Protein folds, fold classifications & structure stability

Magnus Andersson

magnus.andersson@scilifelab.se

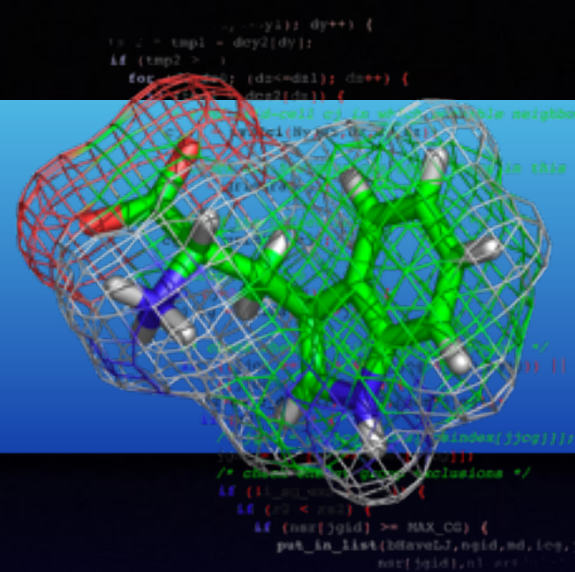
Theoretical & Computational Biophysics

SciLifeLab

KTH VETENSKAP OCH KONST

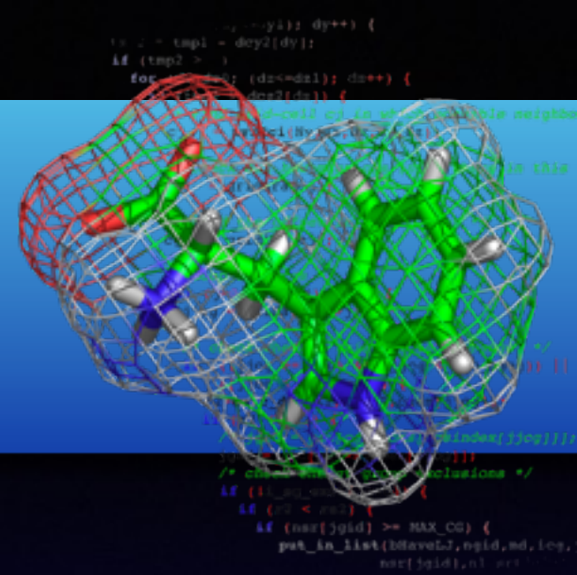
UNIVERSITET STOCKHOLMS

Recap



- Globular proteins
  - α,β,mixed proteins
  - Common supersecondary structure motifs
    - Rossman fold, Greek key motif etc
- Membrane proteins
  - Mostly α-helix, but some β-barrels
  - Stabilized by internal H-bonds in hydrophobic environment
  - Leading research area in Stockholm

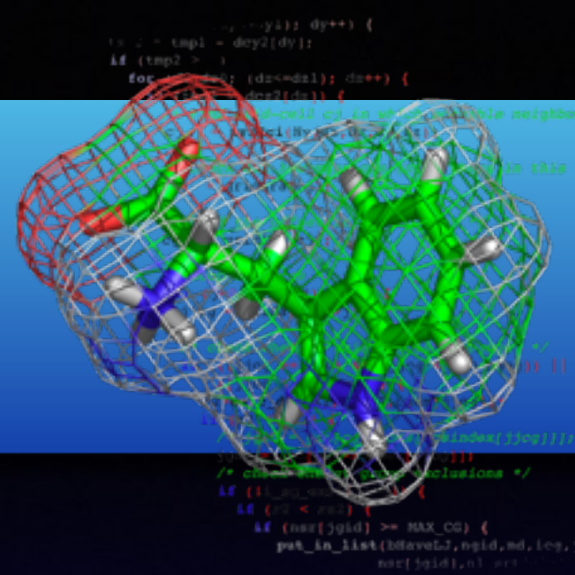
Outline today



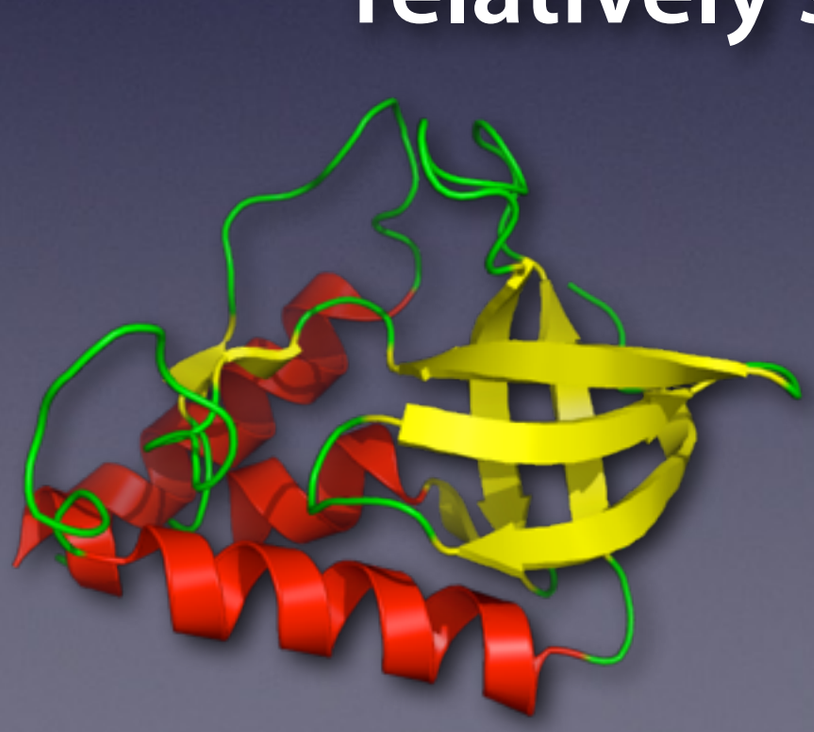
- Fold stability
- Structural evolution
- Protein size variation
  - Why helices/sheets have certain sizes
- Boltzmann statistics for folds - or not?
- Sequence-structure compatibility
- Fold stabilization from residues
- How stable are proteins, and why?

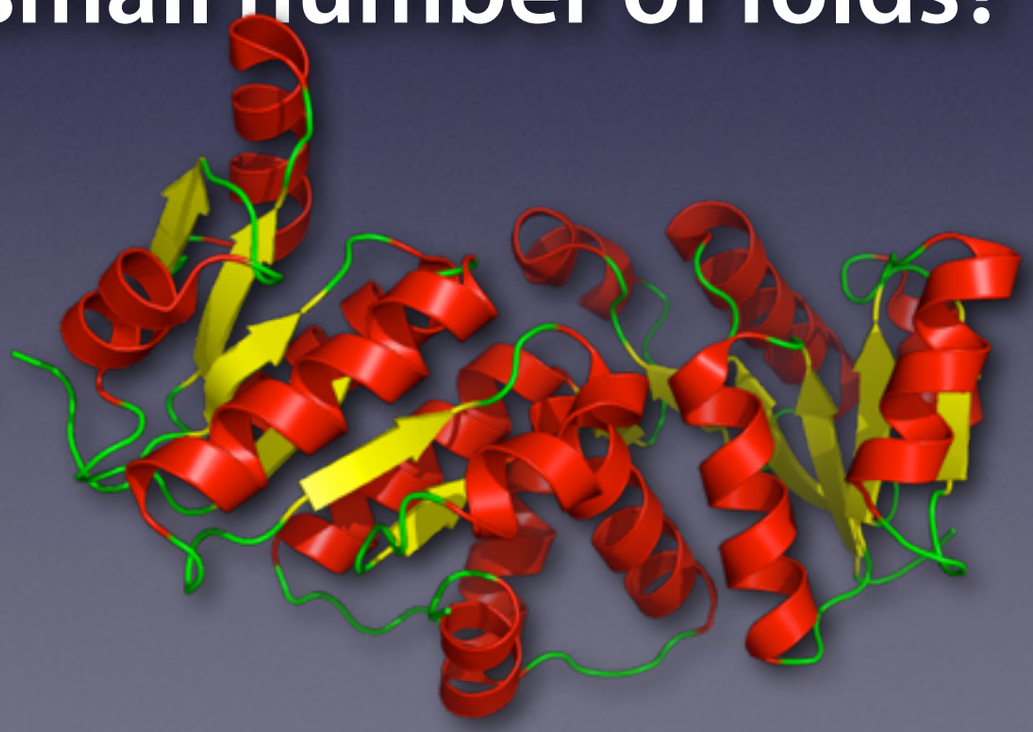
Protein physics book: Chapters 15 & 16

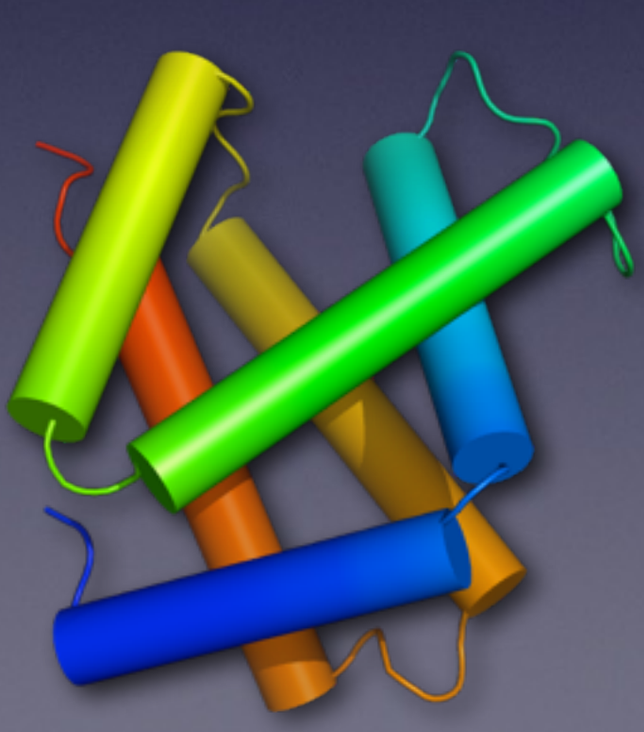
The fold universe



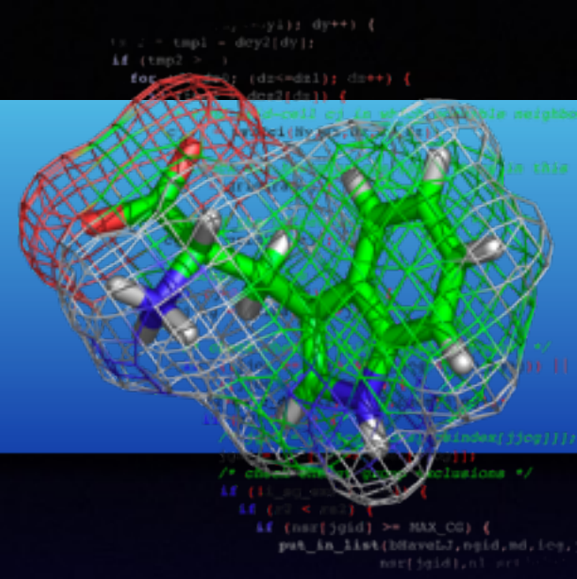
- Why are there so few protein folds?
- Chothia: *"~~1000~~ 1500 folds for the molecular biologist"*
- Why do most sequences seem to fit a relatively small number of folds?



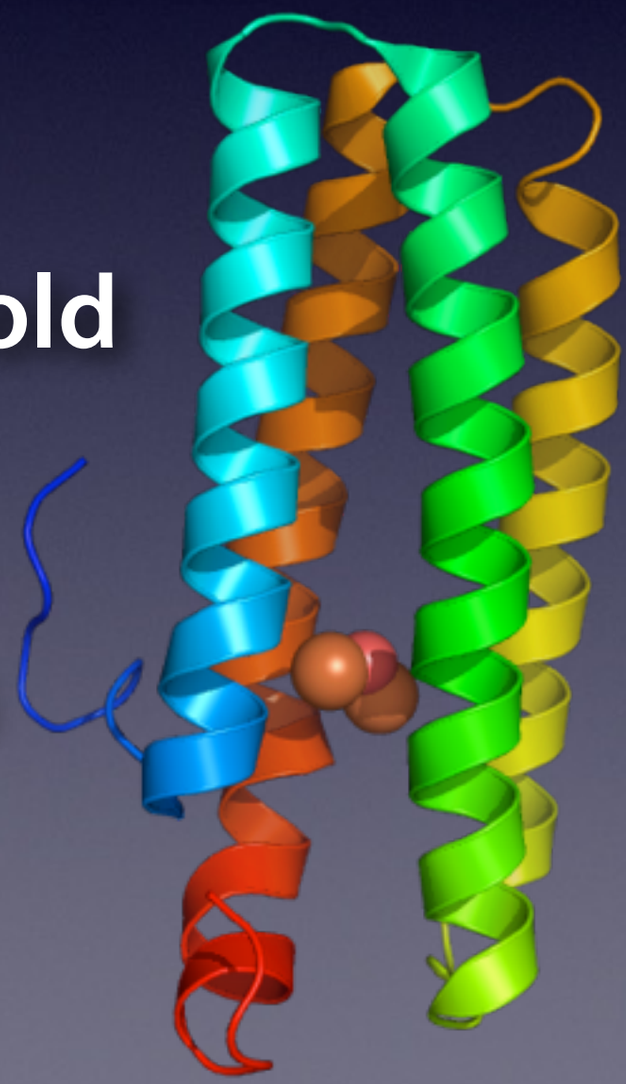




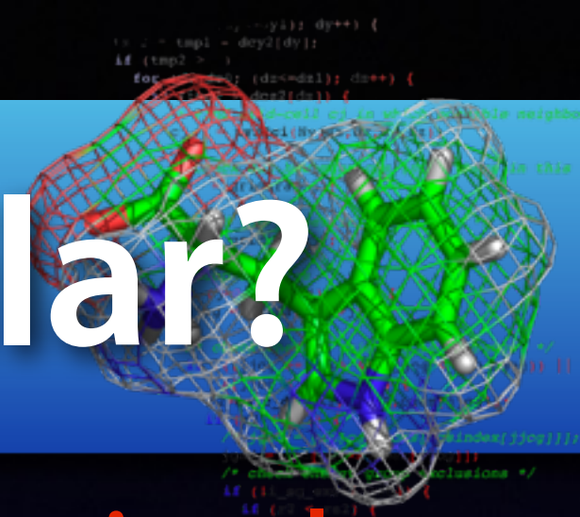
“Typical” folds



- 20% of folds account for 80% of proteins
- Mostly true for RNA too
- Compare with DNA: Only a single fold
- Homologous sequences
- Functional convergence onto folds
- Physical restrictions



Why are proteins similar?



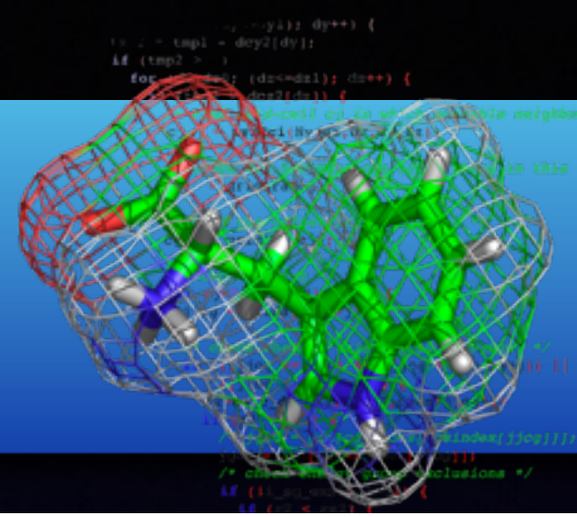
Evolutionary Divergence

Functional Convergence

?

Limited number of possible folds

Folding patterns

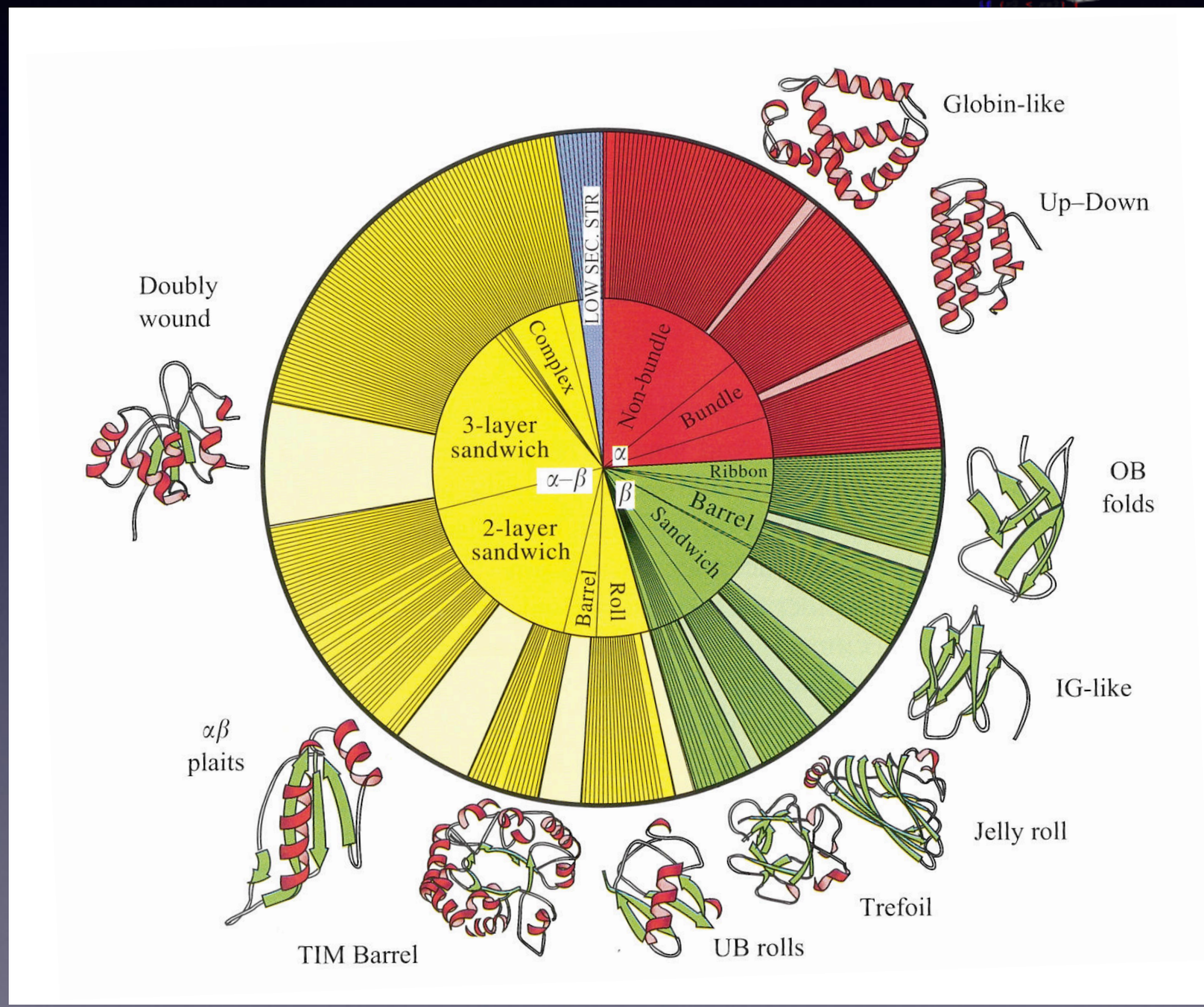


Simple permutations of helices/sheets

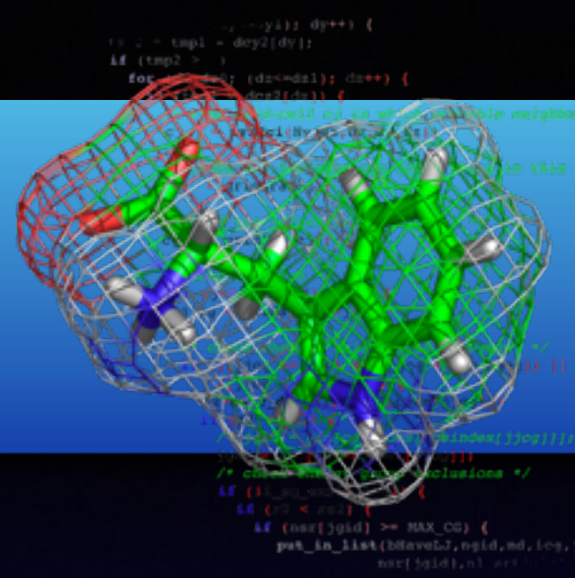
Stable local patterns (lots of h-bonds)

Hydrophobic patterns

Contiguous sheets



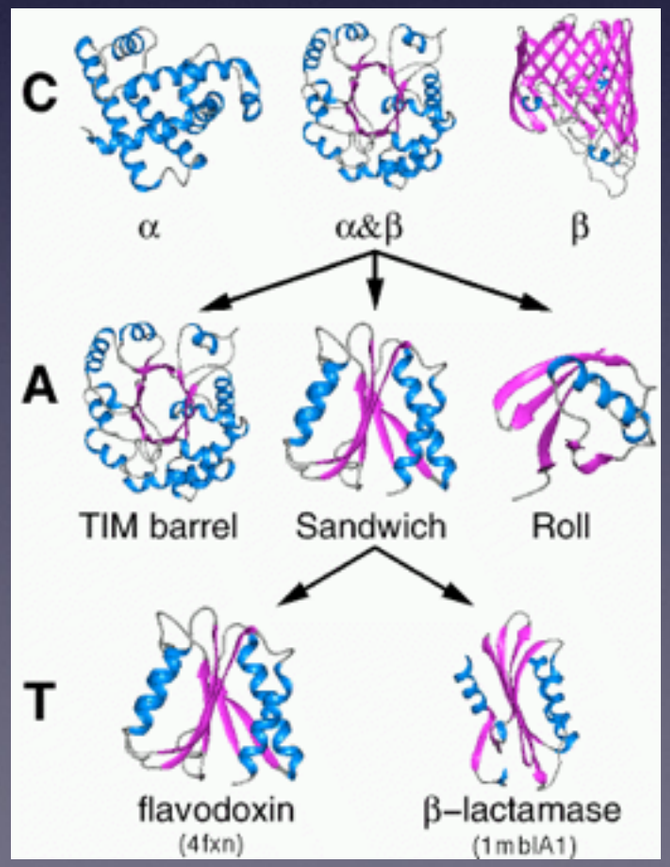
Fold classifications

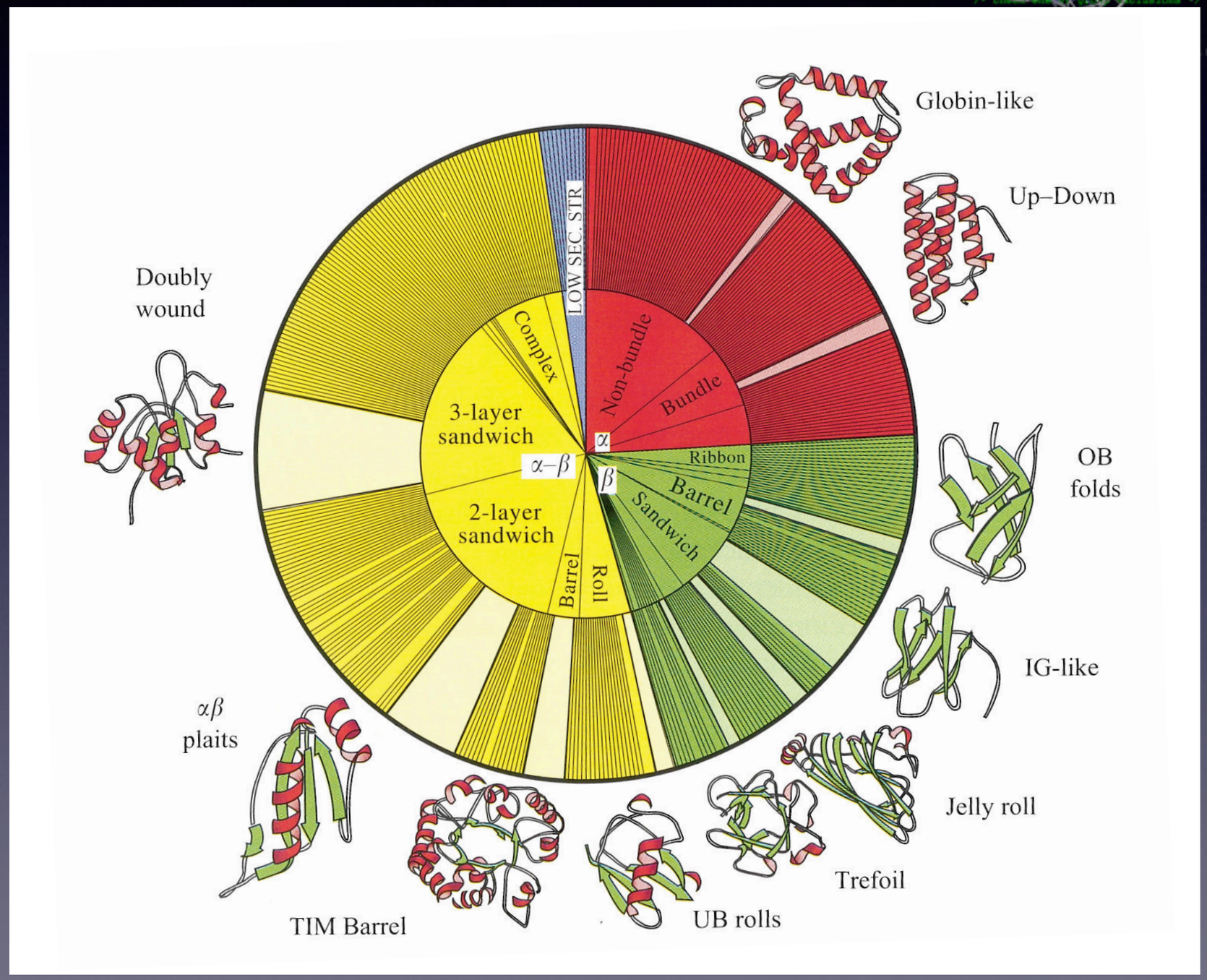


- Structural alignments
- CATH
- SCOP

CATH - 90 % automatic

Class
Architecture
Topology
Homology





CATH - 235,858 domains

Orengo & Thornton

Domain: 1fuiA03

http://cathwww.biochem.ucl.ac.uk/cgi-bin/cath/Domain.pl?domain_id=1fuiA03

CATH

Domain: 1fuiA03

CATH Protein Structure Classification

CATH DHS Gene3D FTP

Home > Top >

Search

Go!

Goto

SSAP Server
CATHEDRAL Server
DHS
Gene3D

Navigation

Home
Top of hierarchy

Cath Update

Not logged in: [login]
DB cathdb_v3_1_0

Domain: 1fuiA03

Version: v3_1_0 | Version: current

Domain: 1fuiA03

Status

The domain has been assigned to a CATH superfamily and does not require any further processing.

Classification (3.20.14.10.1.1.1.1)

Class 3
Alpha Beta

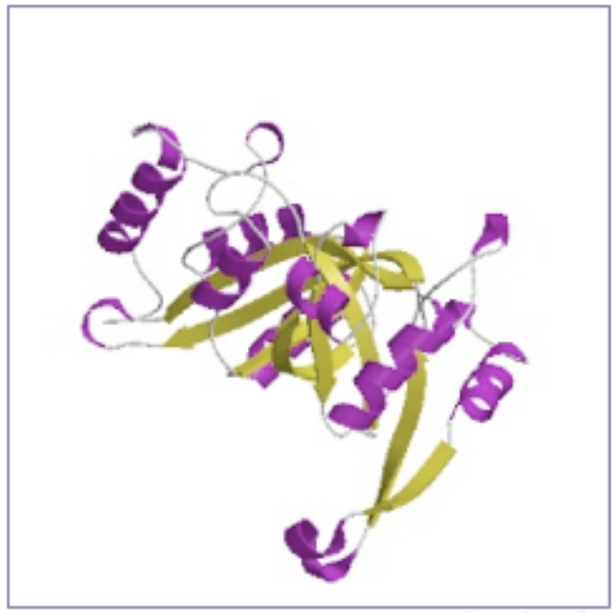
Architecture 3.20
Alpha-Beta Barrel

Topology 3.20.14
L-fucose Isomerase; Chain A, domain 3

Homologous Superfamily 3.20.14.10
L-fucose Isomerase; Chain A, domain 3

S35 Family 3.20.14.10.1

S60 Family 3.20.14.10.1.1



Domain Boundaries

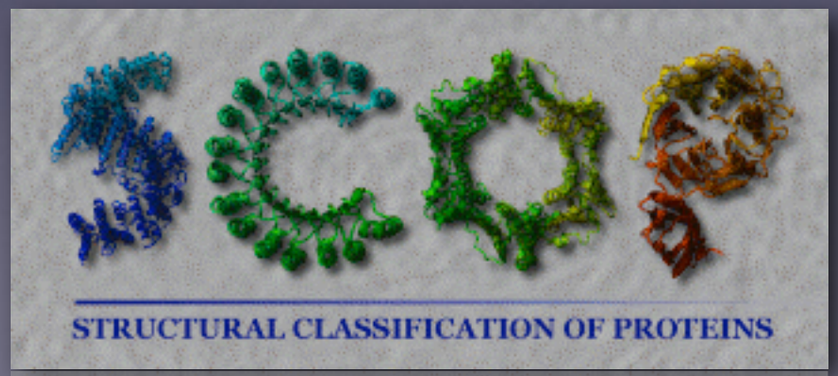
SCOP - 192,710 domains

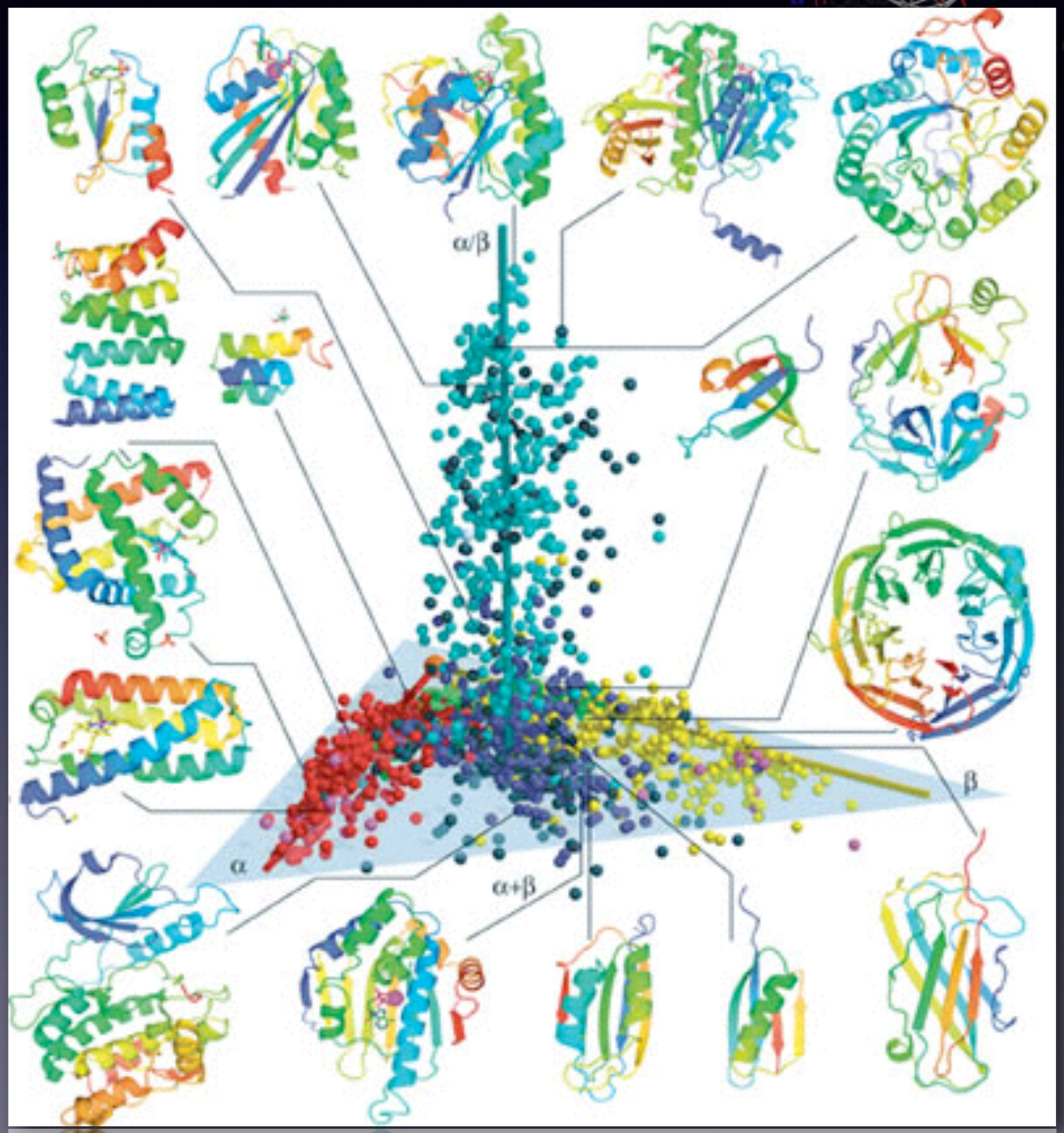
Access methods

- Enter SCOP at the **top of the hierarchy**
- Keyword search of SCOP entries
- SCOP parseable files
- All SCOP releases and reclassified entry history
- SCOP domain sequences and pdb-style coordinate files (ASTRAL)
- Hidden Markov Model library for SCOP superfamilies (SUPERFAMILY)
- NEW Structural alignments for proteins with non-trivial relationships (SISYPHUS)

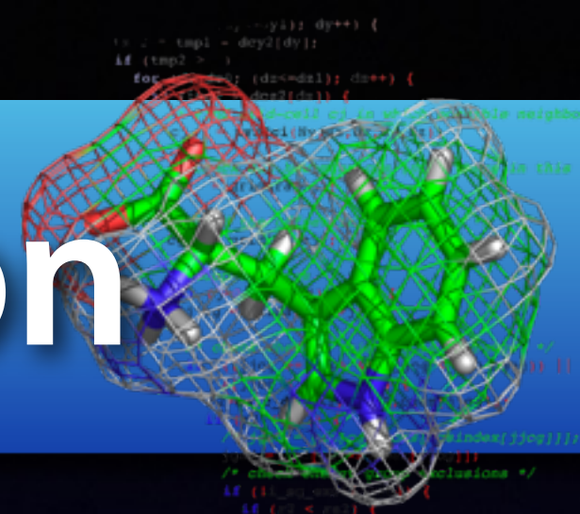
ASTRAL, SUPERFAMILY, etc.

Murzin, Brenner, Chotia



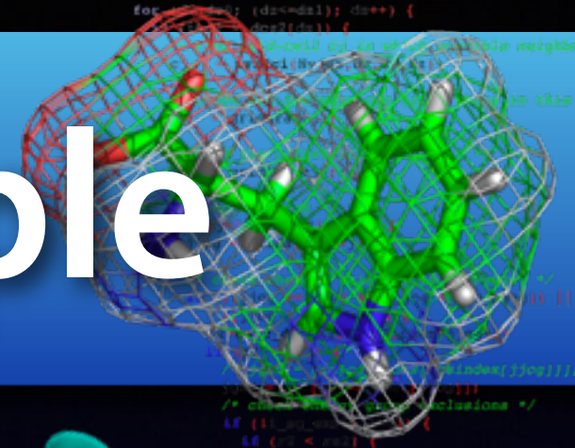


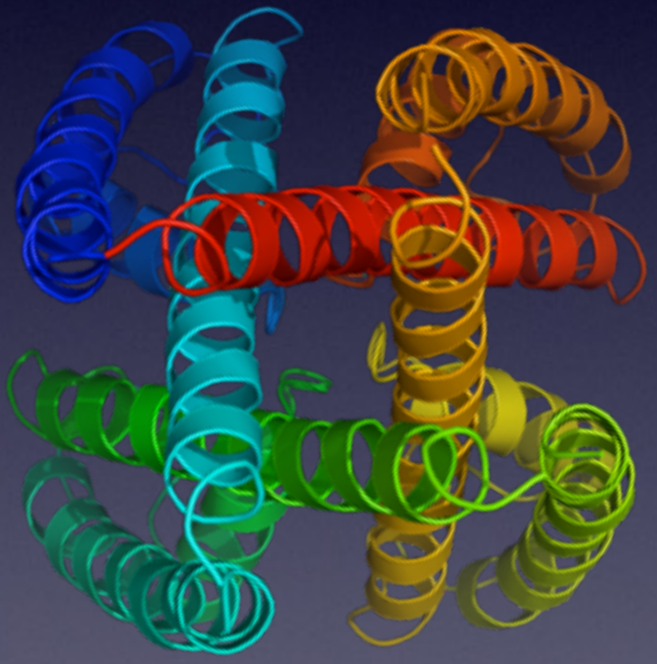
Structural Evolution



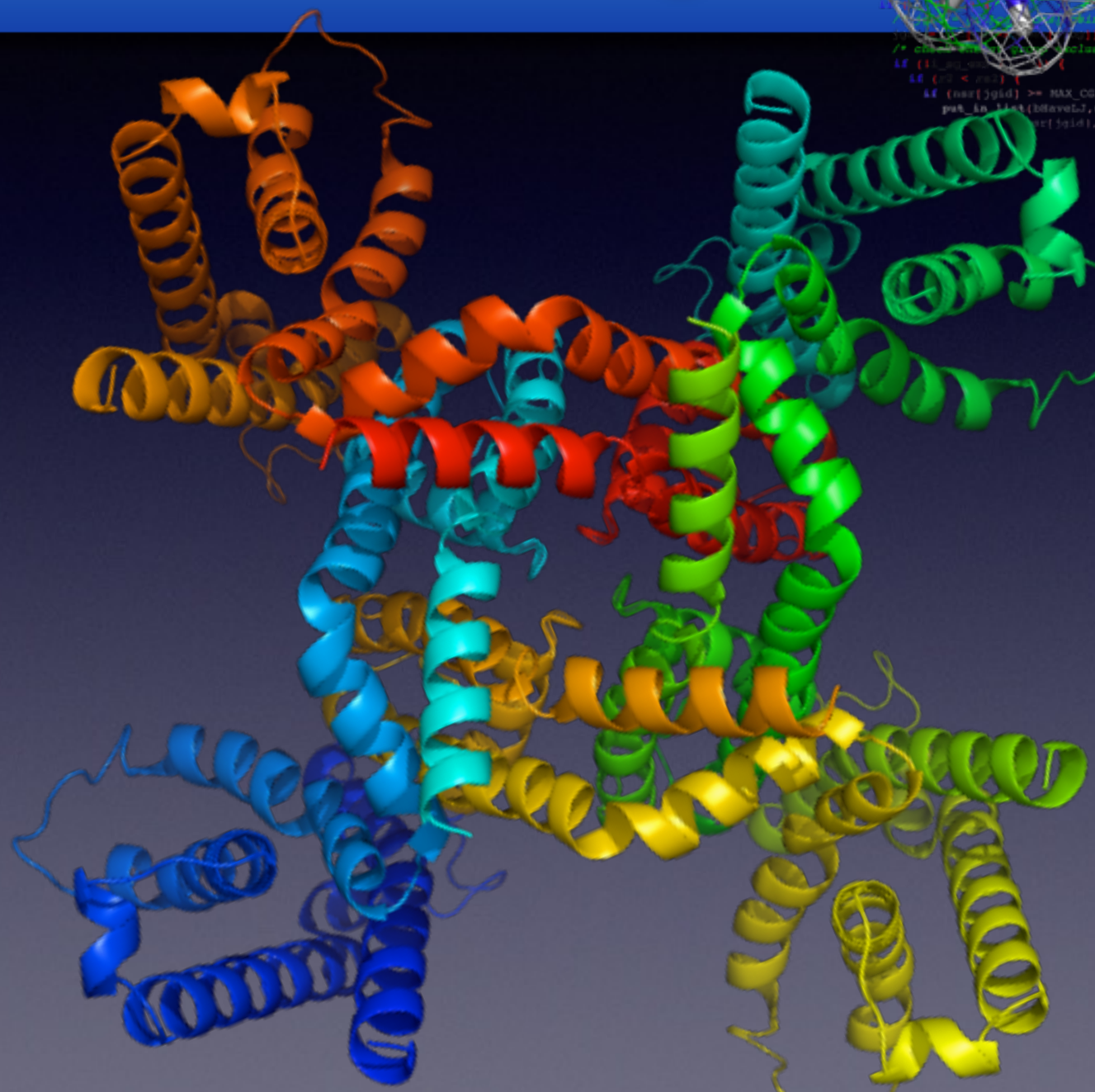
- Llama hemoglobin binds oxygen harder than pony/horse hemoglobin
- Fetal hemoglobin is different from adult!
- Genes can be shut on/off in organisms
- Are eukaryotic/vertebrate proteins more complex than prokaryotic ones?
  - Folding patterns seem to be similar
  - Eukaryotic proteins sometimes have more domains, and they can be larger

K+ channel example



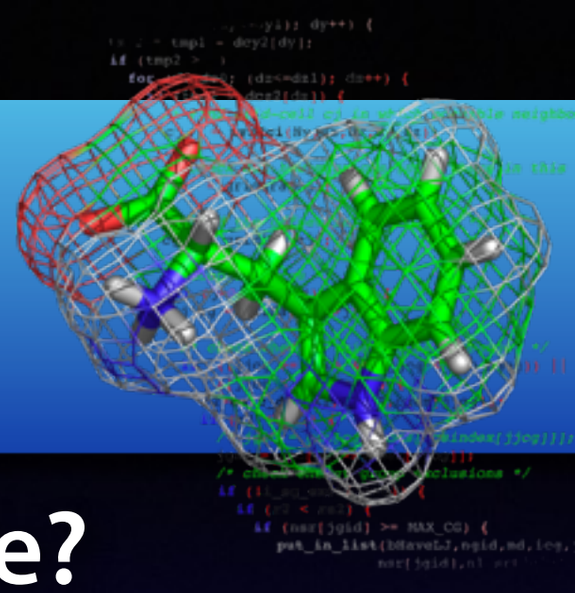


KcsA (bacterial)

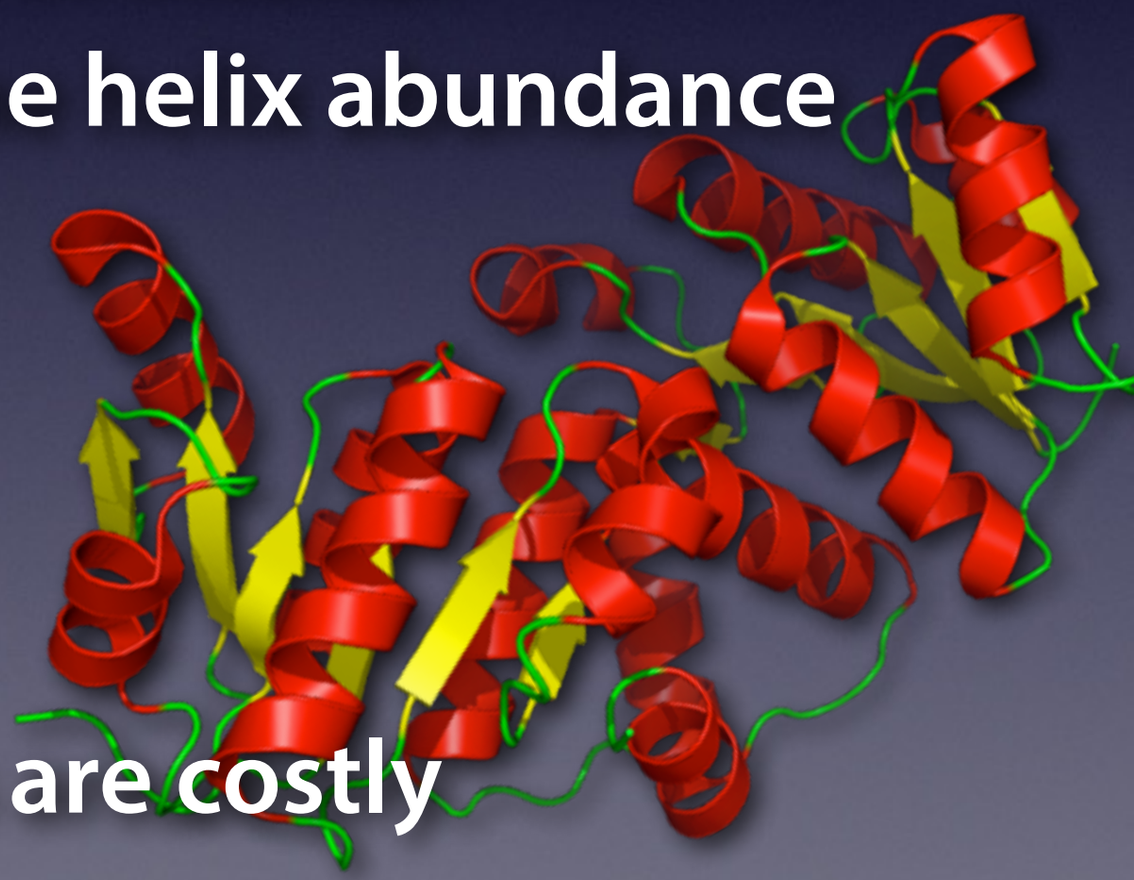


Kv1.2 (eukaryotic)

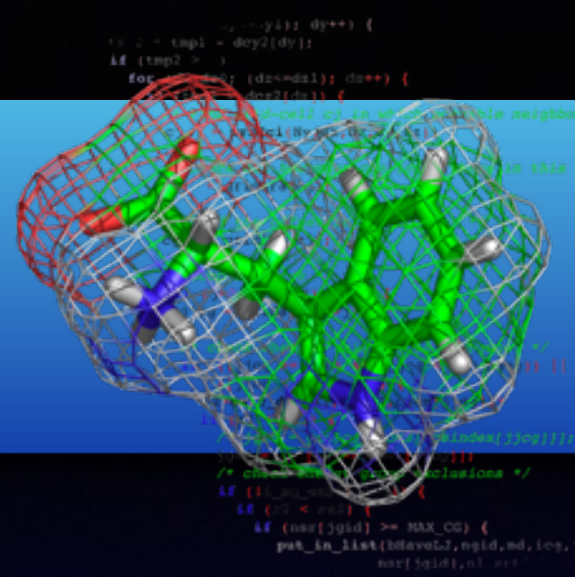
Structural stability



- Why are the common structures stable?
- H-bond saturation!
  - Loops/coil cannot exist in interior
  - Also explains membrane helix abundance
  - Edges of helices/sheet must face water
  - Helix & sheet regions must be separate
- Structure/energy defects are costly

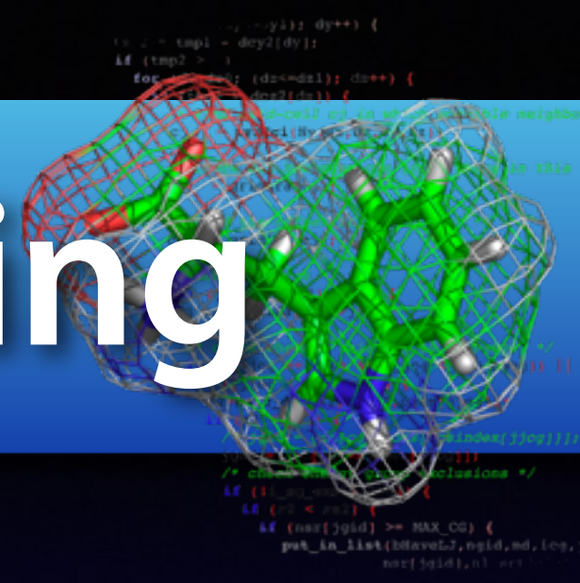


Fold layers



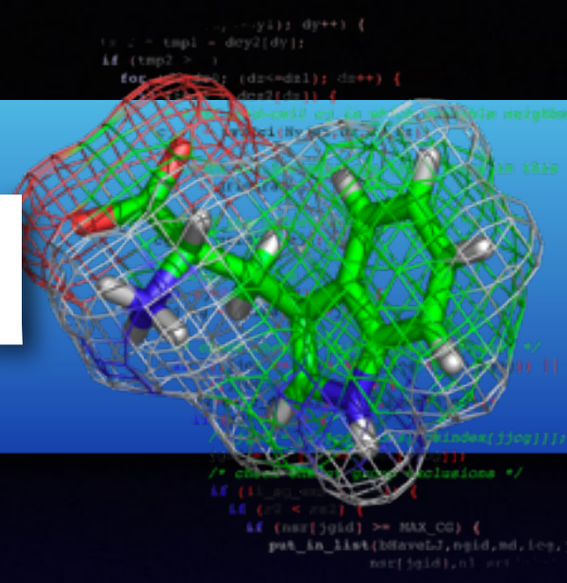
- 1 layer: Not very useful
- 2 layers: Great for shielding
- 3 layers: Rossman fold, double cavities
- 4 layers: Rare, buries hydrophilic aa:s
- 5 layers: Doesn’t occur in practice
- Large proteins by necessity need to be divided into subdomains for stability!

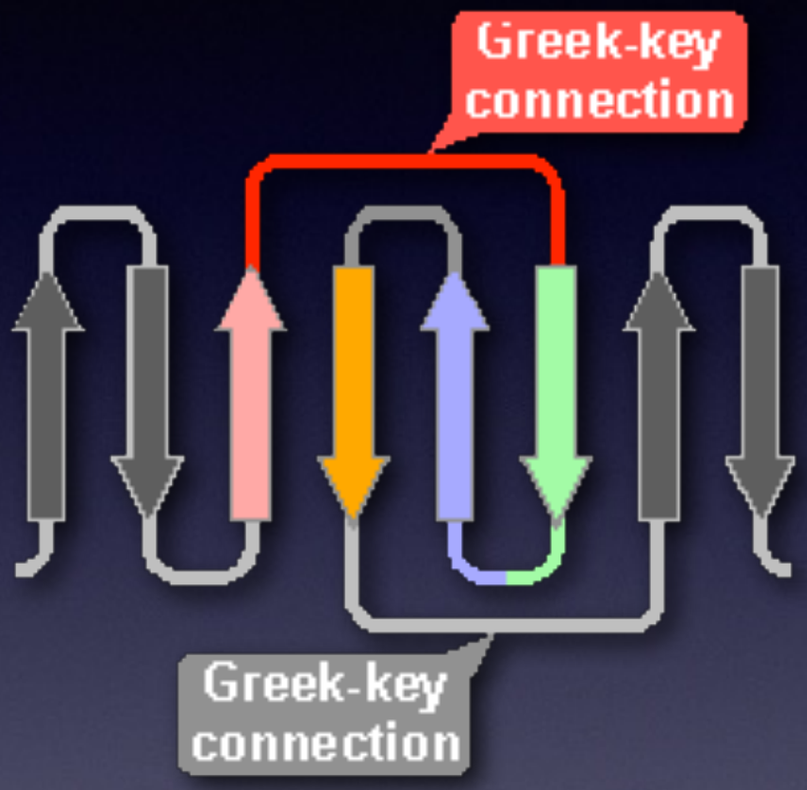
Sequence-fold fitting

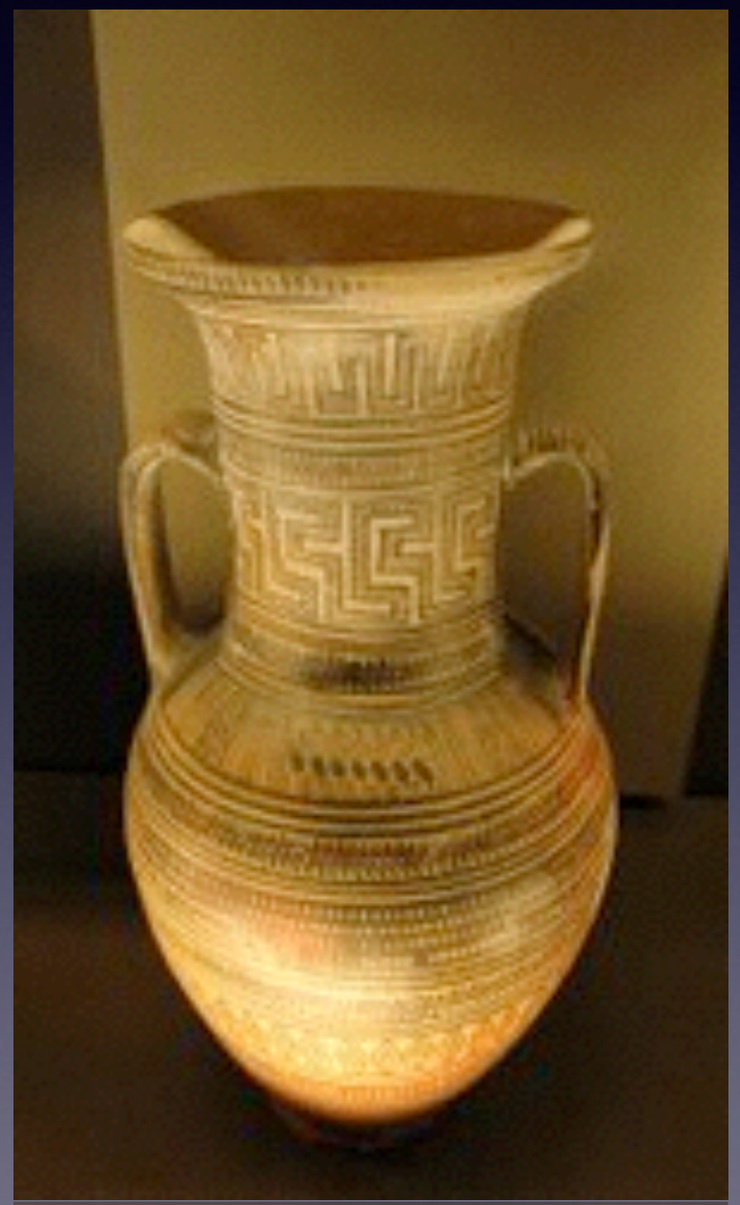


- So, which *sequences* can fit a given fold?
- Simple folds can accommodate lots of sequences - that's why they are common
- A fold with special defects requires special amino acids (e.g. Cys bridges) for stabilization, and can only accomodate a few sequences
- Natural selection at work!

Greek keys, revisited



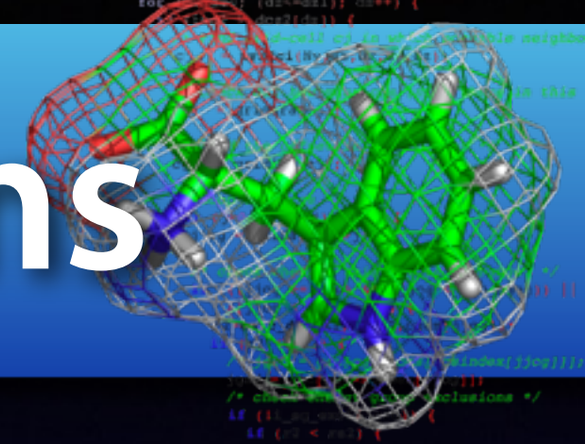




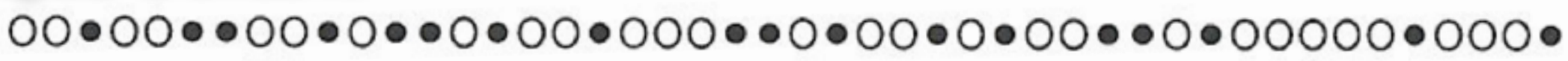
It is not a coincidence that we see this pattern both on vases and in proteins - can you think of why?

(Richardson, Nature 1977)

Sequence patterns



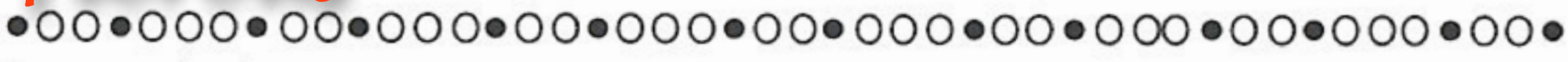
Globular



Membrane

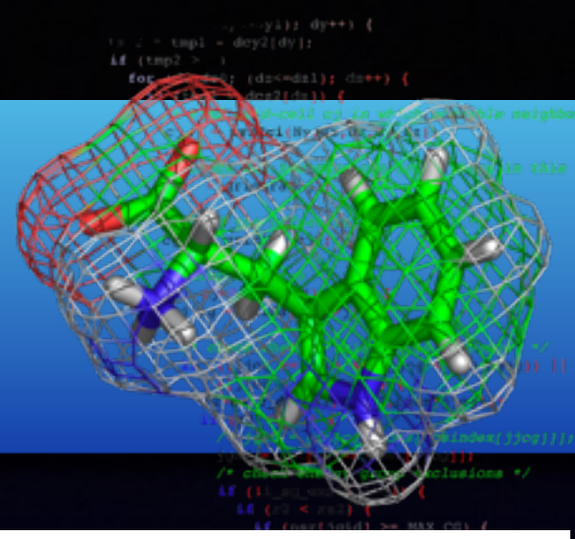
| Hydro-phobic | Hydro-philic |

Fibrous

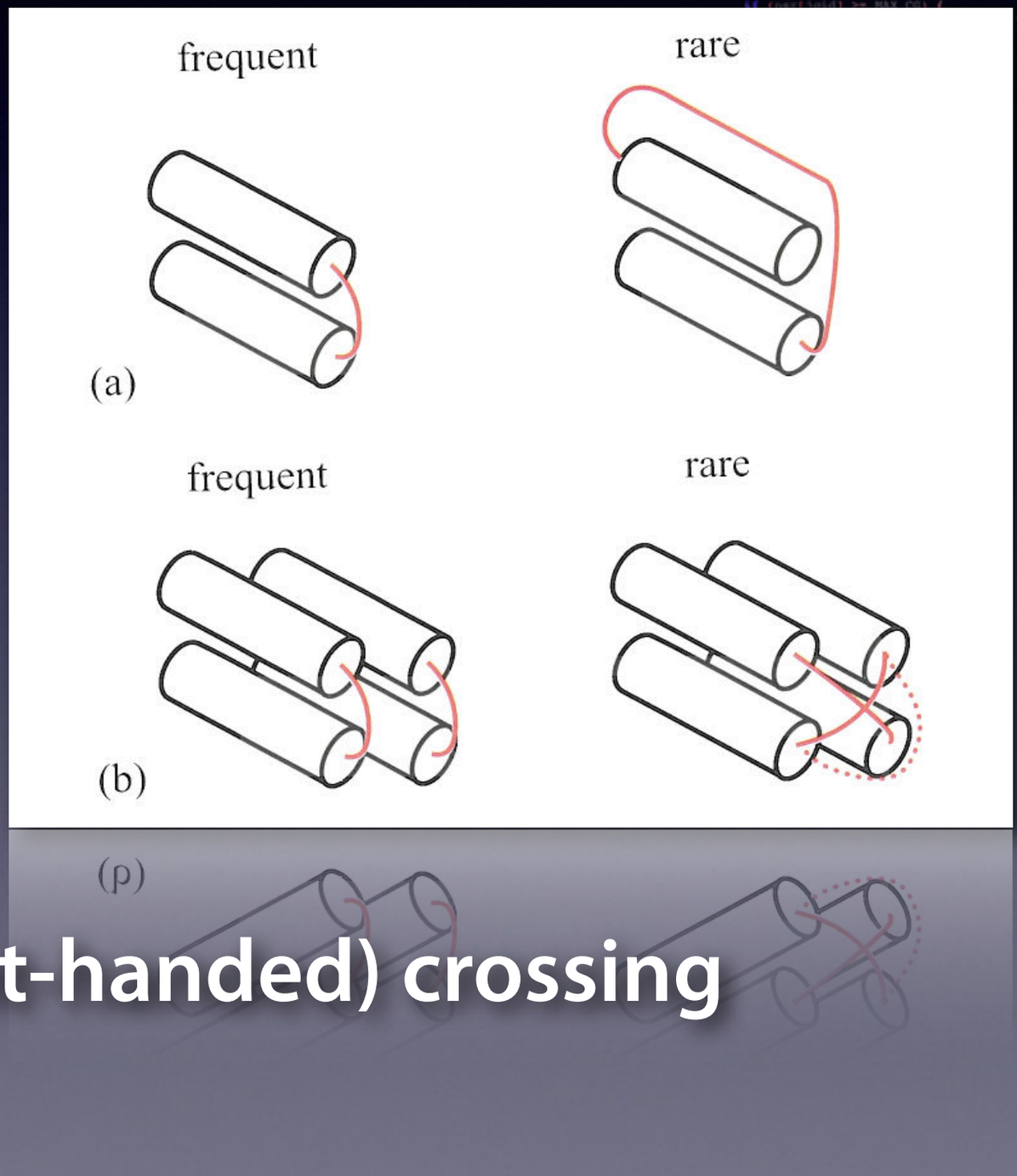


| repeat |

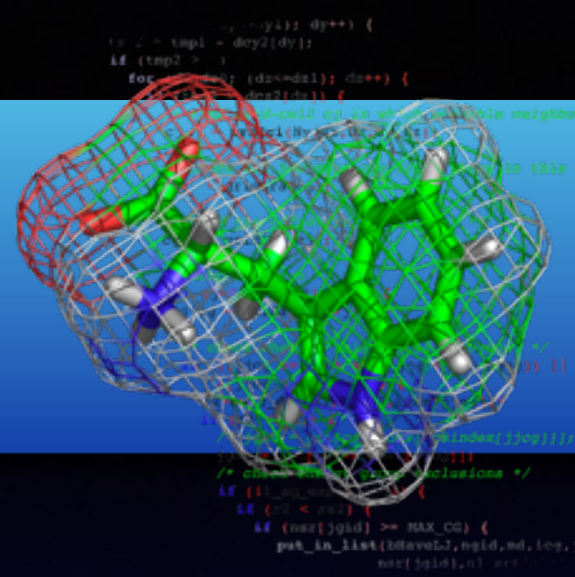
Structural stability



- Why are defects rare?
- Loss of 1-2 h-bonds
- But that would only cost 5-10 kcal/mol?
- Small fraction of total E
- Same for beta sheet (right-handed) crossing

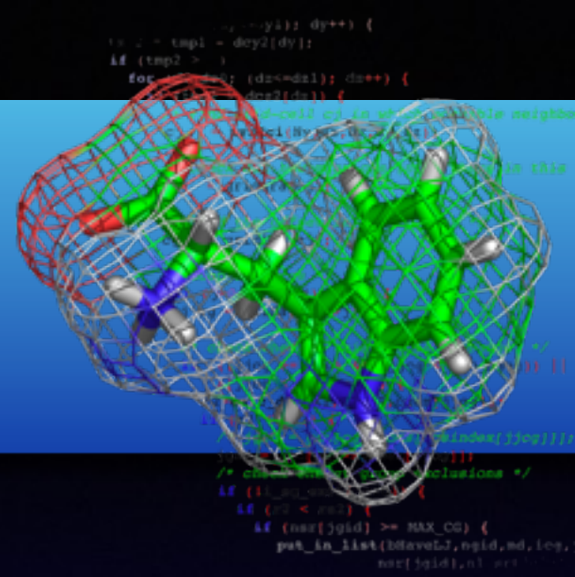


Enthalpy/Entropy



- Chains with limited conformational flexibility can only accommodate few sequences
- Others would have much higher energy
- Chains that can choose between many conformations can accommodate more sequences in low energy states

Boltzmann stats

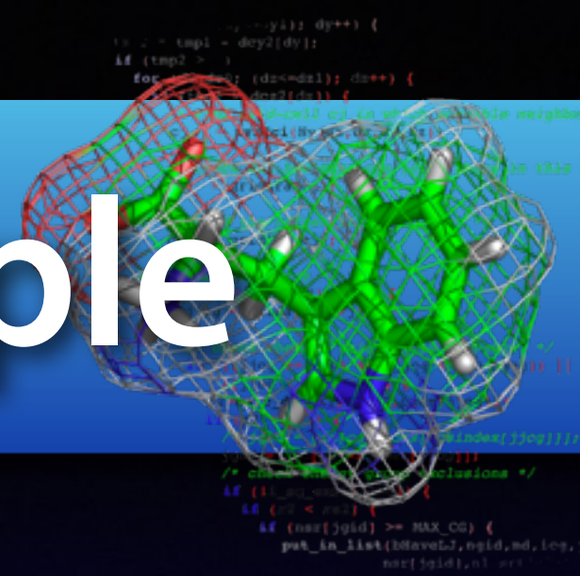


- But we know how to handle this, right?
- Occurence of elements in protein:

$$\rho(r) \propto \exp -\Delta E / kT$$

- Seems to hold up experimentally...
- But it is NOT a Boltzmann distribution!
- Here, the structure is constant, but the question is why many sequences fit it!

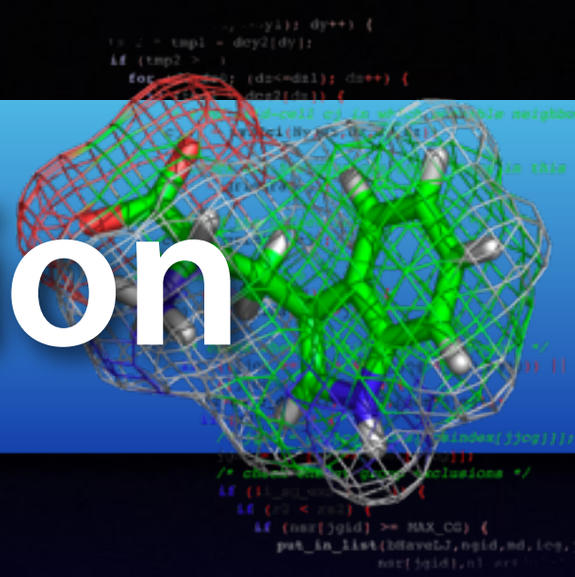
The multitude principle



"The more sequences that can fit a given architecture without disturbing its stability, the higher the occurrence of this architecture in native proteins"

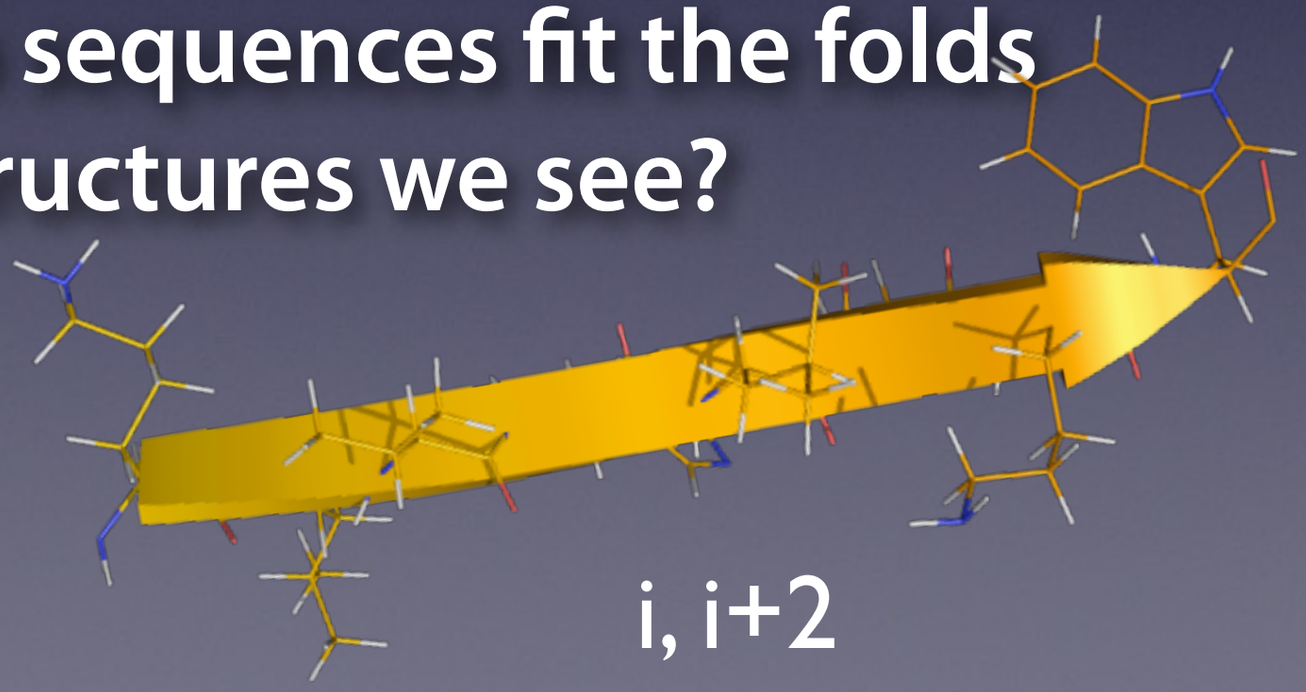
Defective patterns are not impossible, just quite rare!

Sequence stabilization



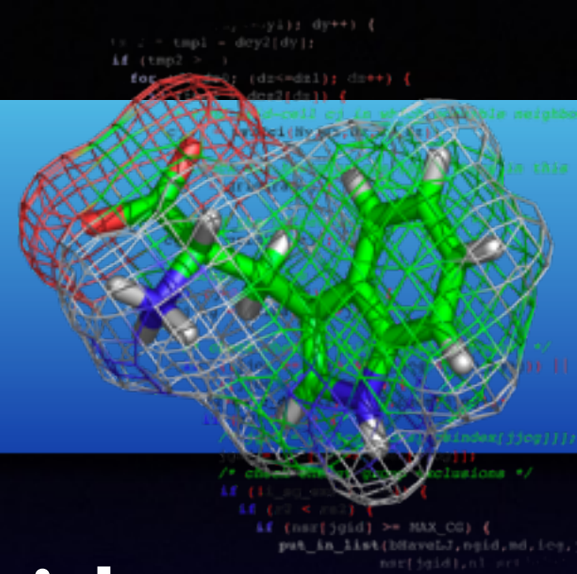
- Limited number of folds for globular proteins
- Approximately equal fractions of hydrophobic/hydrophilic residues (DNA)
- How well do such sequences fit the folds and secondary structures we see?

i, i+3 OR i, i+4



i, i+2

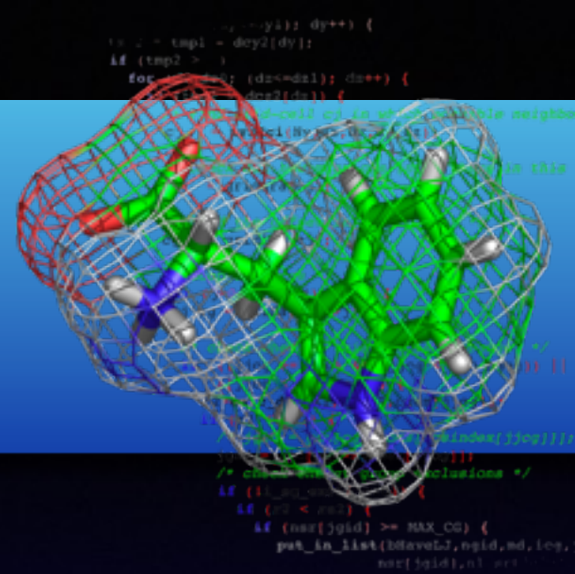
Segment stability



- Let p be the fraction non-polar residues in the sequence
- What is the average number of such groups we will find in a stretch?
- Probability of *r* such groups in a stretch:

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

Segment stability



- **Weighted average:**

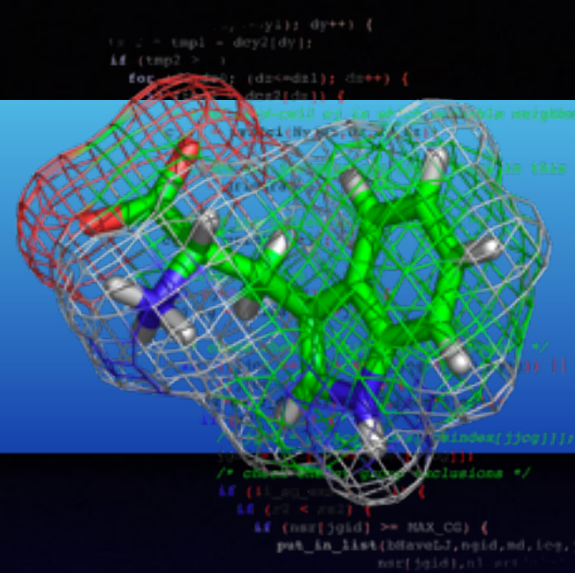
$$\langle r \rangle = \frac{\sum_{r\geq 2}[W(r)r]}{\sum_{r>2} W(r)} = \frac{\sum_{r\geq 2} r p^r}{\sum_{r>2} p^r}$$

$$\sum_{r=1}^{n} p^r = \frac{p(1-p^n)}{1-p}$$

$$\langle r \rangle = 2 + \frac{p}{1-p}$$

about 3 for p=0.5!

Helix/sheet length

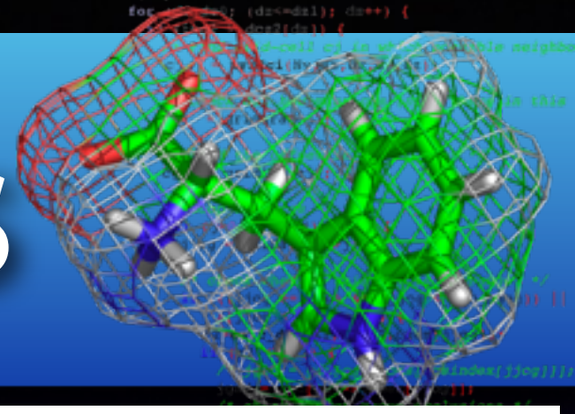


- 3 units of the typical repeat?
- Alpha helix: 3*3.6 = 11 residues
- Beta sheet: 3*2 = 6 residues
- Fits quite well with observed lengths!
- Similarly, average loop length:

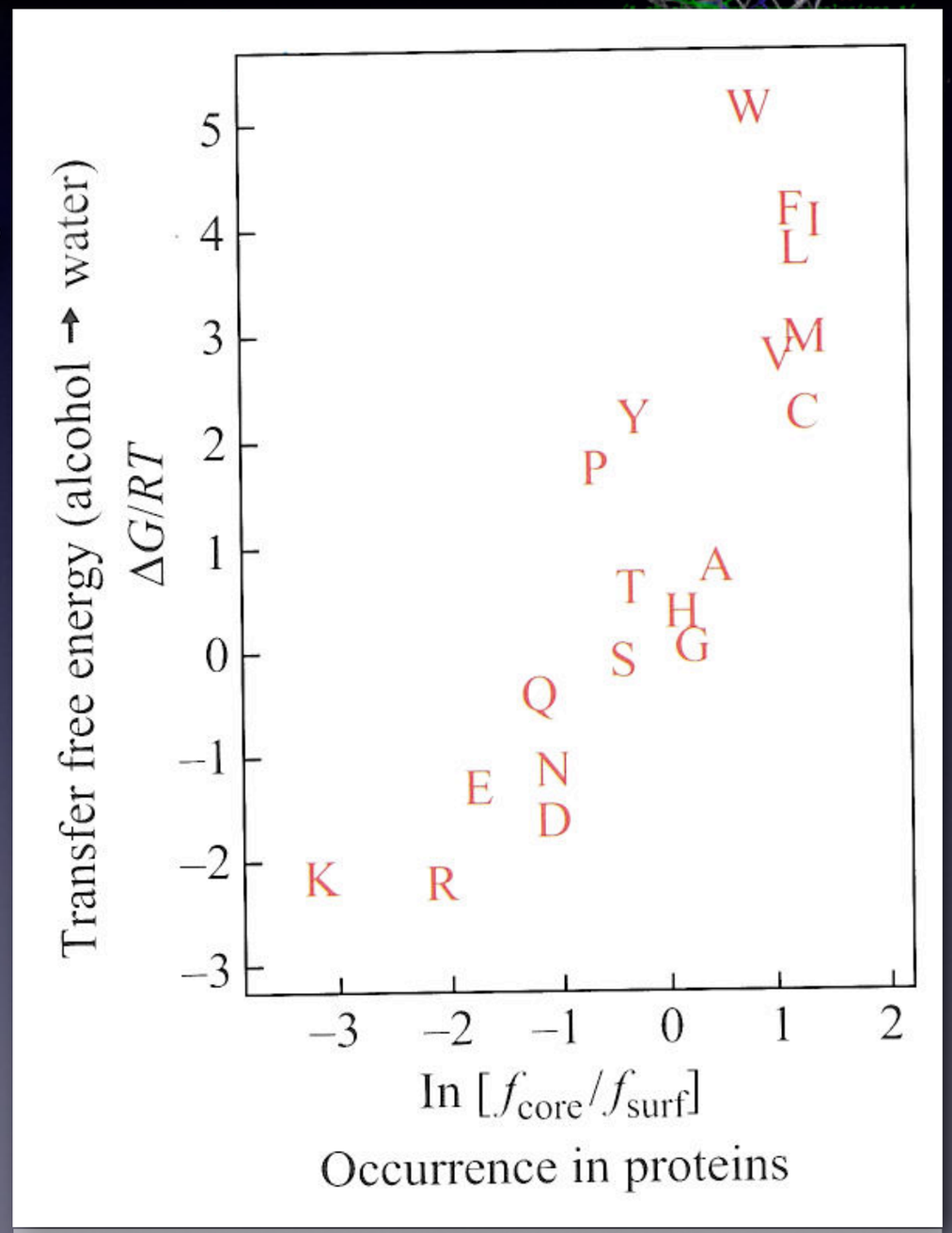
$$\langle r \rangle = 3 + \frac{1}{2p^2}$$

- Even random sequences can form 1 layer!

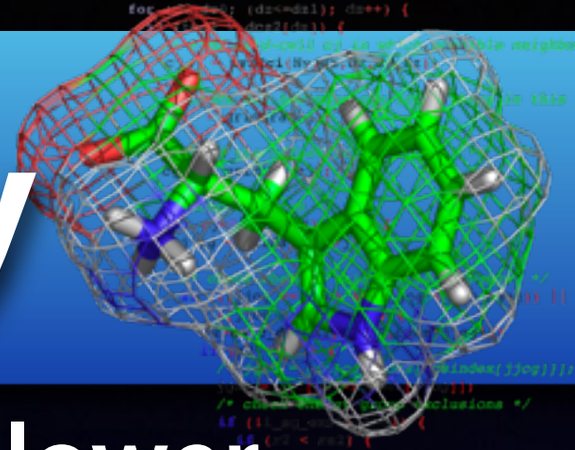
Stability energetics



- Why are energy defects of ~1kcal important for stability?
- What does it have to do with a Boltzmann distribution?
- hydrophobic/hydrophilic residue distribution in structures obey it reasonably well too!?

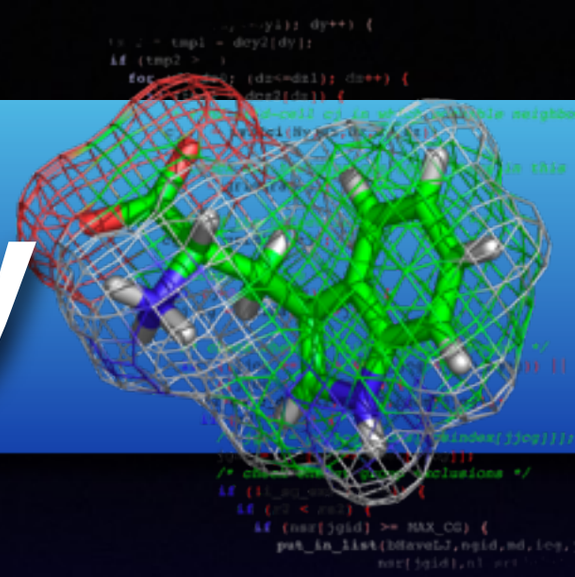


Native fold stability



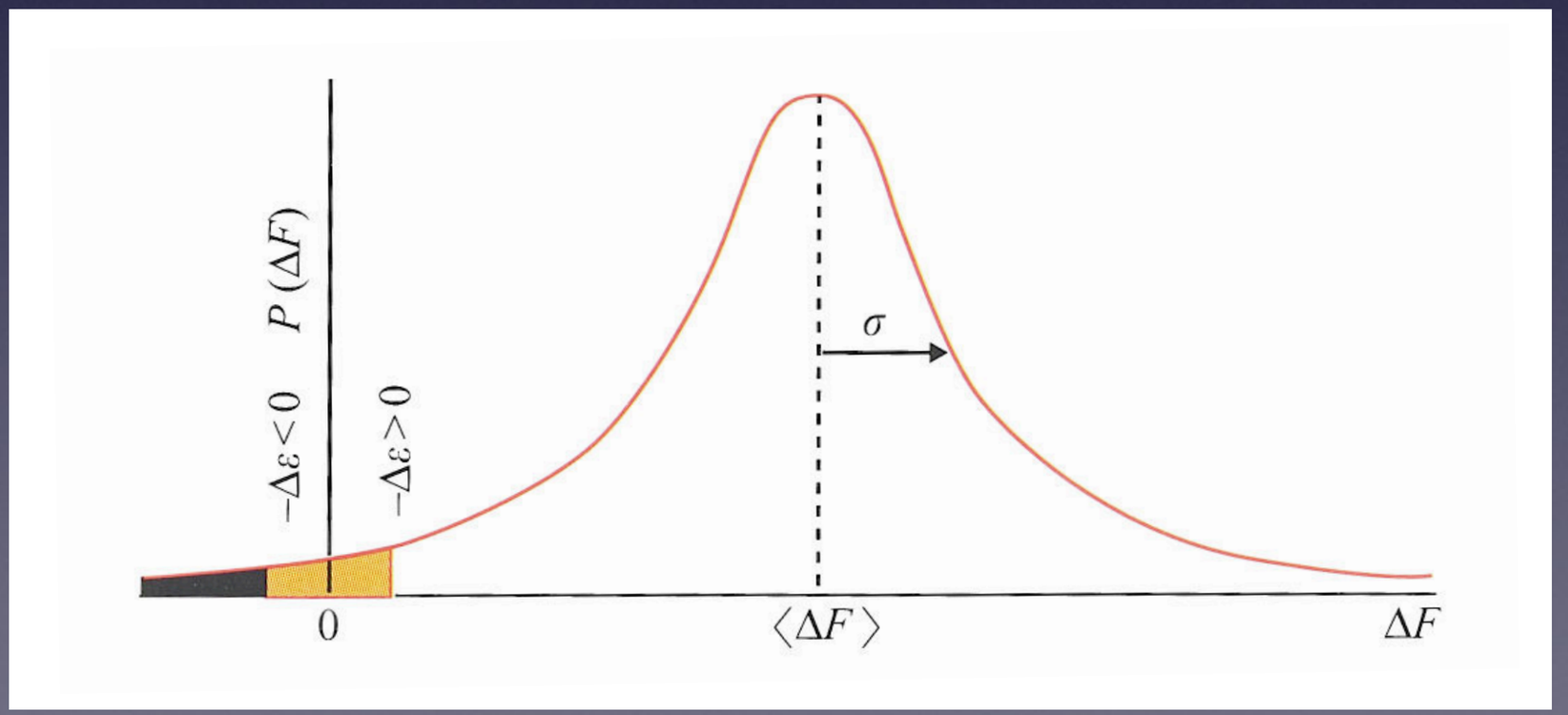
- Native state is stable if free energy is lower (by kT) than for all other states
- Consider Ser <-> Leu mutations
- Transfer from oil (protein inside) to water:
  - Ser: Δε=0 kcal/mol Leu: Δϵ=+2kcal/mol
- Fold with Ser inside also works with Leu
- But fold with Leu works for more seqs!
- Rest of chain: ΔF Total: ΔF+Δε

Native fold stability

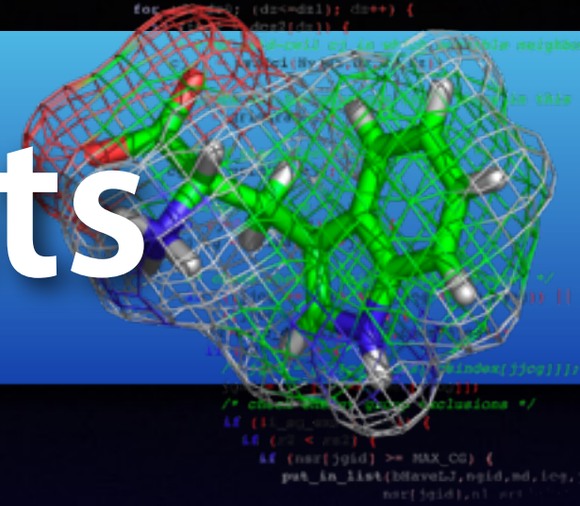


- Stable fold if ΔF < -Δε :

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



Quasi-Boltzmann stats



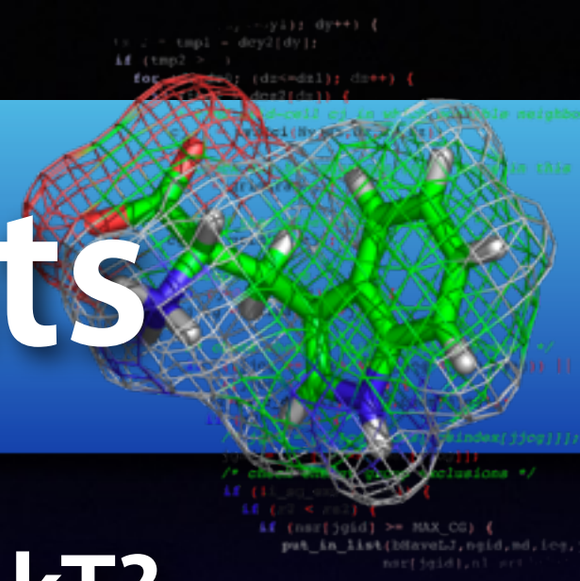
- **Stable fold if ΔF < -Δε :**

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

$$\approx C\exp\left[-\frac{\Delta\varepsilon}{\sigma^2/\langle\Delta F\rangle}\right]$$

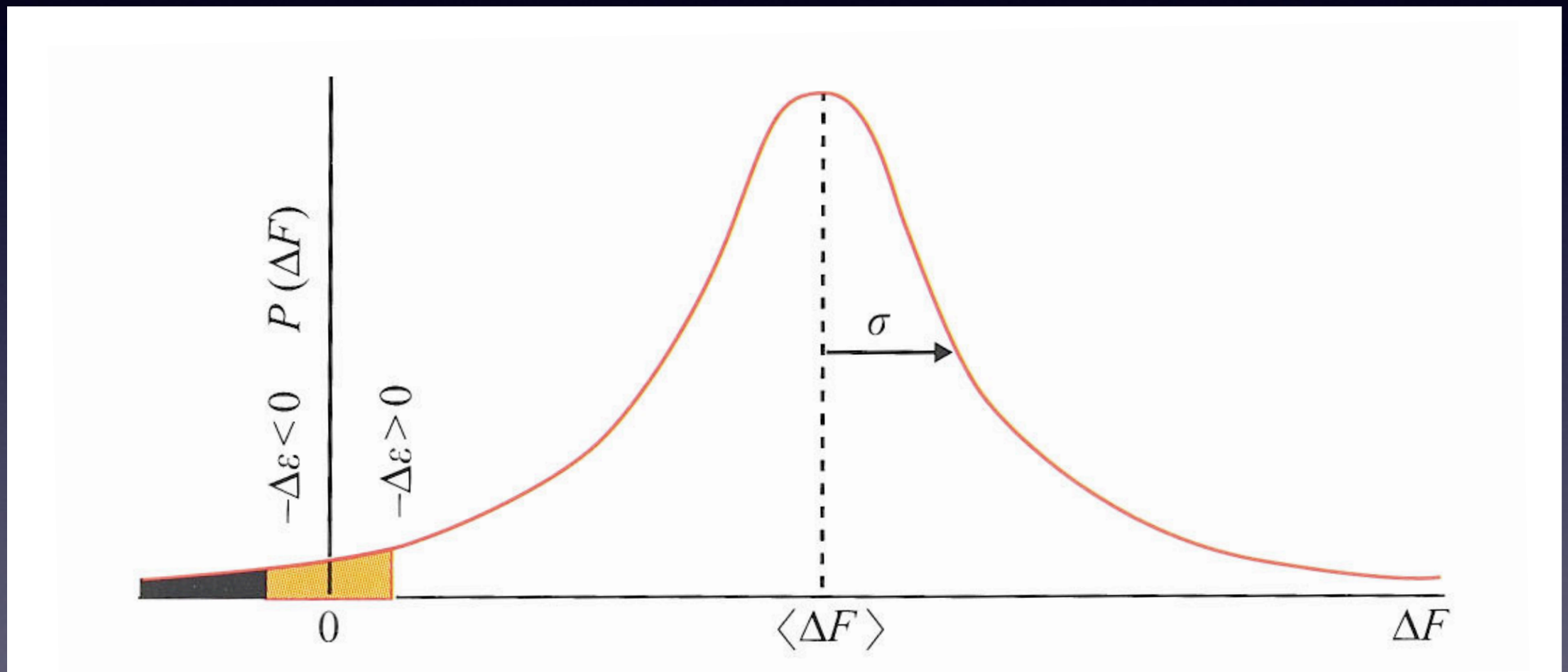
Note the similarity to the Boltzmann distribution!
Increasing Δε reduces the number of stabilizing sequences exponentially

Quasi-Boltzmann stats



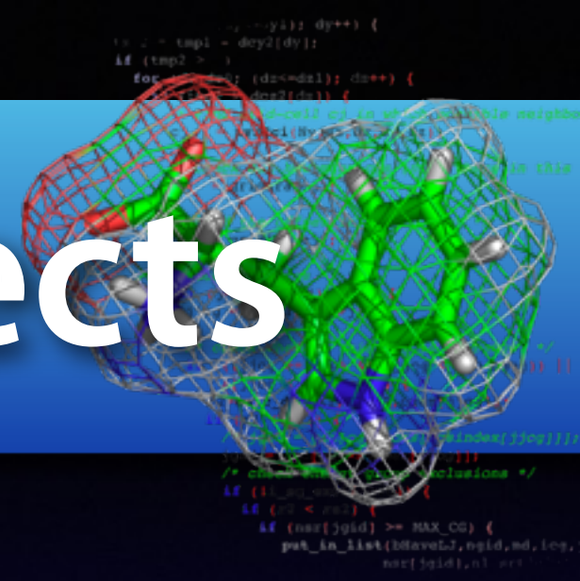
- What does $\sigma^2/\langle F \rangle$ mean rather than kT?
- Both σ^2 and $\langle F \rangle$ are proportional to size
  - The quotient is *size-independent*
- Thus: protein stabilization energy is not dependent on the size of the protein!
- Chain energy or "characteristic energy"
- Think of it as kT_C, with T_C around 350K
- Energy defects should be compared to kT_C rather than the entire protein energy!

Good vs. bad sequences



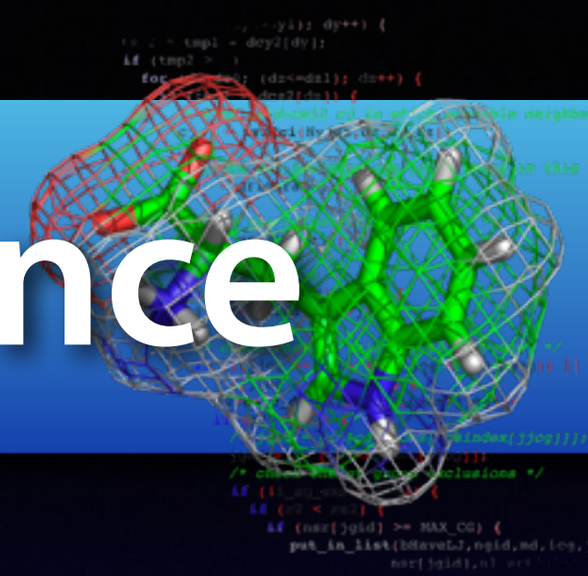
Most sequences do not fold into stable structures!

Entropic packing effects

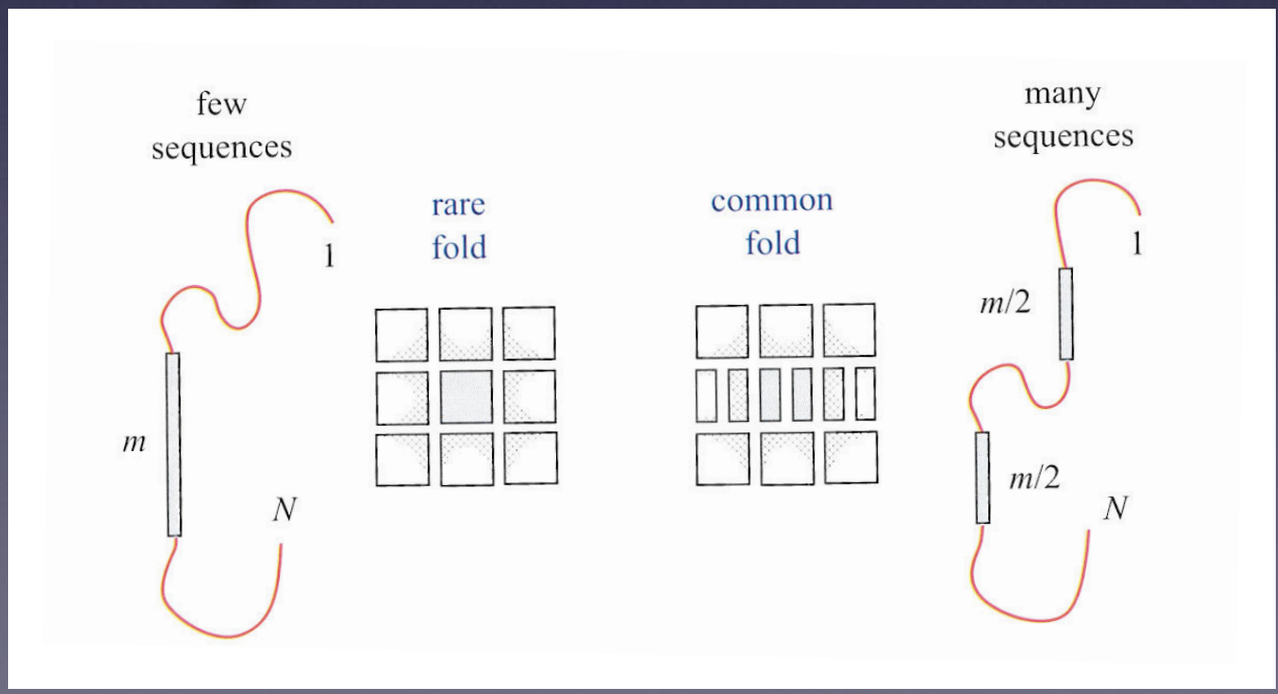


- Example: Left- vs. right-handed sheets
- Structures with more conformational freedom can accommodate more sequences
- Higher density of these states in P(ΔF) means they will be more likely to appear in stable folds
- Same quasi-Boltzmann effect as for the energy distribution before!

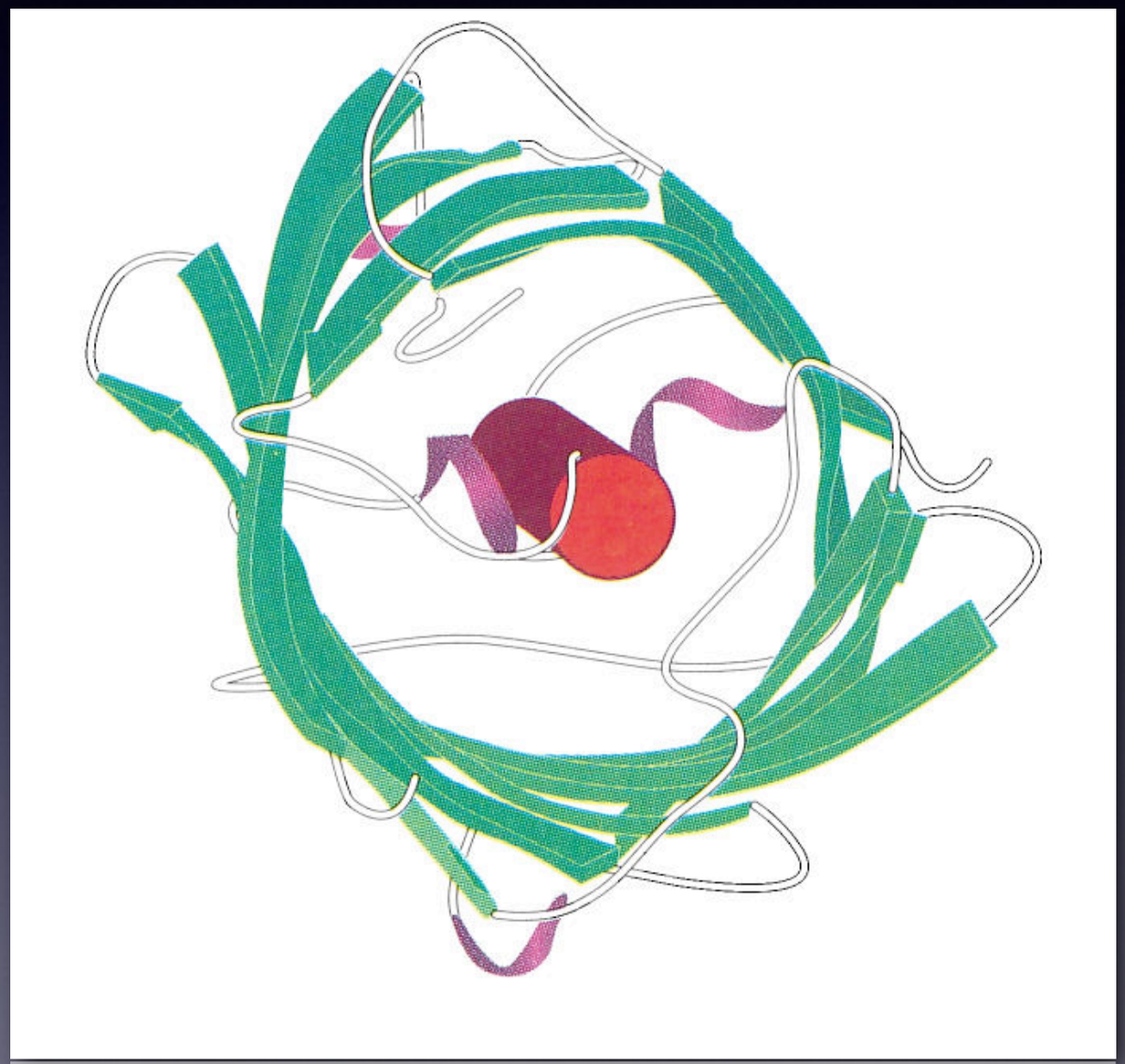
Helix/sheet occurence

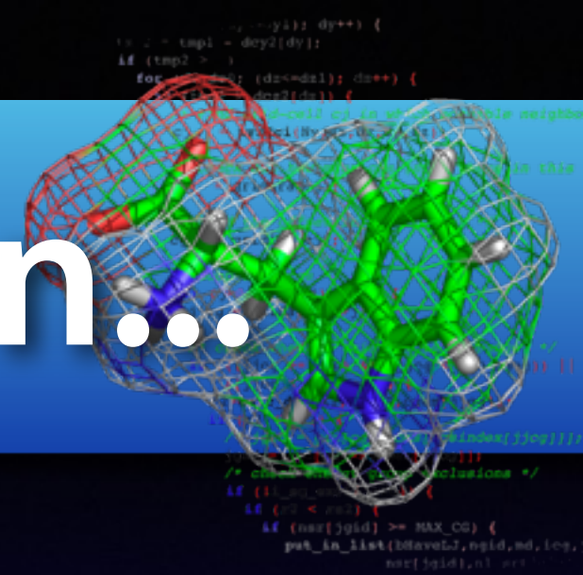


- Which is more common in the protein interior, sheets or helices?
- Sheet: n residues per length
- Helix: 2n residues per length
- Interior must be hydrophobic
- Many more ways to place two small blocks inside!

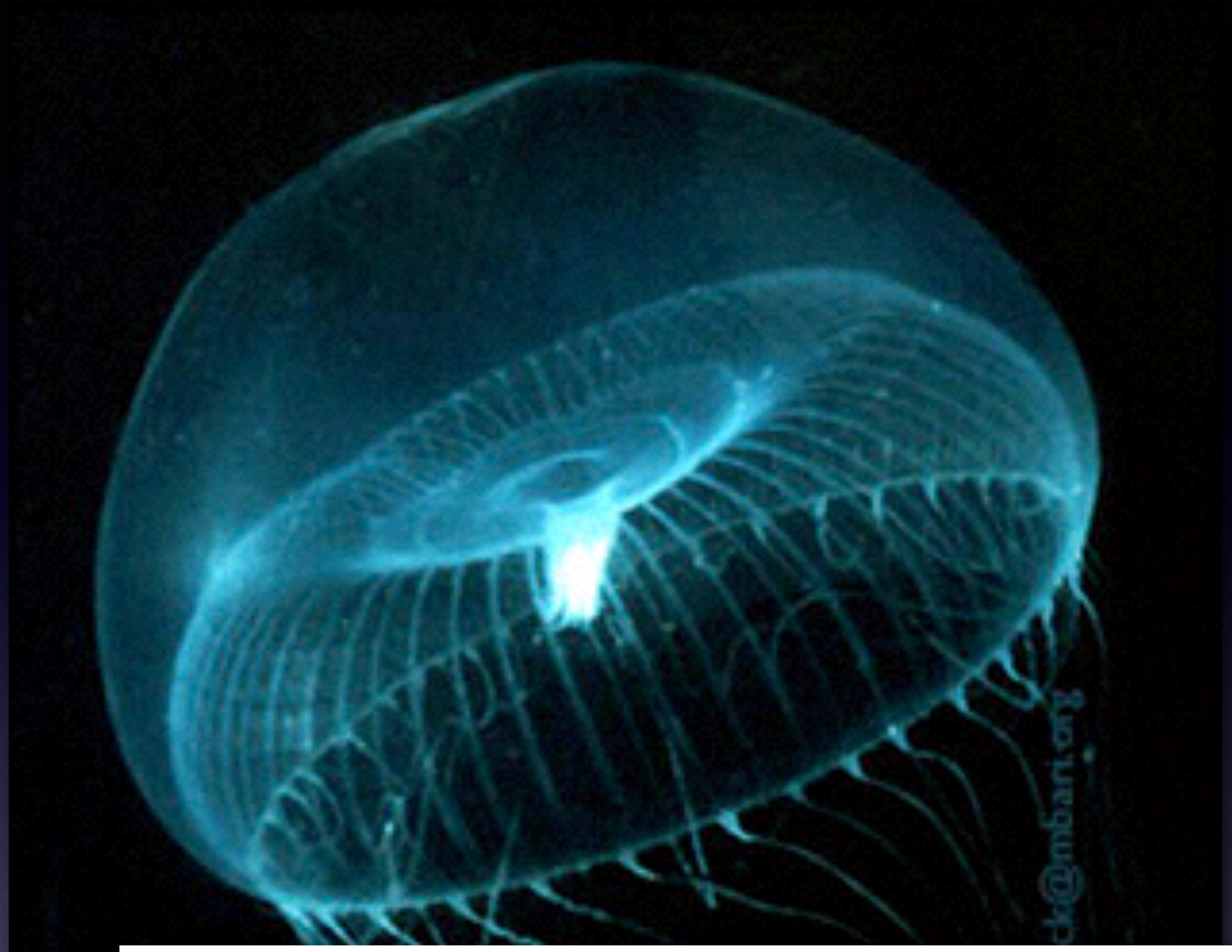


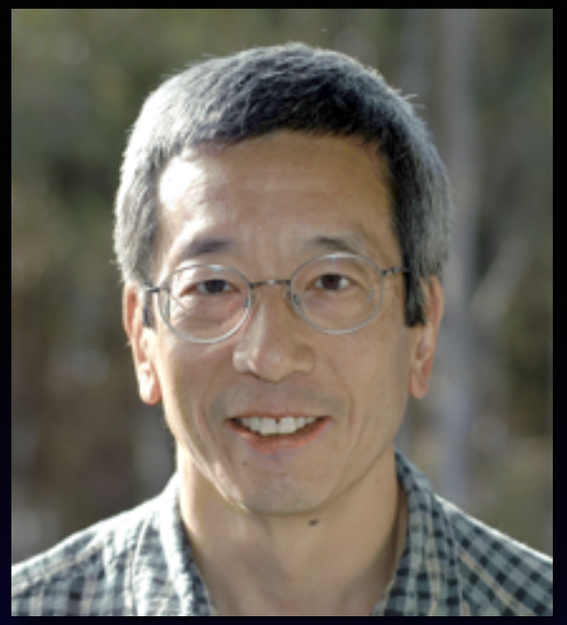
GFP is an exception...





Green Fluorescent Protein





Fluorescent peptides highlight peripheral nerves during surgery in mice

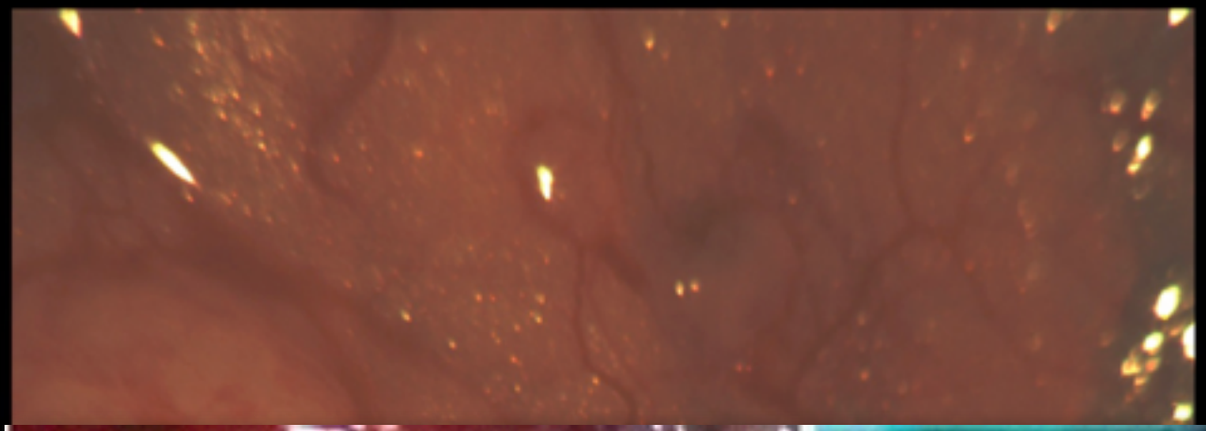
Michael A Whitney, Jessica L Crisp, Linda T Nguyen, Beth Friedman, Larry A Gross, Paul Steinbach, Roger Y Tsien & Quyen T Nguyen

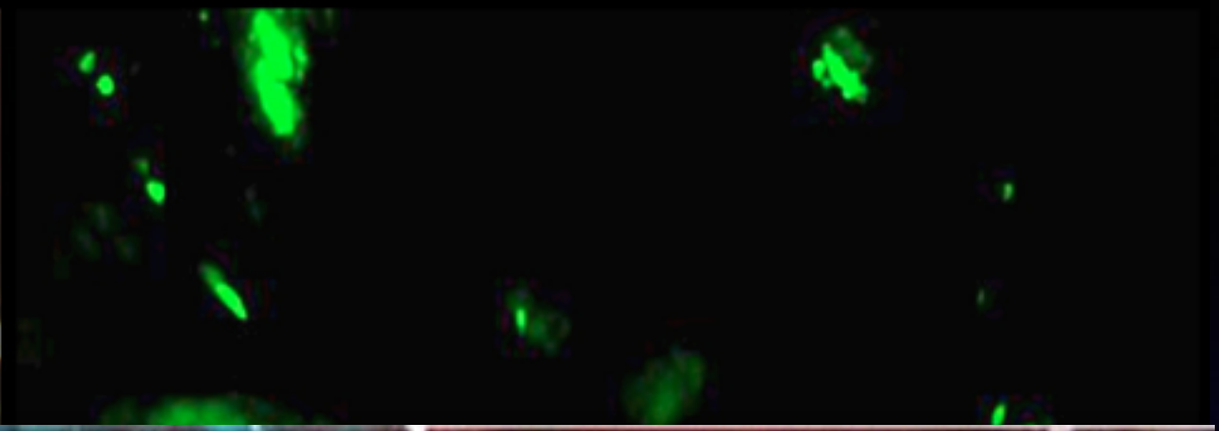
Affiliations | Contributions | Corresponding author

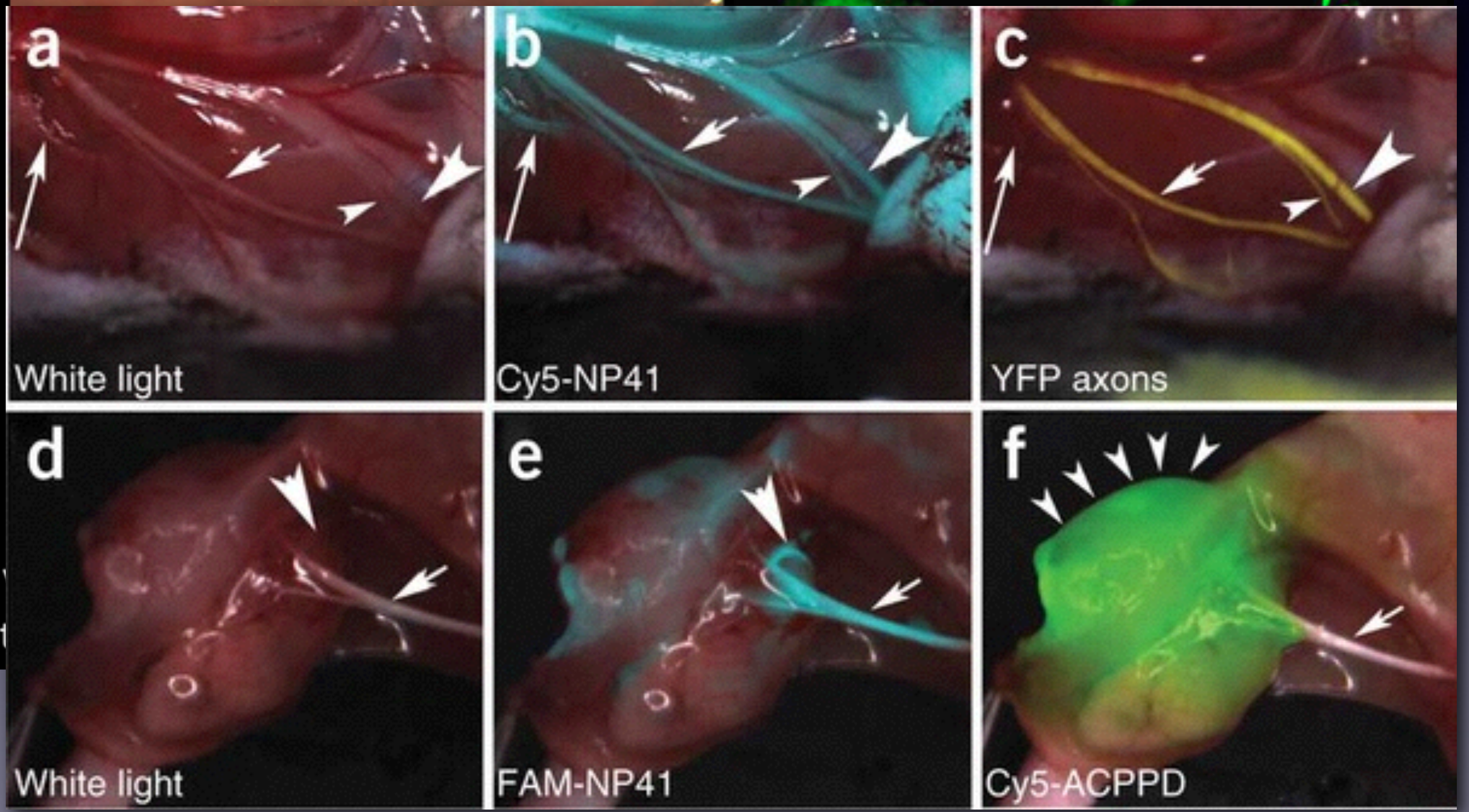
Nature Biotechnology **29**, 352–356 (2011) | doi:10.1038/nbt.1764
Received 03 August 2010 | Accepted 04 January 2011 | Published online 06 February 2011

Surgeon's former view

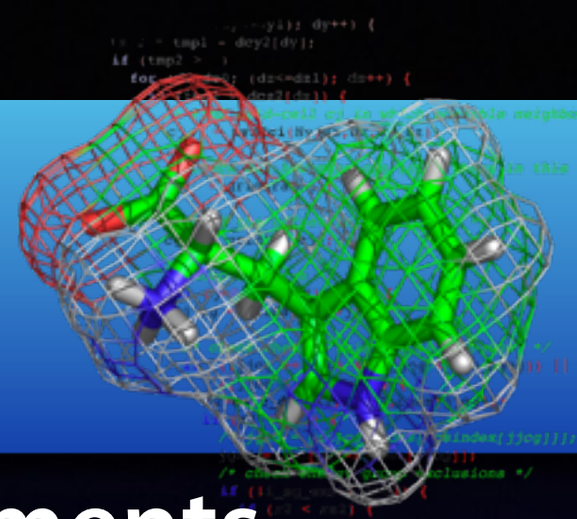
Surgeon's new view







Summary

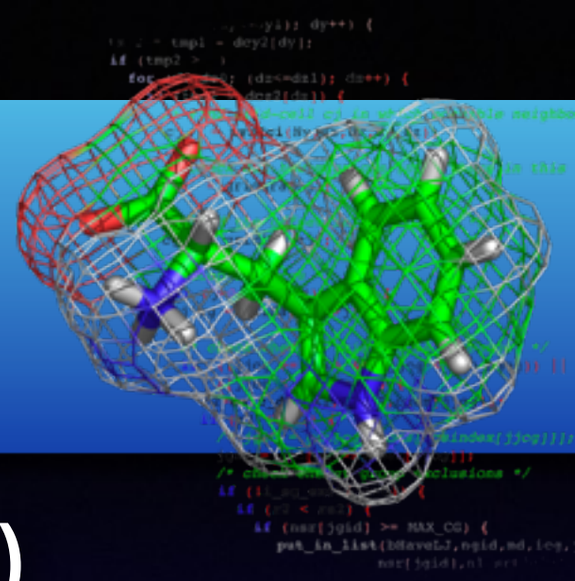


Probability of observing structural elements in randomly created stable globules depends on the amount of sequences that stabilize the fold:

$$\rho(r) \propto \exp -\Delta G / kT_C$$

This is not because of the Boltzmann distribution (no equilibrium), but it has the same shape and a typical temperature.

Summary



- Structure classification (SCOP, CATH)
- Structural evolution
- Size of helices/sheets
- Sequence-structure compatibility
- Protein folds are stabilized by only tens of kcal/mol, regardless of size
- Compare to characteristic energy kT_C
- It will be very hard to design *de novo* folds
- Read chapters 15 & 16!