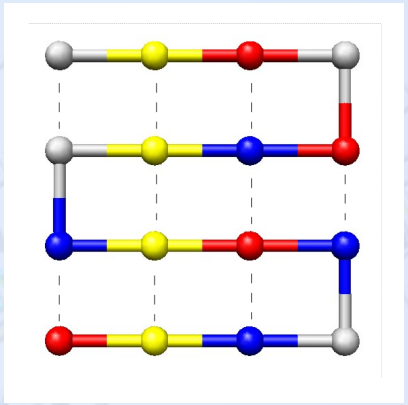


$$P_{\text{acc}} = \min\left\{1, e^{-1/q(d_{\text{new}} - d_{\text{old}})}\right\}$$

How to derive a potential?



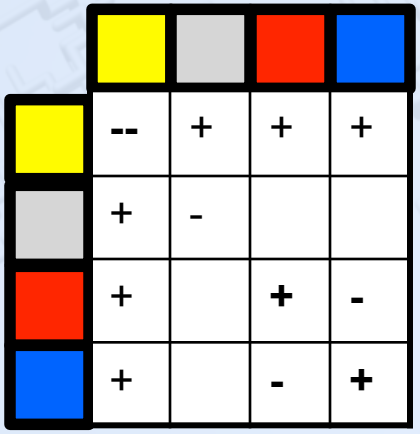




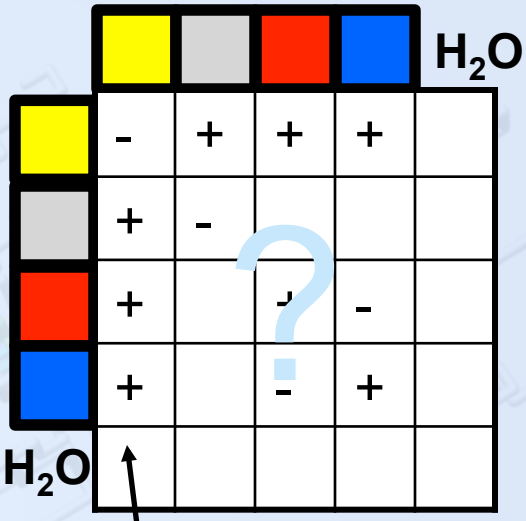
$$E = \frac{1}{2}\sum_{i}^{N}\sum_{j}^{N}\epsilon_{a(i),a(j)}C_{ij}$$

Can we use experimental biological data?

“Knowledge Based” Amino Acid Pair Potentials



hydrophobic
polar (hydrophilic)
negative charge
positive charge

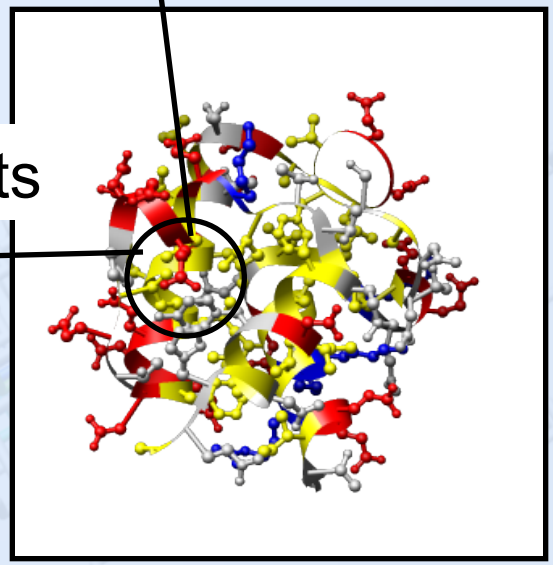


Miyazawa & Jernigan 1993 Protein Eng 6
Betancourt & Thirumalai 1999 Protein Sci 8

observed contacts

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

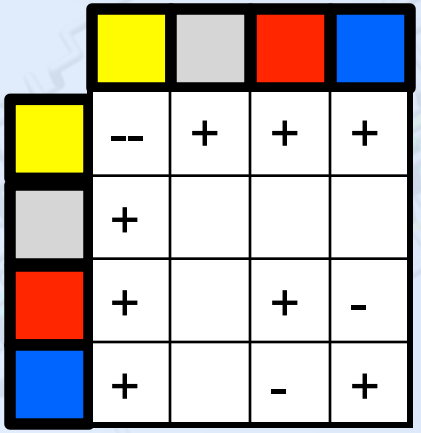
expected contacts

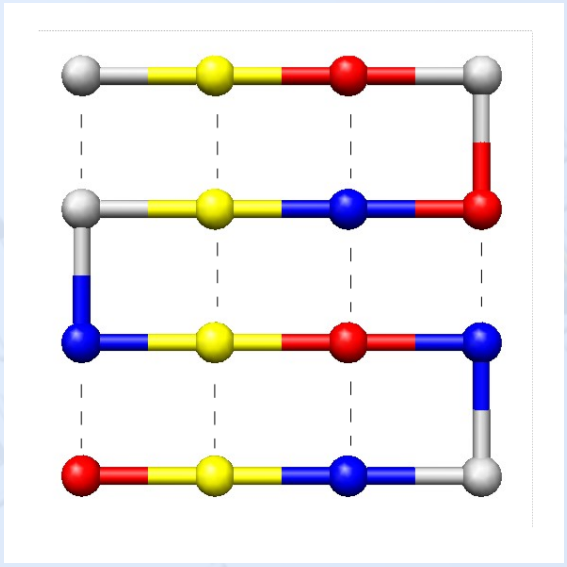


“Knowledge Based” Amino Acid Pair Potentials

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

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



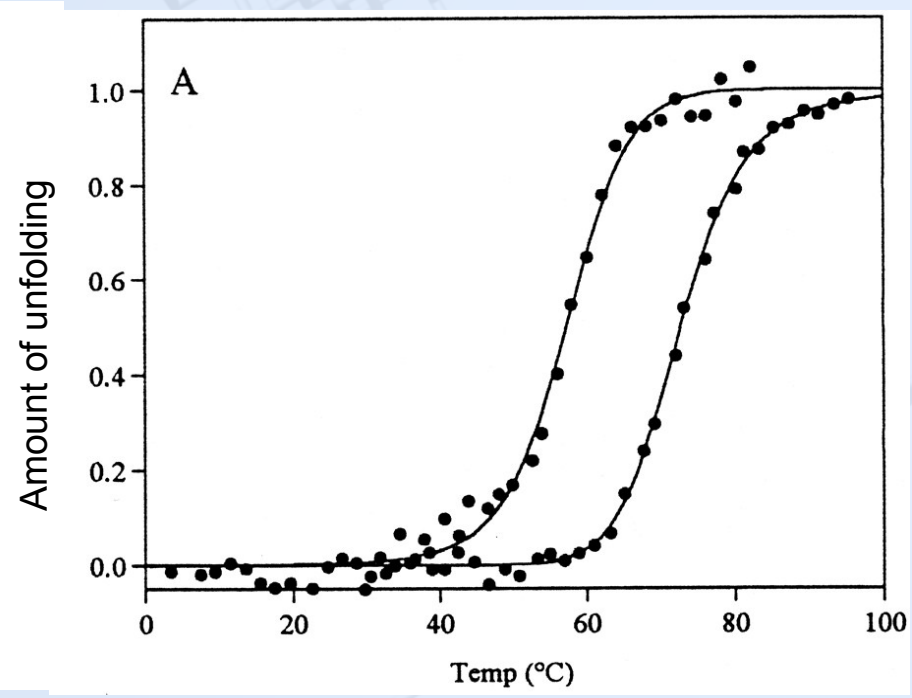


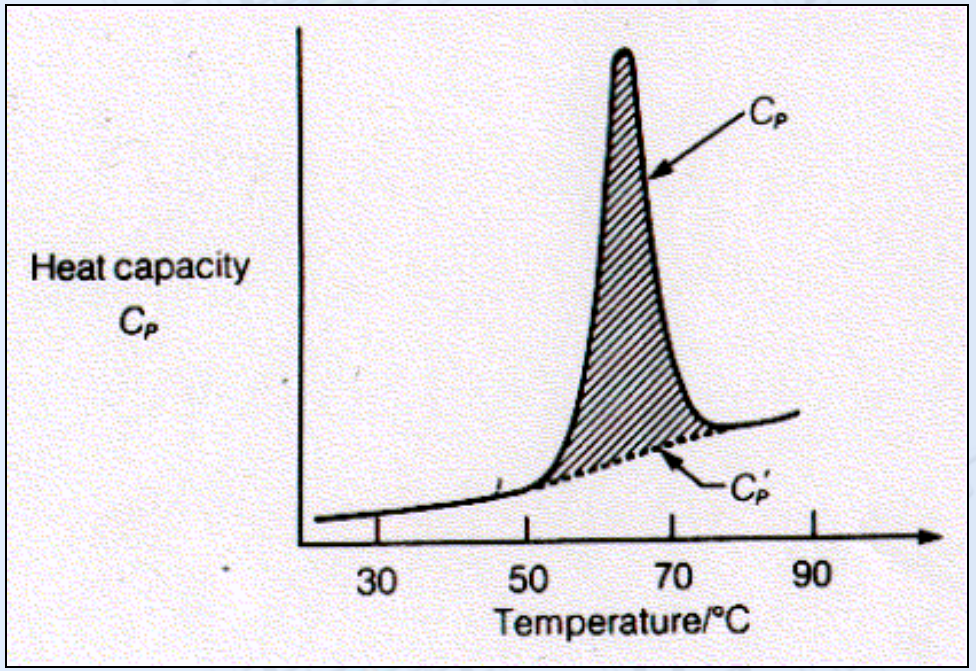
“Knowledge Based” Amino Acid Pair Potentials

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

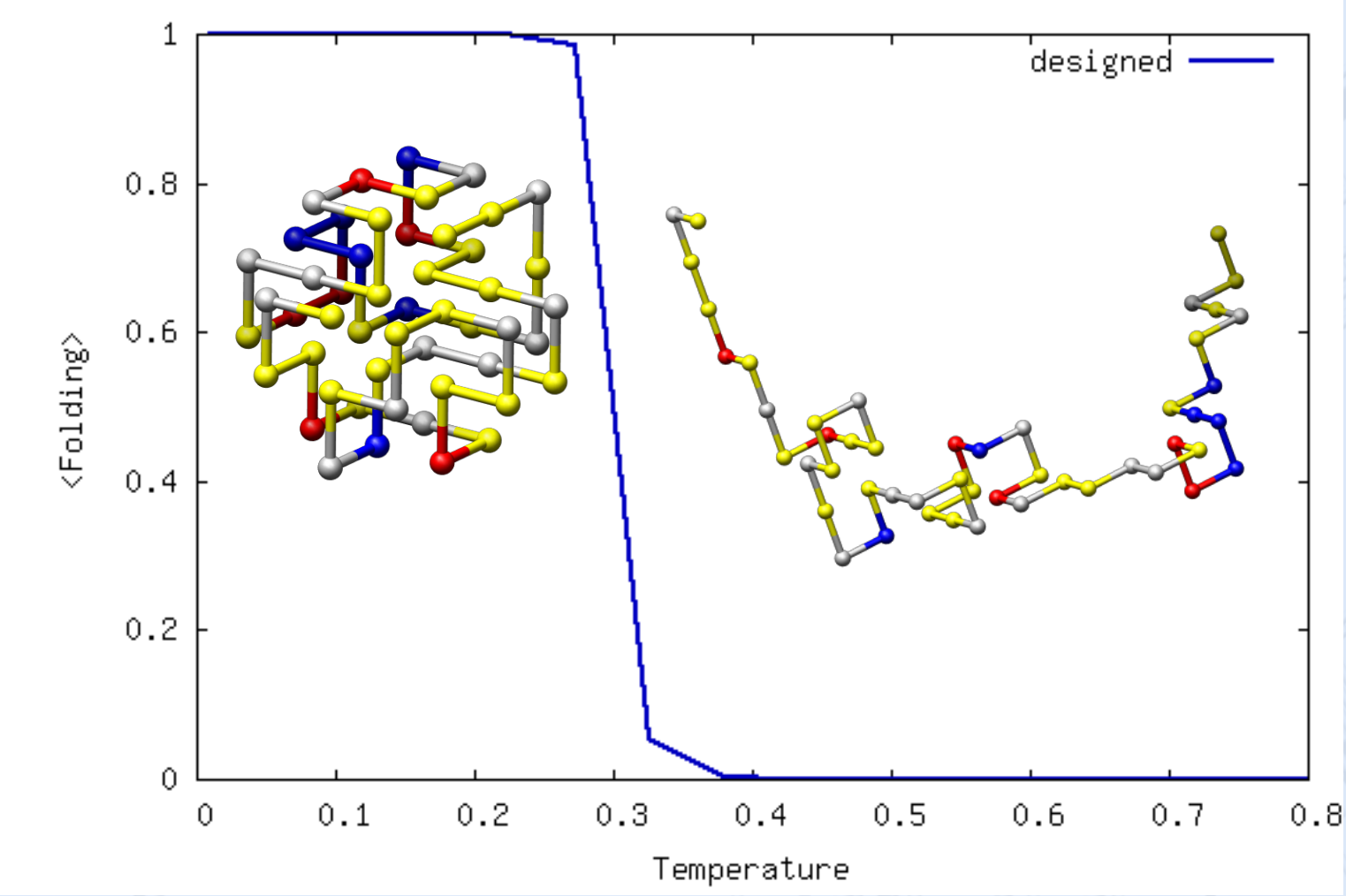
- How could biology (evolution) affect these results?
- How could we prevent this?

Are experimental result captured by the model?

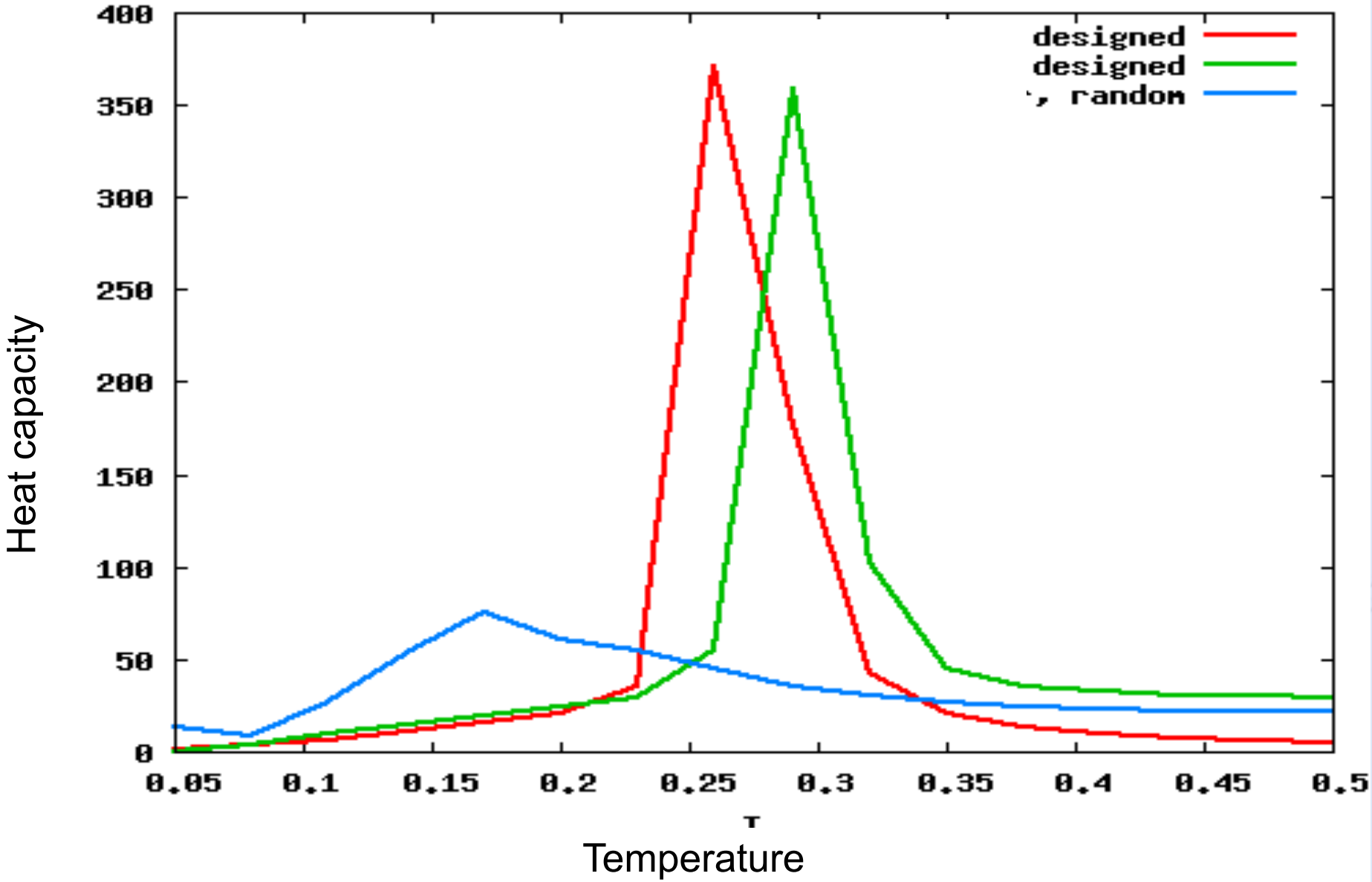




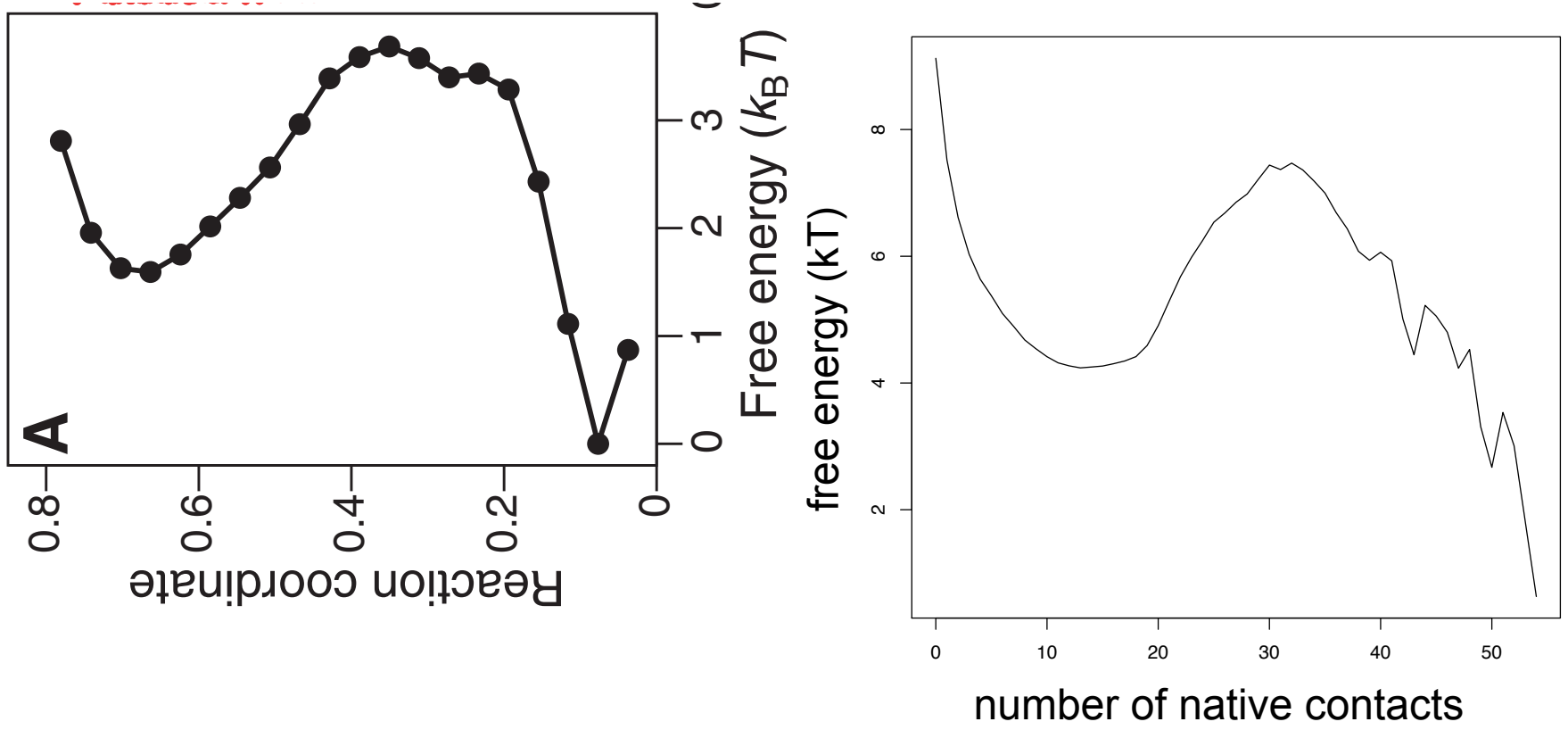
Folding Specificity on the Lattice



Foldable, with high specificity

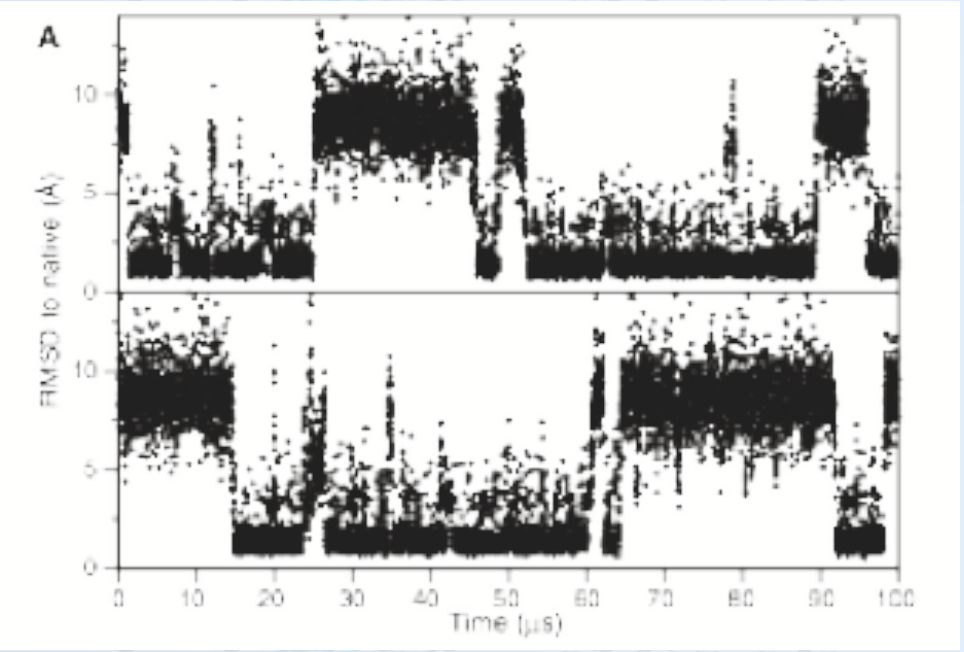


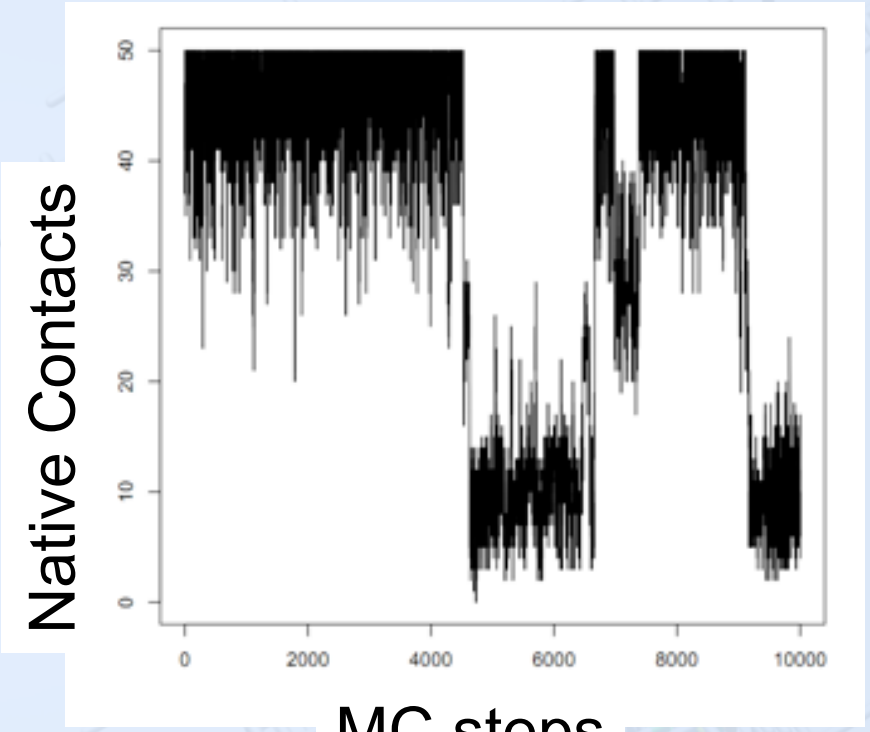
Full atom vs coarse grained folding



Shaw, D. E., et al. (2010) Science, 330

Full atom vs coarse grained folding

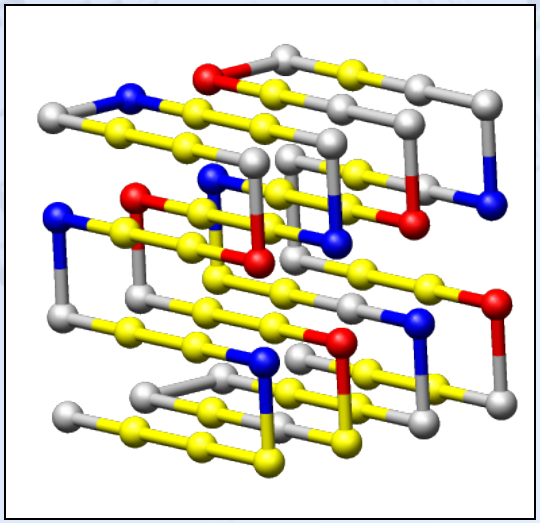


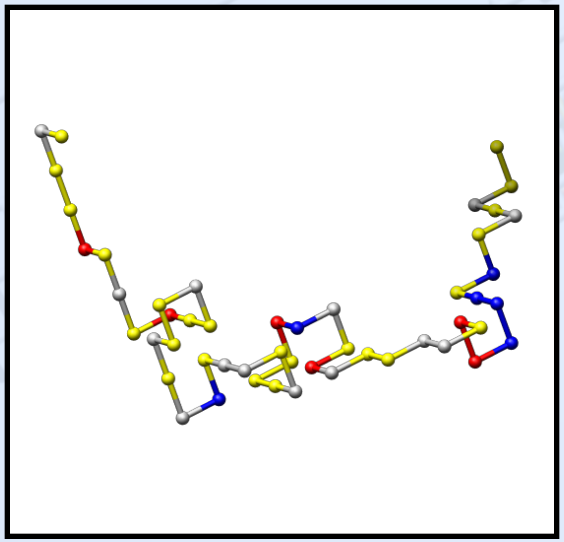


Shaw, D. E., et al. (2010) Science, 330

(thanks to Erik van Dijk)

Cubic Lattice Model





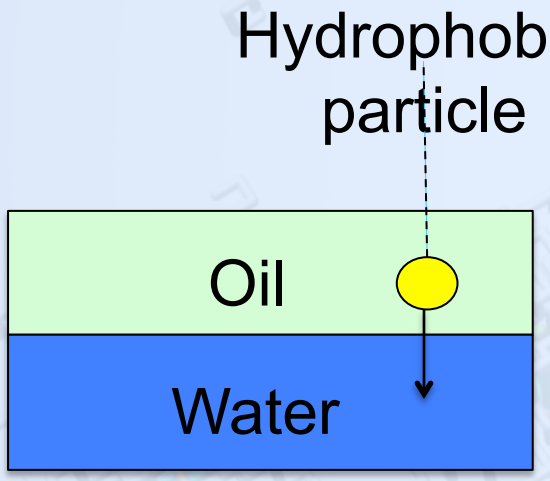
- Cheap & simple
  - Use for right purpose
- Can model:
  - General trends
  - Folding specificity
  - Heat capacity
  - Binding and unbinding
- Not captured:
  - Secondary structure
  - Hydrophobic effect (cold denaturation)
  - Specific proteins

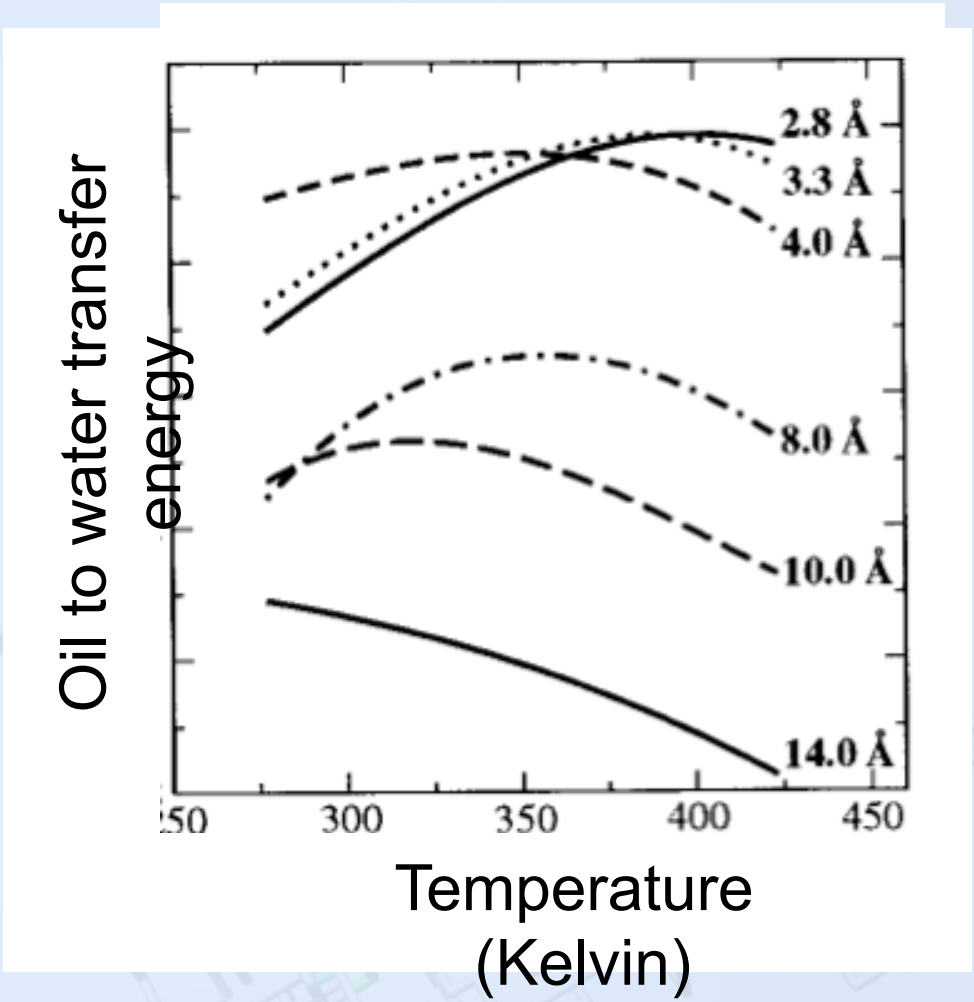
Shakhnovich & Gutin 1993 PNAS 90

Coluzza et al 2003 Phys Rev E 68

Emergent behaviour (entropy & enthalpy)

Hydrophobic force has a maximum around 70 - 80 °C



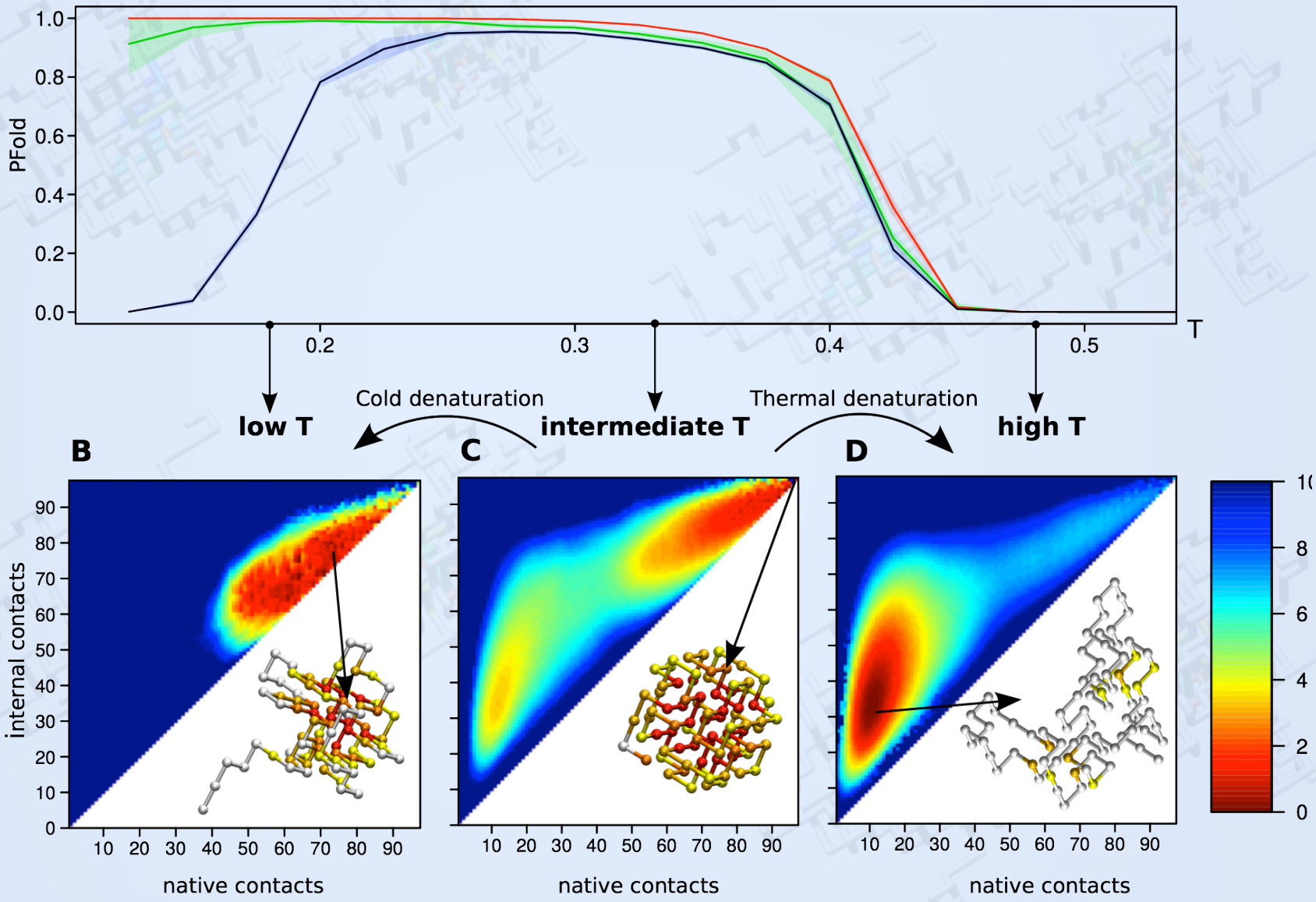


Widom 2003, Phys. Chem. Chem. Phys., 2003,

Gallagher 2003, JACS

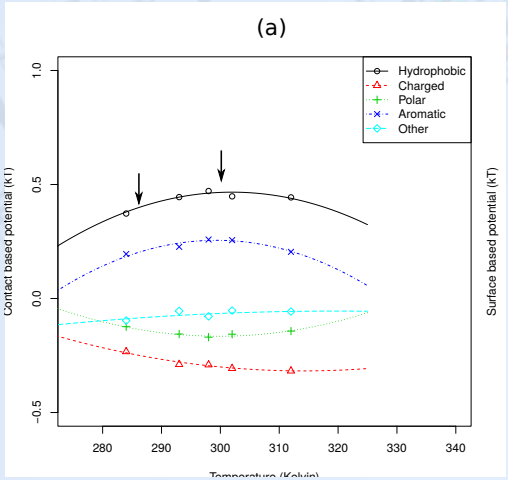
Huang and Chandler PNAS 2000

Use a lattice model to investigate cold denaturation

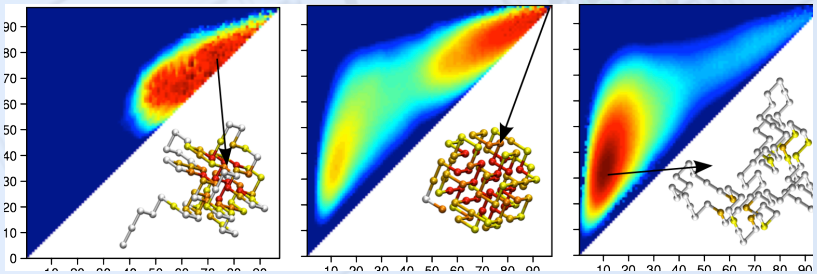


Impact of understanding hydrophobicity

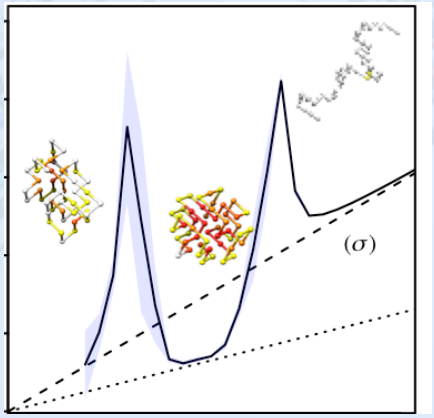
Measure hydrophobic temperature dependence



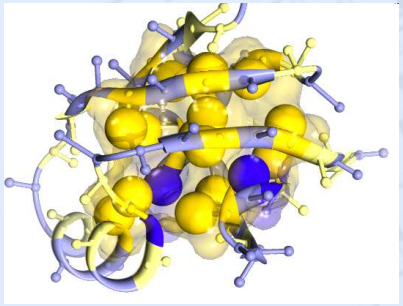
We can explain cold denaturation

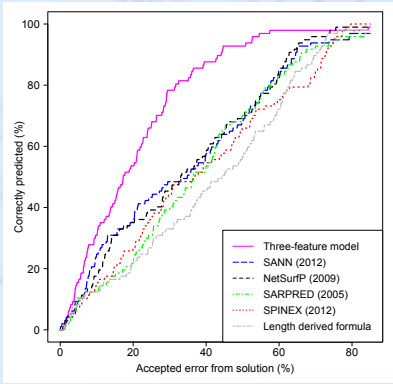


We can understand heat capacity curves



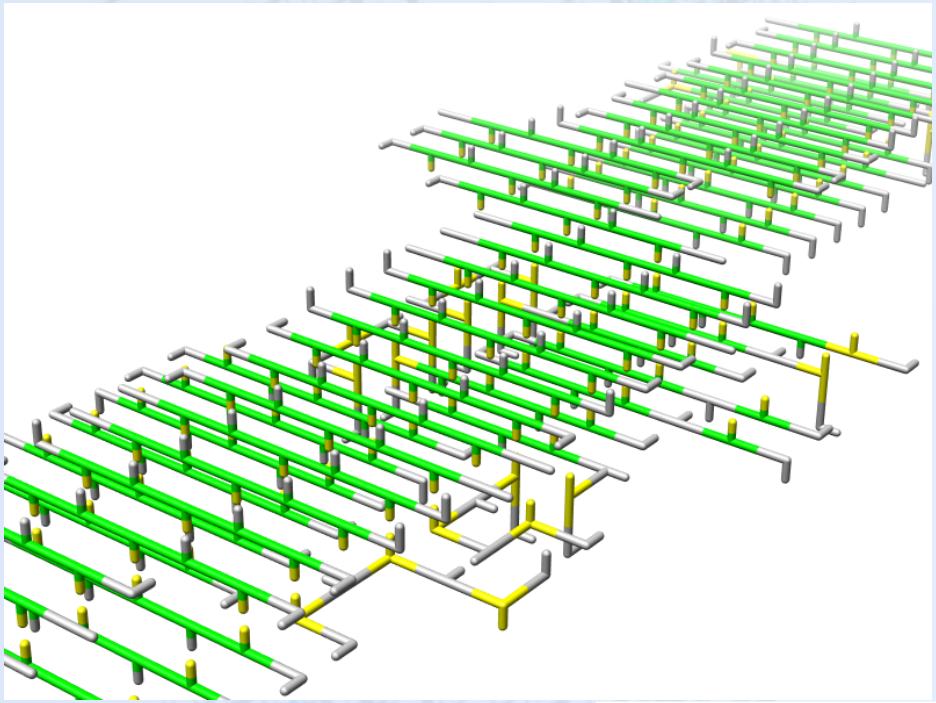
Predict total hydrophobic surface area





Model can reproduce formation of fibres

Abeln, S., Vendruscolo, M., Dobson, C. M., & Frenkel, D. (2014). A Simple Lattice Model That Captures Protein Folding, Aggregation and Amyloid Formation. PLoS ONE, 9(1), e85185

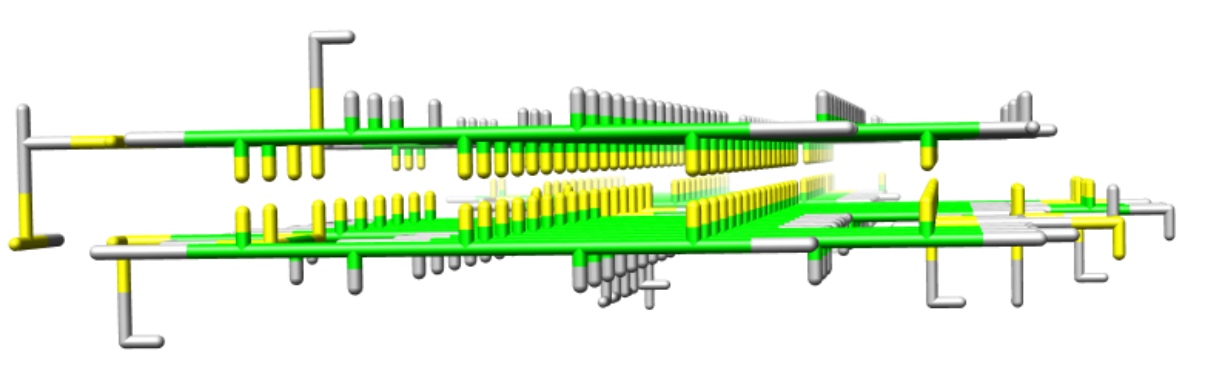


- Fibres
  - Formation depends on sequence properties
  - Hydrophobic inner layers
  - Fast simulation

hydrophobic

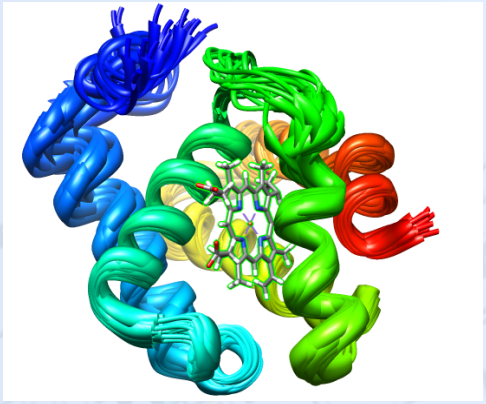
polar (hydrophilic)

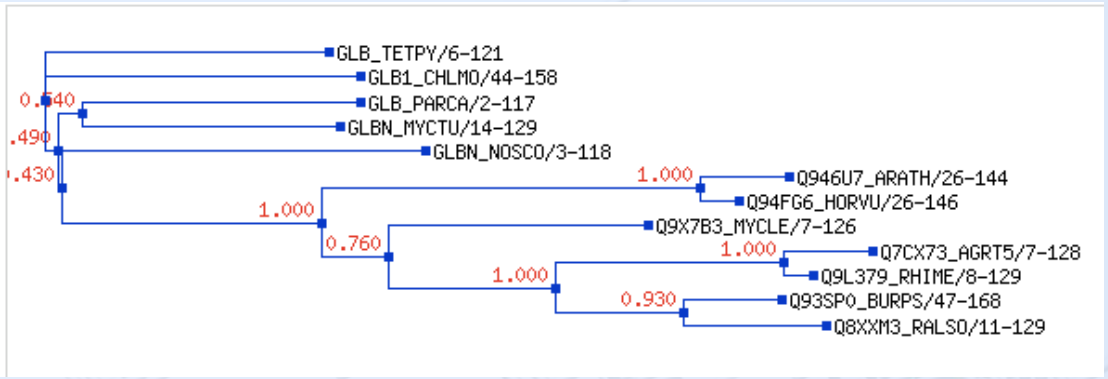
beta-strand

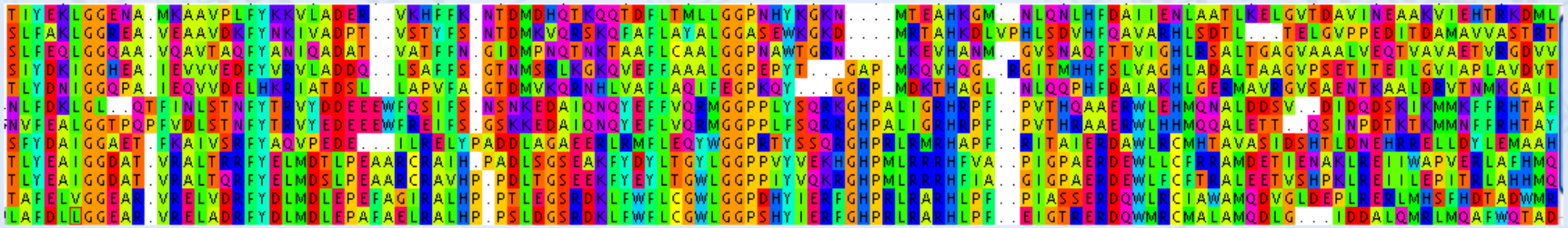


Biology & Physics:

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

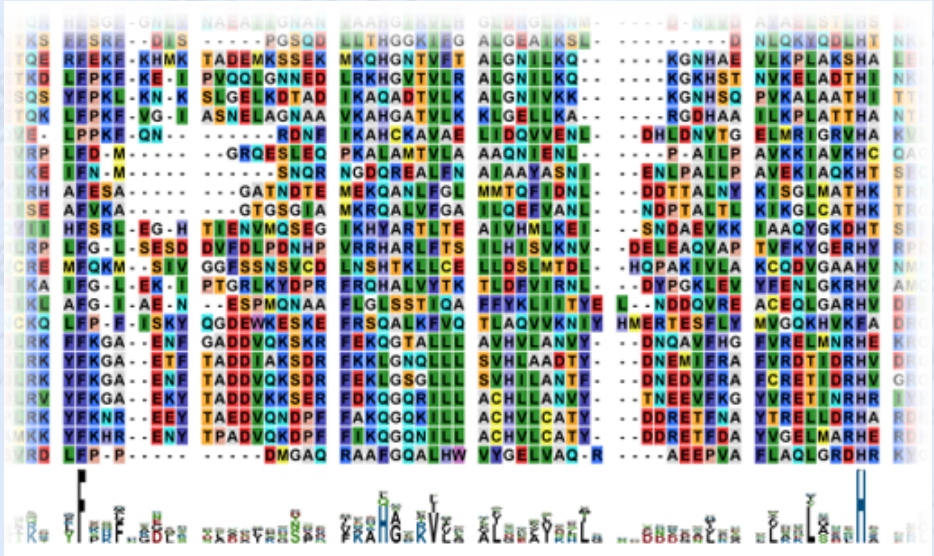






Lessons for full atomistic simulations - do not forget about evolution

- 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”



References

Abeln, S., & Frenkel, D. (2008). Disordered flanks prevent peptide aggregation. *PLoS Comput. Biol.*, *4*(12), e1000241. doi:10.1371/journal.pcbi.1000241

Abeln, S., & Frenkel, D. (2011). Accounting for protein-solvent contacts facilitates design of nonaggregating lattice proteins. *Biophys. J.*, *100*(3), 693–700. doi:10.1016/j.bpj.2010.11.088

Abeln, S., Vendruscolo, M., Dobson, C. M., & Frenkel, D. (2014). A Simple Lattice Model That Captures Protein Folding, Aggregation and Amyloid Formation. *PLoS ONE*, *9*(1), e85185. doi: 10.1371/journal.pone.0085185

Ni, R., Abeln, S., Schor, M., Cohen Stuart, M. A., & Bolhuis, P. G. (2013). Interplay between Folding and Assembly of Fibril-Forming Polypeptides. *Physical Review Letters*, *111*(5), 058101. doi: 10.1103/PhysRevLett.111.058101

Ni, R., Kleijn, J. M., Abeln, S., Cohen Stuart, M. a., & Bolhuis, P. G. (2015). Competition between surface adsorption and folding of fibril-forming polypeptides. *Physical Review E*, *91*, 1–5. doi: 10.1103/PhysRevE.91.022711

van Dijk, E., Hoogeveen, A., & Abeln, S. (2015). The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures. *PLOS Computational Biology*, *11*(5), e1004277. doi:10.1371/journal.pcbi.1004277

van Dijk, E.; Varilly, P.; Knowles, T.; Frenkel, D.; Abeln, S. Consistent treatment of hydrophobicity in protein lattice models accounts for cold denaturation. ArXiv e-prints 2015, [arXiv:cond-mat.soft/ 1511.06590] – accepted PRL