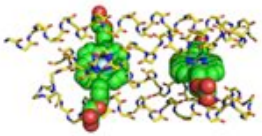
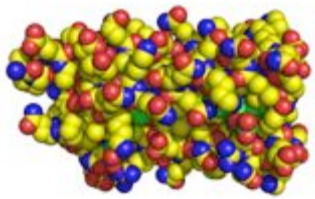
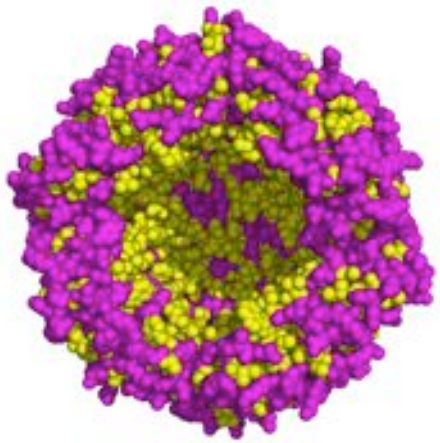
Theoretical approaches to designing and understanding proteins

Jeffery G. Saven

University of Pennsylvania
Department of Chemistry

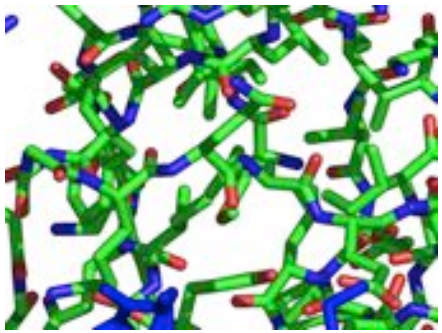





Design Challenges of Beratan

1. How to explore—*and characterize*—chemical space?
2. Discovering structures with excellent properties
3. How to capitalize on iterating theory and experiment?
4. Metrics for diversity?
5. Geometry—*and energetics*—of underlying design landscape

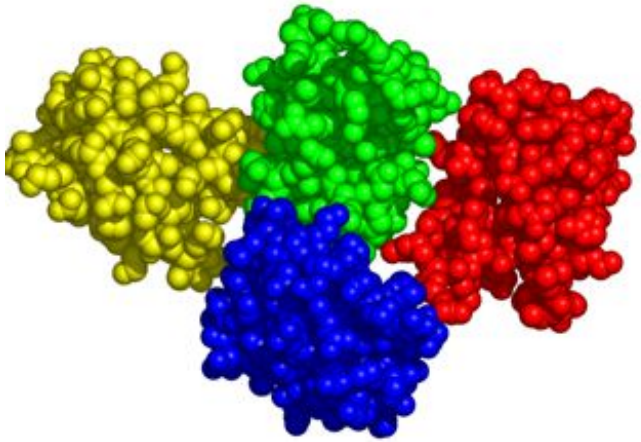
Proteins: many length scales and functions

- Multiple environments: solution, membranes, surfaces…
- Many functionalities
  - Solution: catalysis, recognition, signals, pigments…
  - Membranes: channels, energy transduction, light harvesting, signaling…



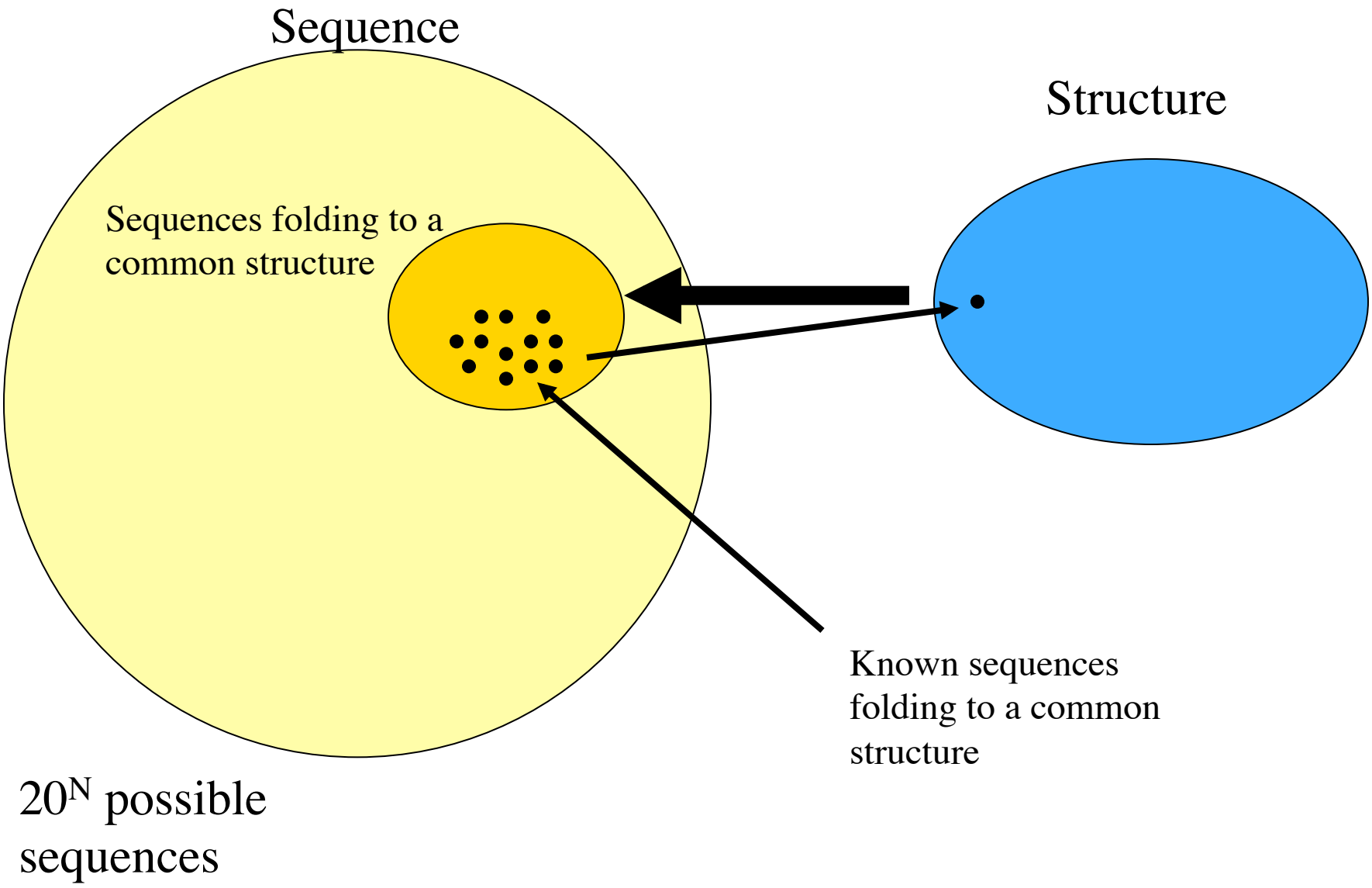
Å

nm



> 10 nm

Sequence degeneracy

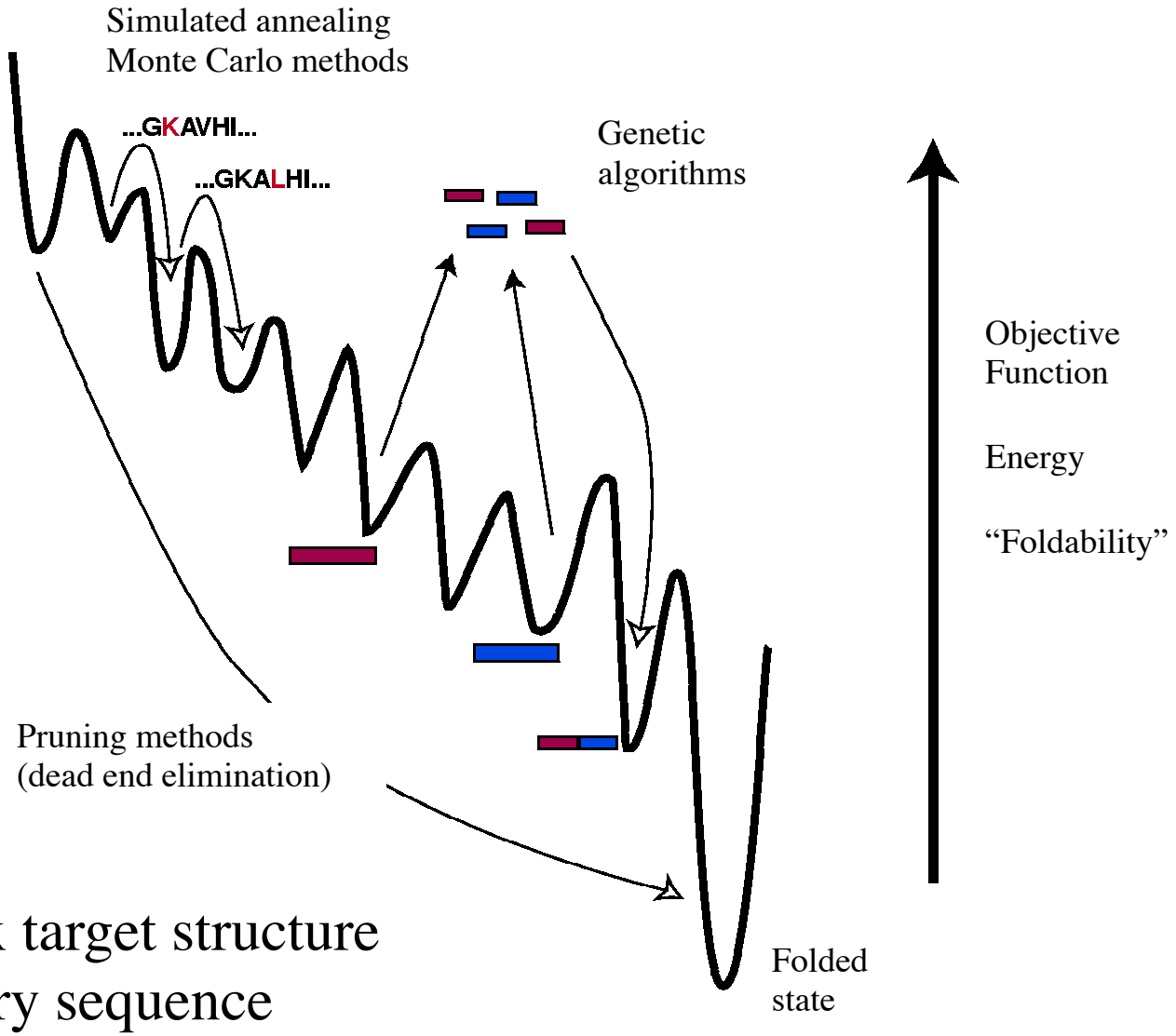


Elements of protein design

- Template structures:
  - redesign of existing structure
  - novel structure
- Monomer (residue) degrees of freedom
  - Backbone structure
  - amino acids (identity)
  - amino acid conformations (rotamers)
- Energy functions: quantifying interactions within a structure
- Folding criteria and negative design
  - Energy based: sequence structure compatibility
  - Competing structures
- Characterizing sequences
  - Searching sequence space with optimization based methods
  - Estimating the frequencies of the amino acids

A. Lehmann, et al, in *Protein Folding and Misfolding* (ed. V. Munoz; Royal Society of Chemistry), 2008. S. Park et al, *Annual Reports of the Royal Society of Chemistry, Section C, (Physical* Chemistry), 2004, pp. 195-236. J. G. Saven, *Chemical Reviews* 2001. 101 (10): 3113-3130. Samish et al, *Annu. Rev. Phys. Chem* 2011.

Sequence design: search methods



Fix target structure
Vary sequence

S. Mayo (Caltech), H. Hellinga (Duke); D. Baker (UW Seattle); T. Alber (UC Berkeley), P. Kim, B. Tidor, A. Keating (MIT); P. Harbury (Stanford); J. Desjarlais (Xencor); C. Floudas (Princeton); L. Lai (Beijing); S. Takada (Kyoto)…

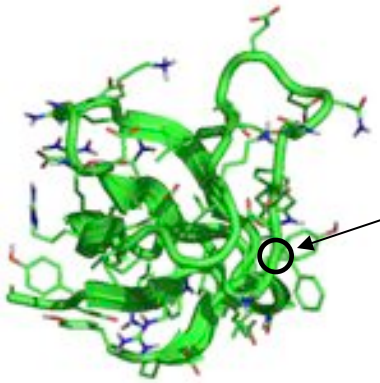
Difficulties protein design

- Large numbers of degrees of freedom
  - 20 amino acids, most have multiple side-chain conformations: 100's of states per residue position: $(20 \times 100)^{100}$ configurations
  - Calculations can become demanding even for small proteins
- Imperfect energy functions
- "Imperfect" structures
- Optimization?
  - Natural proteins are marginally stable
  - Biological function may compete with stability
  - Scoring functions are approximate
  - Sequence variability
- Incomplete information guides directed design of protein sequences
  - Potentials contain (partial) information about stabilizing forces: van der Waals interactions, electrostatics, structural propensities,...
  - A *probabilistic approach* seems appropriate to broadly understand sequences consistent with a chosen structure…
  - …and combinatorial experiments (10^7 sequences) …

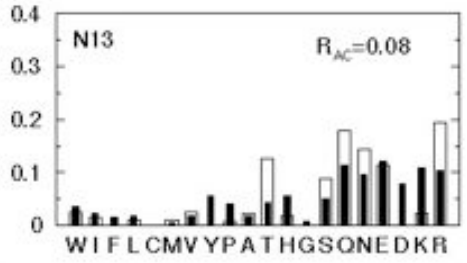
Theory and Design of Proteins (and Self-Organizing Macromolecules)

- Methods for probabilistic protein design
  - Input:
    - –Target tertiary and quaternary structure
    - –Features, e.g., well-packed, hydrophobic interior
    - –Atomistic energy functions
    - –Physical, synthetic and functional constraints on sequences
  - Output: Site-specific probabilities of the amino acids for a given structure
  - Advantages:
    - Large structures and diversity
    - Application to de novo protein design and combinatorial design
    - Transferable to nonbiological systems

Target structure



Atomistic model of structure and monomers

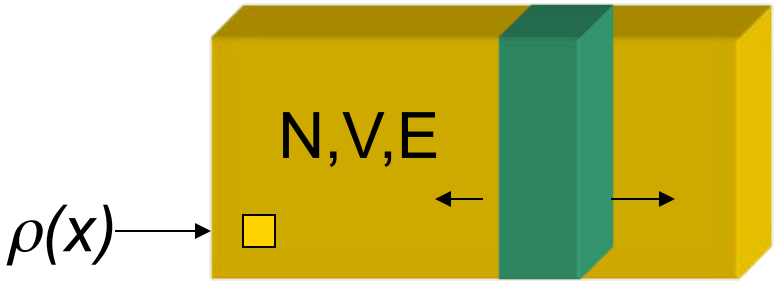


Site-specific monomer probabilities

Thermodynamic analogy

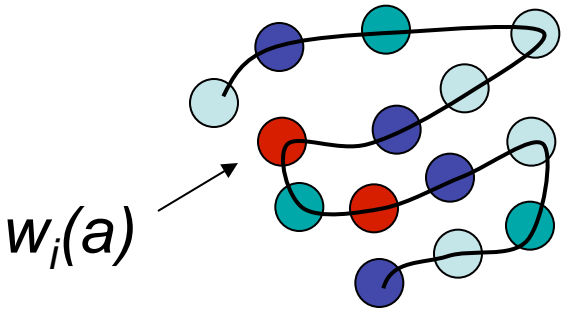
<u>Thermodynamic ensemble</u>

Ω=no. configurations



<u>Sequence ensemble</u>

Ω=no. sequences



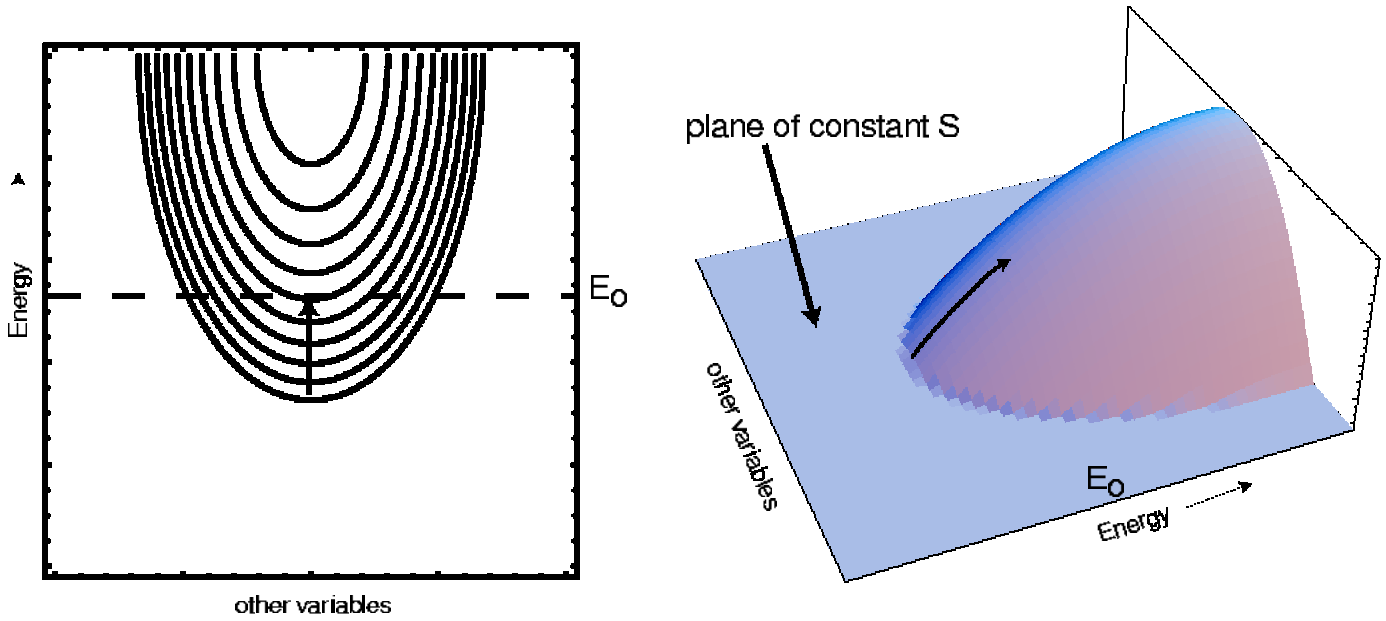
Apply methods from statistical thermodynamics to estimate probabilities (effective thermodynamic quantities: T, E, S…)

- Solve for probabilities $w_i(a)$ subject to constraints on sequences
  - Self-consistent field methods based on entropy maximization

    H. Kono and J. G. Saven. *J Mol. Biol.*,. 306: 607-627 (2001).
    J. Zou and J. G. Saven, *J. Mol. Biol.,* 296: 281-294 (2000).
- Sample sequences and count frequencies of amino acids
  - Efficient (biased with replica exchange) Monte Carlo methods

    X. Yang and J. G. Saven, *Chem. Phys. Lett.*, 401: 205-210 (2005).
    J. Zou and J. G. Saven, *J. Chem. Phys*. 118: p. 3843–3854 (2003).

Entropy maximization

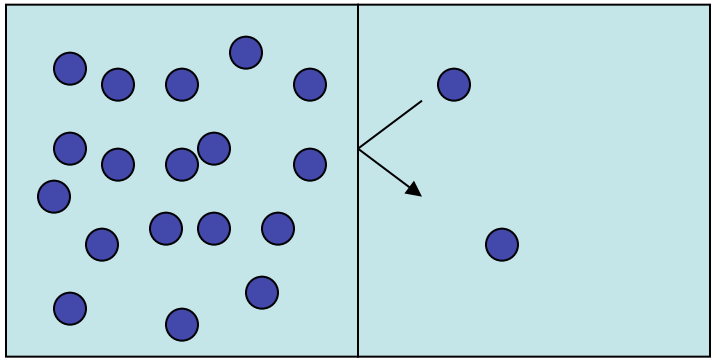


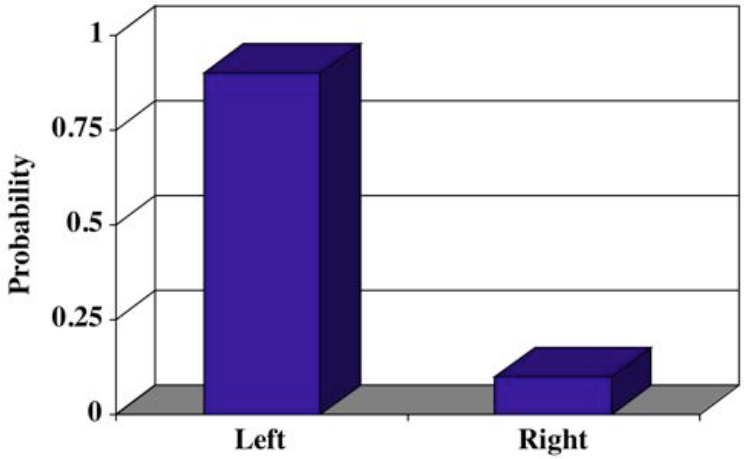
- Sequence ensemble: other variables include amino acid probabilities

Entropy maximization

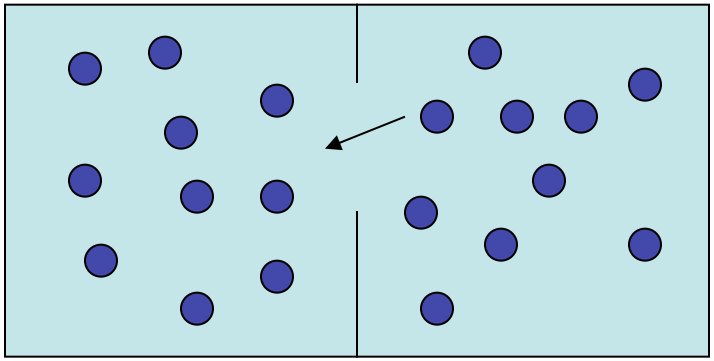
Fix Number, Volume, and Energy

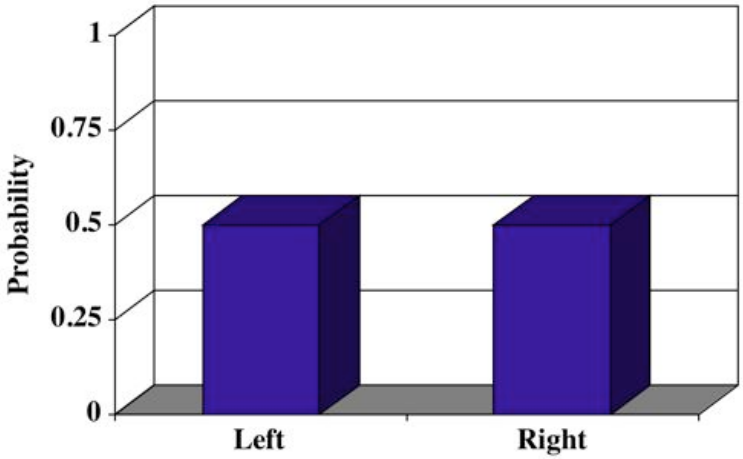
Case 1: permeable wall





Case 2: impermeable wall



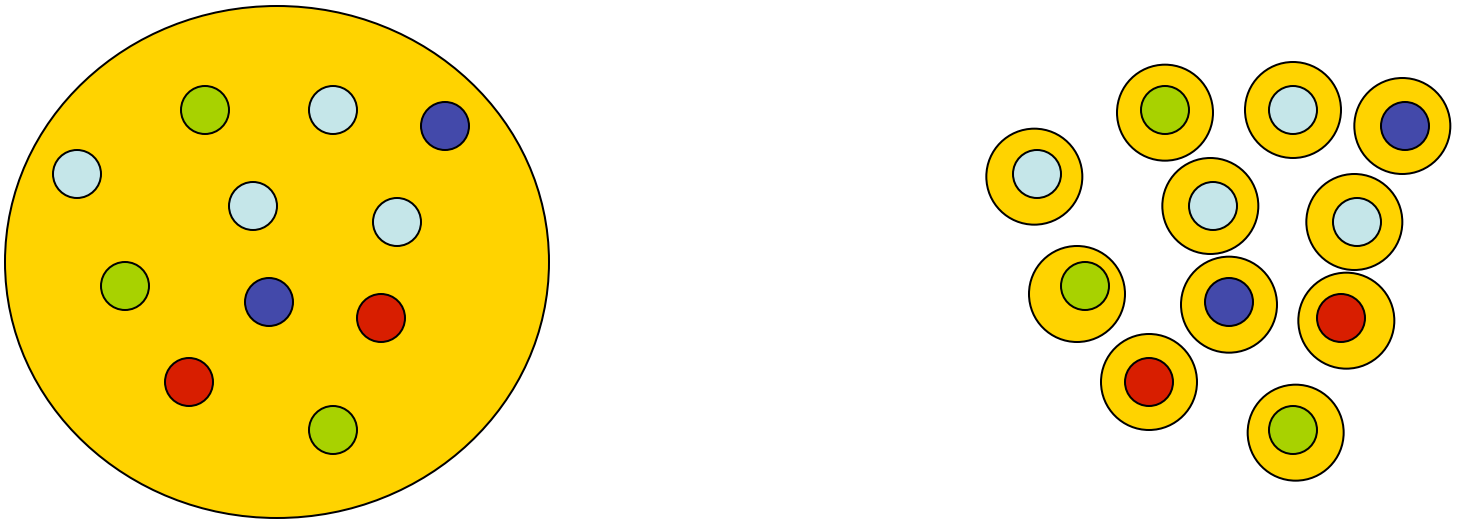


$$S_1(E) < S_2(E)$$
$$\Omega_1(E) < \Omega_2(E)$$

Factorization (Hartree) approximation

$$W(\mathbf{a}) \approx \prod_i w_i(a_i)$$

$$\ln \Omega = S = -\sum W(\mathbf{a}) \ln W(\mathbf{a}) \approx -\sum w_i(a_i) \ln w_i(a_i)$$



…constraints on sequences couple site probabilities

Maximize Effective Entropy

- Maximize sequence entropy for fixed structure

$$S = -\sum_i \sum_a w_i(a) \ln w_i(a)$$

Sum over "sites"

Sum over "states" at each site

- Solve for probabilities $w_i(a)$
  - i = position in sequence
  - a = amino acid "state"
- Maximize subject to constraints on sequences
- Sequences are not enumerated

Maximize entropy subject to constraints

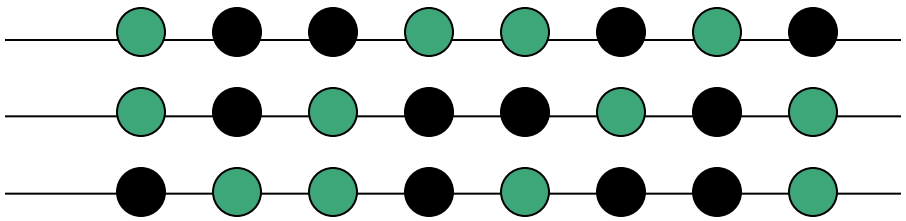
Probabilities are normalized

$$\sum_a w_i(a) = 1$$

Overall Energy of sequence in target structure

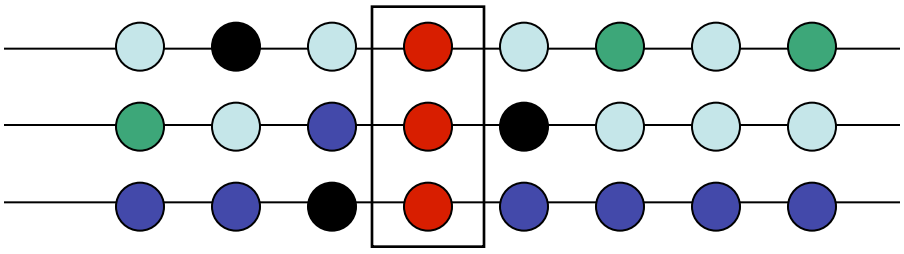
$$E = \langle E \rangle_{seq}$$

Overall composition: number of each amino acid



$$\sum_i w_i(a) = n(a)$$

Patterning constraints



$$w_i(a) = 1$$

Self-consistent, entropy maximization

Sequences are not enumerated

Solve for probabilities $w_i(a)$: a = amino acid state, i = position in sequence

Maximize subject to physical and synthetic constraints on sequences: E_i, f_i

–Constrain effective energies E_i (low energy sequences for target structure)

–Other possible constraints:

- Pattern amino acids: hydrophobic inside, hydrophilic outside
- Specify identities and/or conformations of functionally important residues

$$V(\{w_i(a)\}) = S - \beta_1 E_1 - \beta_2 E_2 - \ldots - \lambda_1 f_1 - \lambda_2 f_2 - \ldots$$

$$S = -\sum_i \sum_a w_i(a) \ln w_i(a)$$

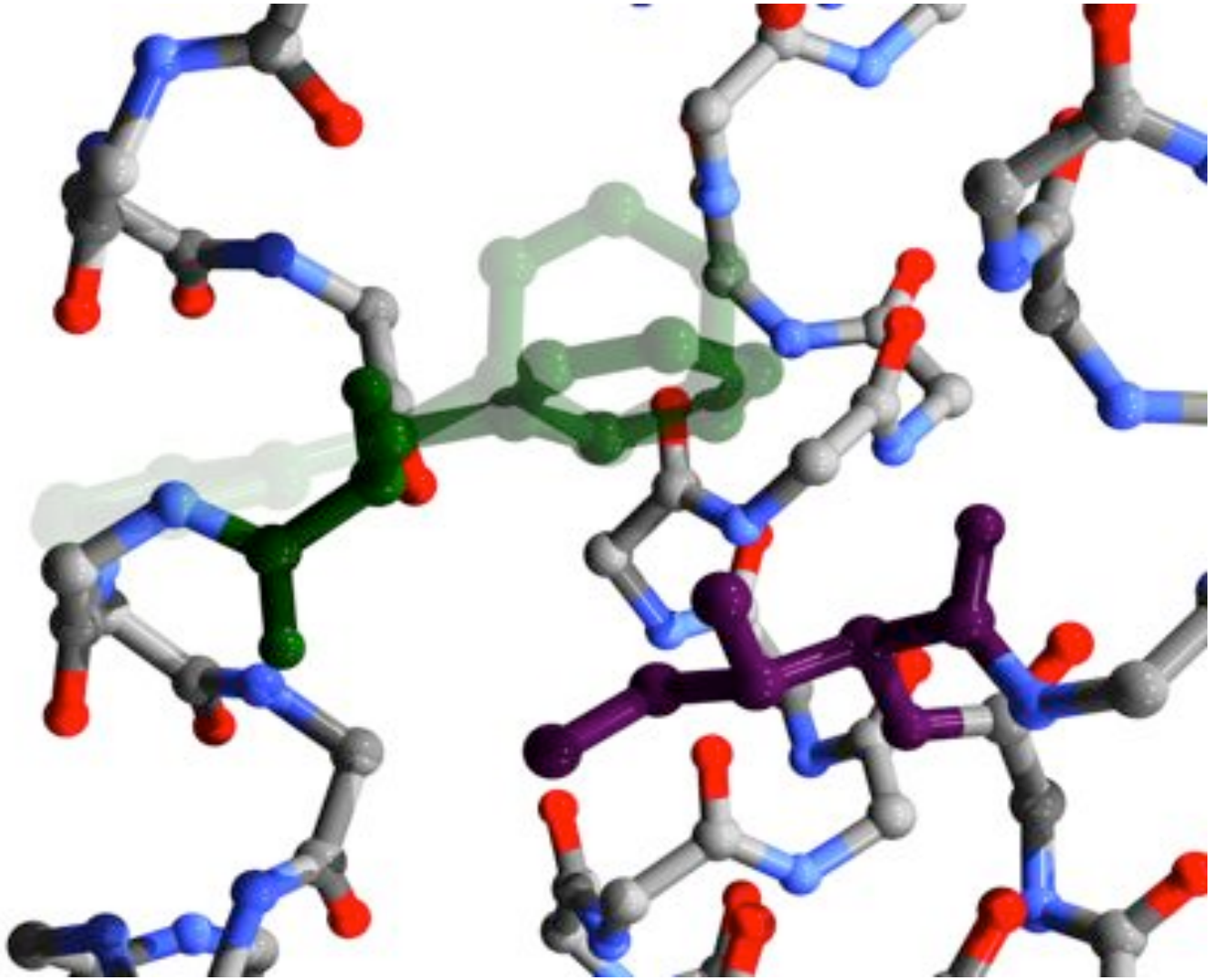
$$E_1 = E_{folded}(\{a\}) - F_{unfolded}(\{a\})$$ Atomic potential energy

$$E_2 = E_{solvation}(\{a\})$$ Solvation energy

$$\frac{\partial V}{\partial w_i(a)} = 0, \text{ and } E_i = \langle E_i \rangle_{sequence}$$

H. Kono and J. G. Saven. *J Mol. Biol.*,. 306: 607-627 (2001).
J. Zou and J. G. Saven, *J. Mol. Biol.*, 296: 281-294 (2000).

$$E(a_1,\ldots,a_N) = \sum_i \gamma^{(1)}(a_i) + \sum_{i<j} \gamma^{(2)}(a_i,a_j)$$



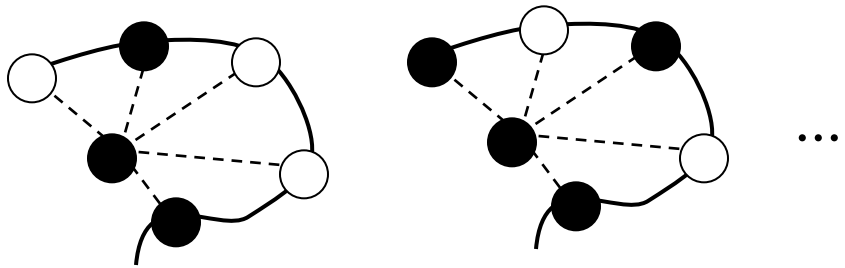
Local average energy

Energy of sequence $(a_1, \ldots, a_N)$ in structure

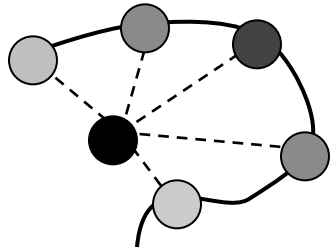
$$E(a_1,\ldots,a_N) = \sum_i \gamma^{(1)}(a_i) + \sum_{i<j} \gamma^{(2)}(a_i,a_j)$$

Local energy of a at site i

$$\varepsilon_i(a) = \gamma^{(1)}(a) + \sum_j \gamma^{(2)}(a,a_j)$$



$$\varepsilon_i(a) \approx \langle \varepsilon_i(a) \rangle = \gamma^{(1)}(a) + \sum_j \sum_{a_j} \gamma^{(2)}(a,a_j) w_j(a_j)$$



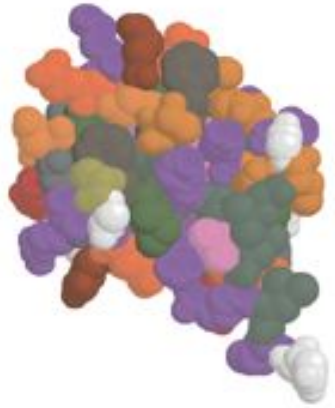
Average over sequences

Atomistic Models of Proteins

- Amino acid and side chain conformation
- "Energy"

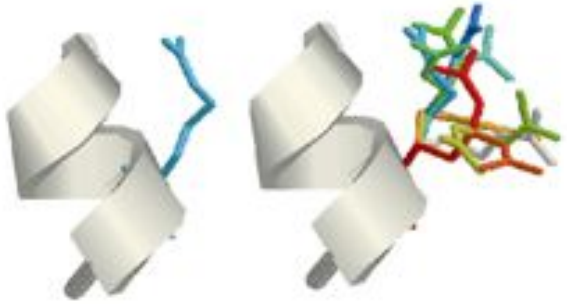
  Atomic interactions (AMBER)
  Solvation (Hydrophobic effect)

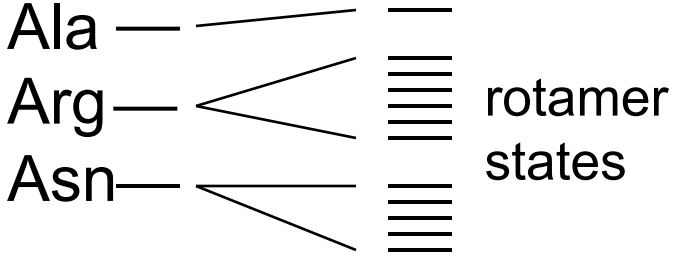
  [Kono & Saven, *J. Mol. Biol,* 306:607 (2001)]



Discrete conformational states for amino acids (rotamers)

Dunbrack & Cohen, *Protein Sci.,* 6:1661 (1997)

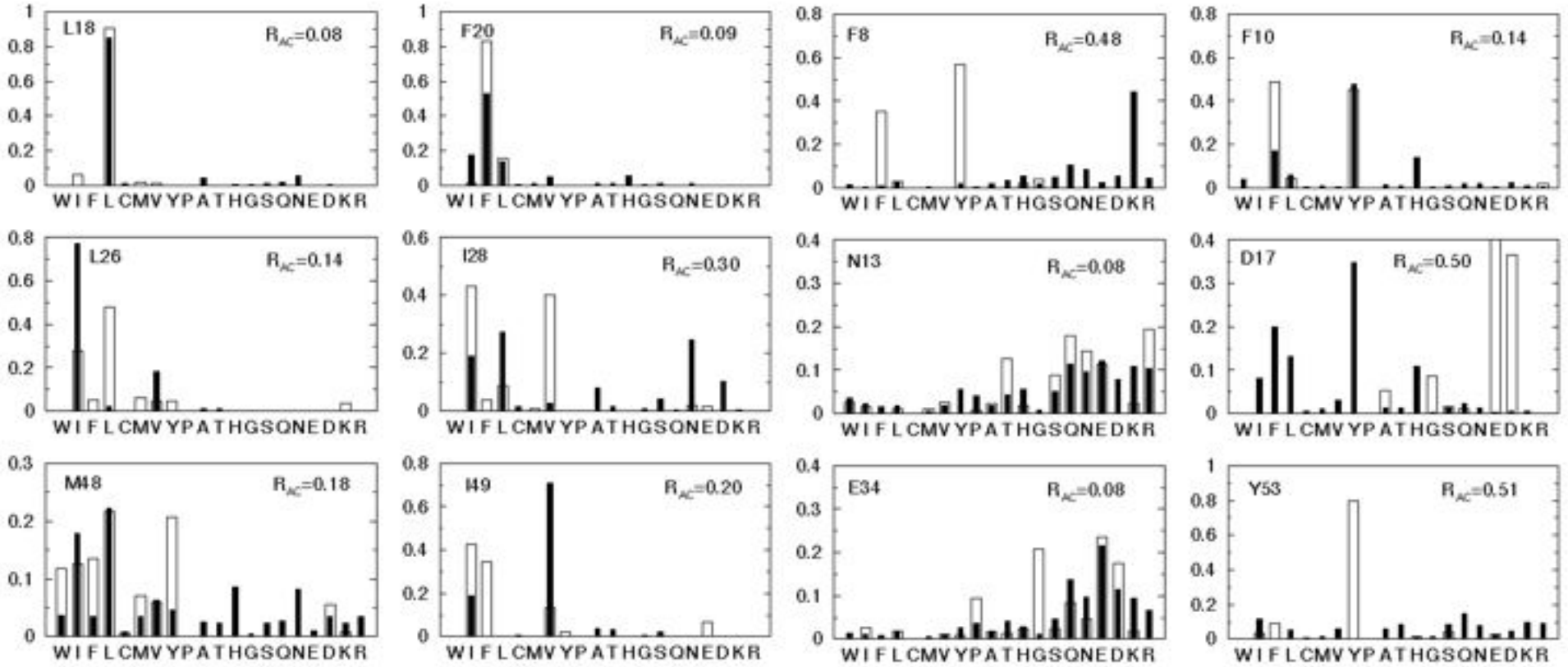




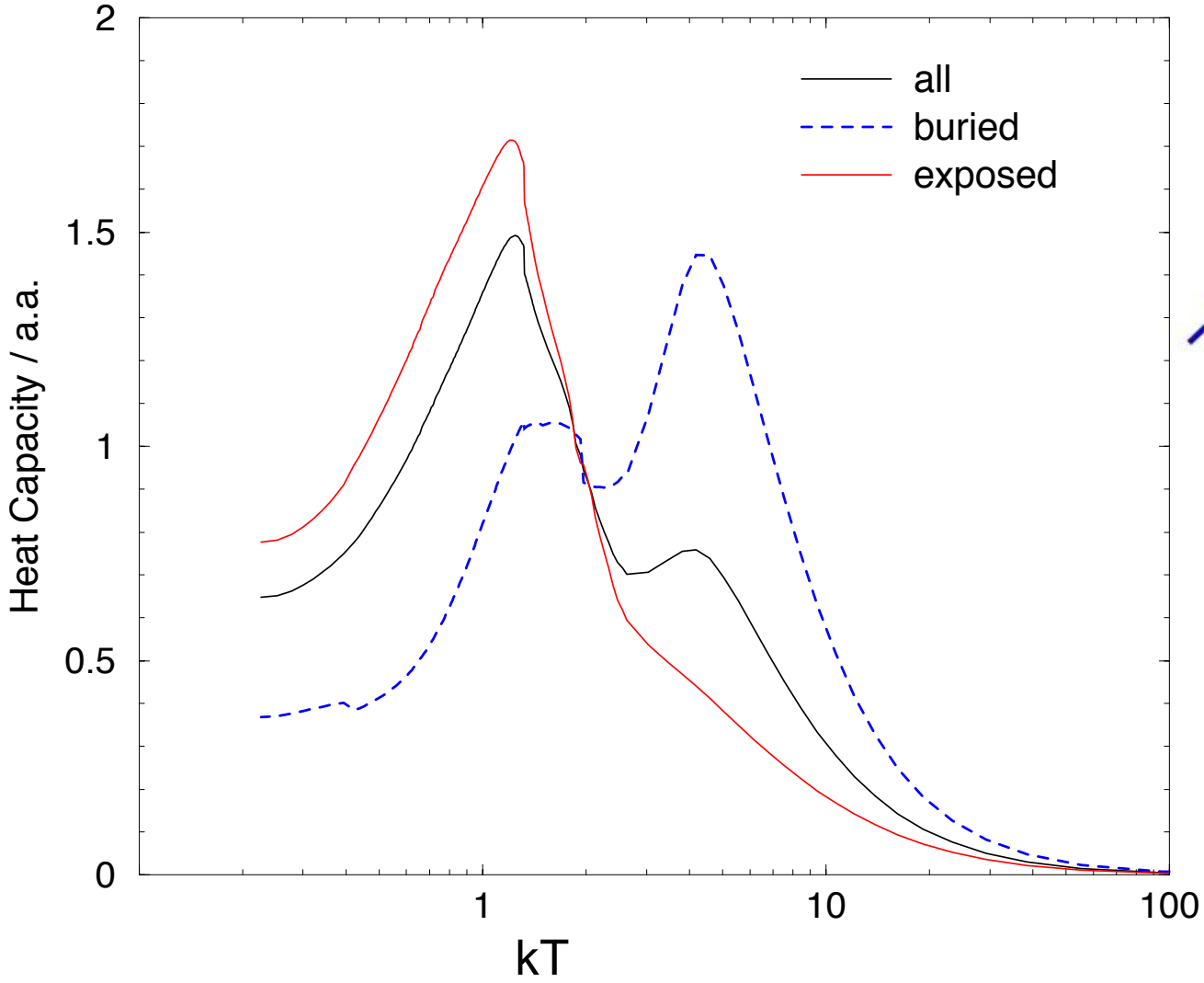
$$w_i(a) = \sum_{r^a} w_i(a, r^a)$$

Relative entropy: SH3 domain

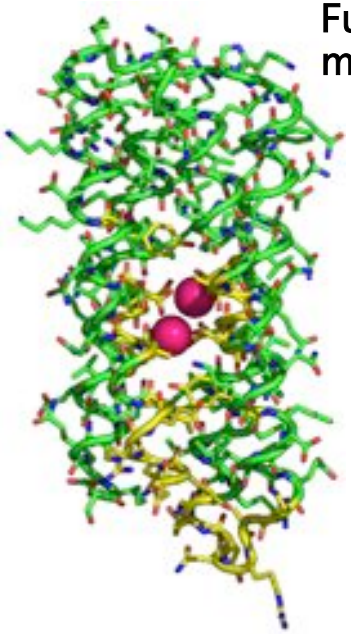
- 57 residue protein
-Allow all amino acids at each position
-Compare with multiple sequence alignment



Effective heat capacity C_v plotted against effective temperature T

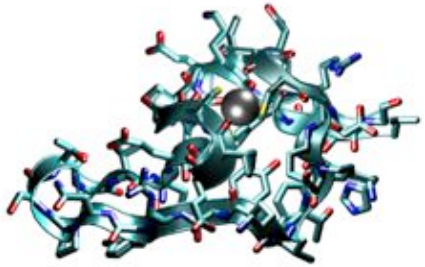


SH3 domain

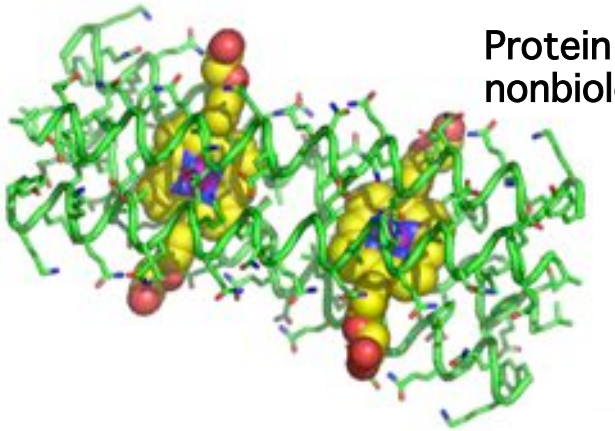


J. Mol. Biol., 2003. 334: 1101

Functional metalloproteins

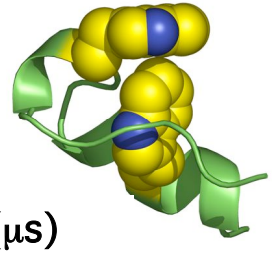


JACS, 2005. 127: 5804

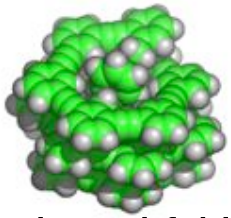


Protein complexes with nonbiological cofactors

JACS, 2005. 127: 1346

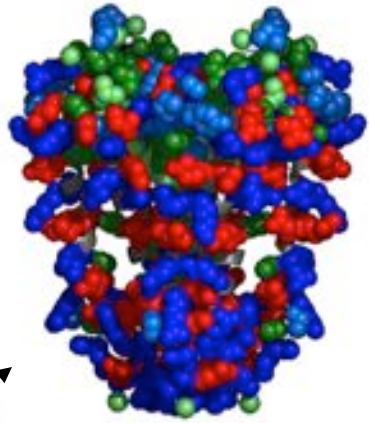


Ultrafast (μs) folding proteins



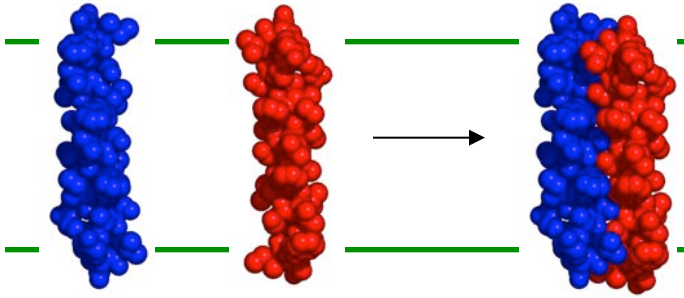
Nonbiological foldamers

J Phys Chem B, 2004. 108: 11988



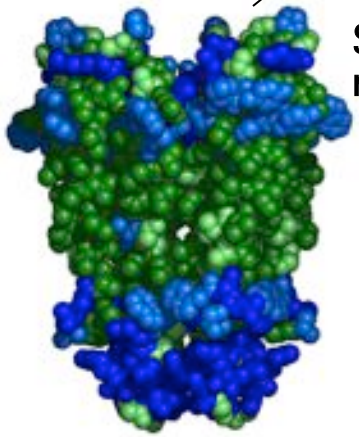
Solubilized integral membrane proteins

PNAS, 2004. 101: 1828

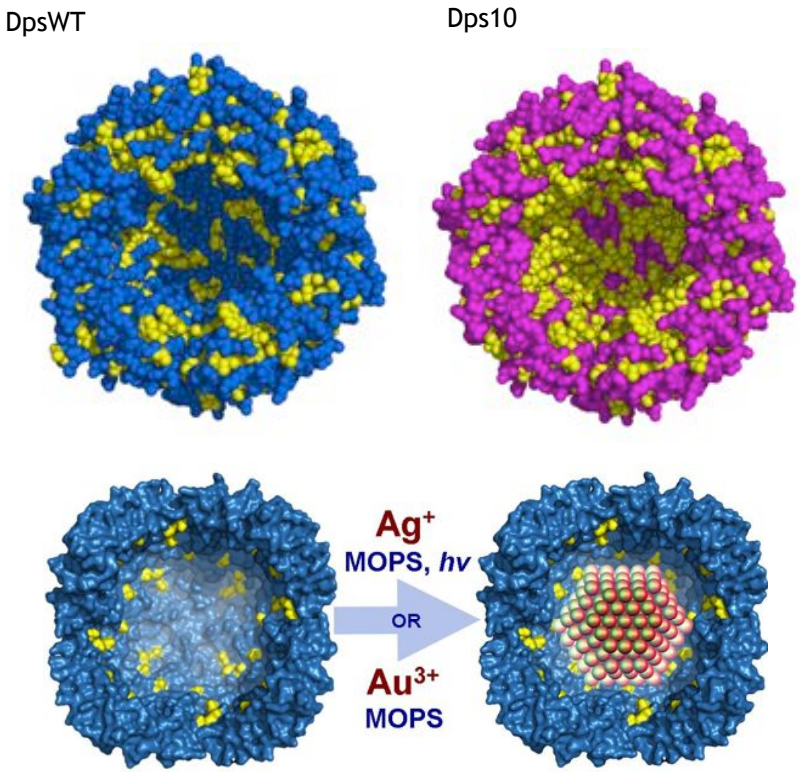


Membrane protein association

Membrane bilayer

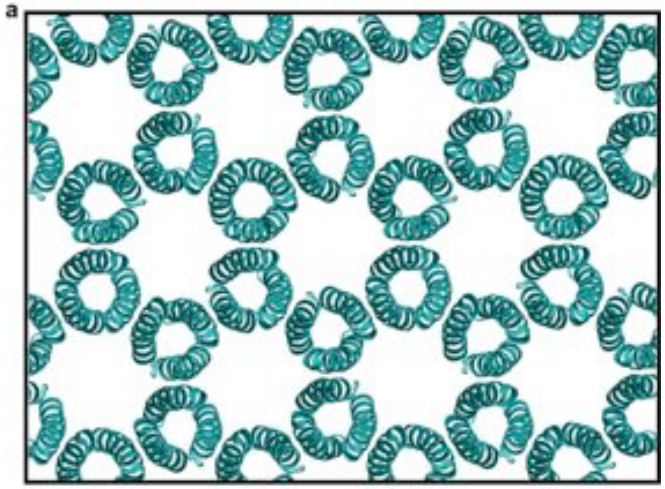


Computational design of protein assemblies



Designing proteins with icosohedral symmetry (ferritins)

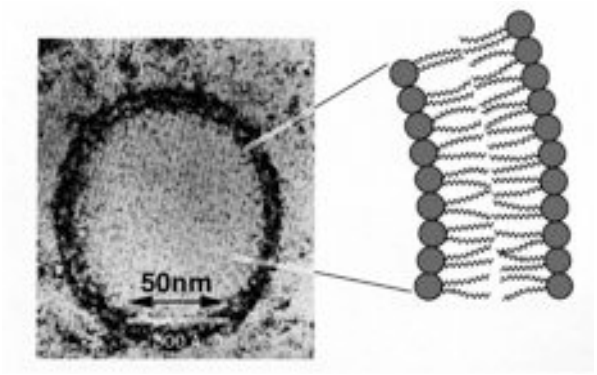
JACS, 2006.
Biochemistry, 2008.

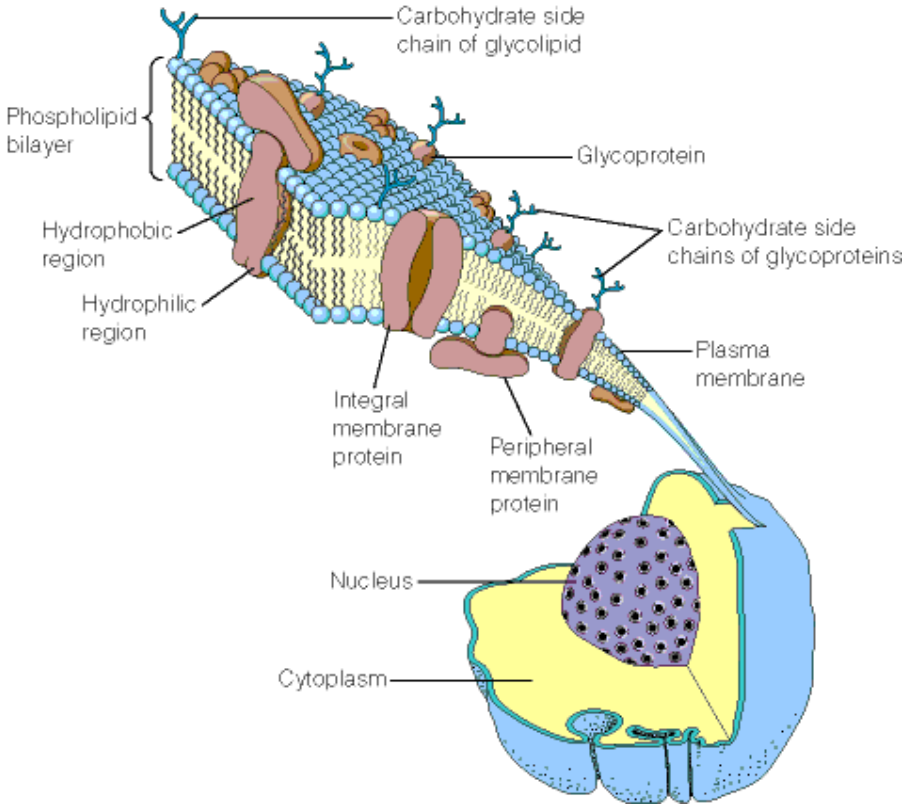


Designing protein crystals

Water-Solubilization of Membrane Proteins

Membrane proteins





Membrane proteins

- Few structures
- 30% of genome
- 30-50% of drug targets
- Poor expression, low solubility
- Often solubilized for biophysical studies and structure determination

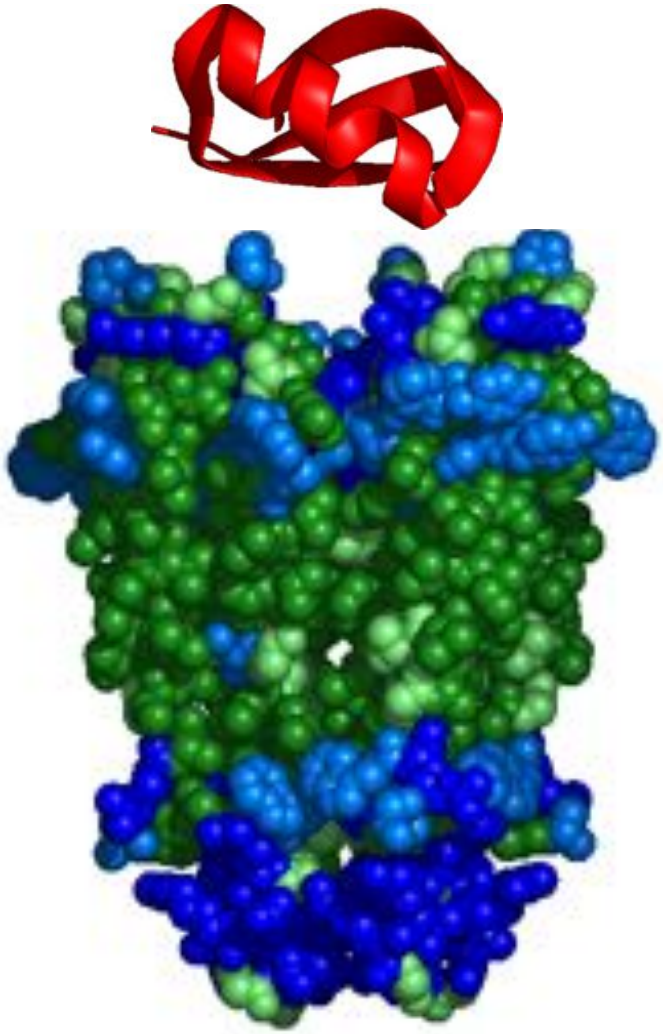
Water solubilization of membrane proteins: KcsA

Humanized version of a bacterial K-channel, tKcsA.

tKcsA: extracellular loops contain residues (from human K-channels) required for binding to a scorpion toxin, agitoxin 2.

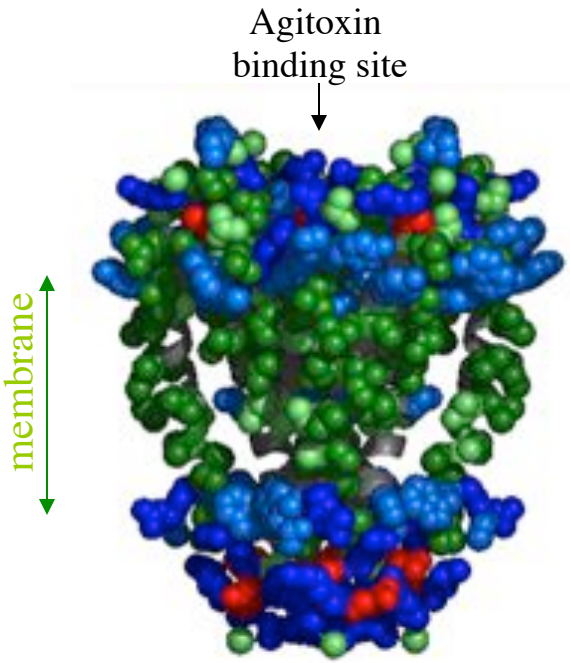
Selective toxin binding

R. MacKinnon, S. L. Cohen, A. Kuo, A. Lee, B. T. Chait, *Science* **280**, 106-109 (1998)



Designing WSK: soluble variant of the potassium channel KcsA

KcsA structure (Mackinnon & coworkers)
Tetramer of 103 residue helical protomers (412 total residues)
Computationally design 29 exterior amino acids of each subunit

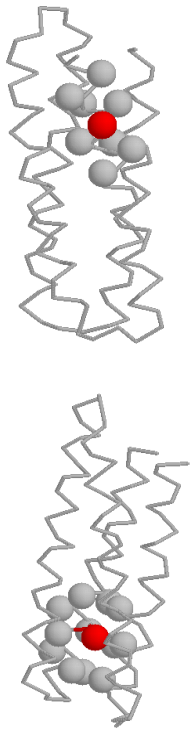


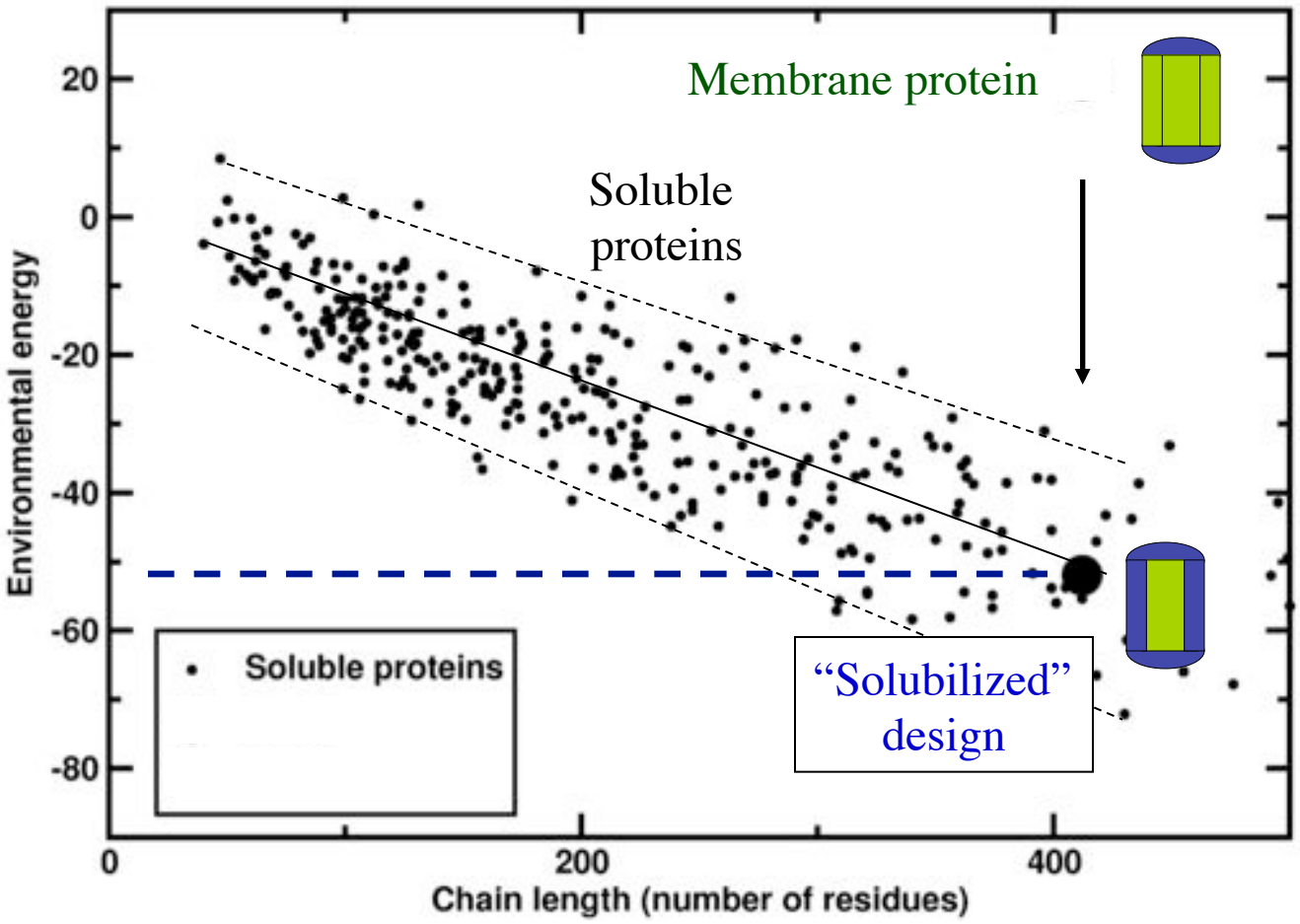
wild type

A. Slovic, H. Kono, J. Lear, J. G. Saven, and W. F. DeGrado, *Proc. Natl. Acad Sci., 101*: 1828-1833 (2004)

Solubilization of membrane protein

Solvation "energy"





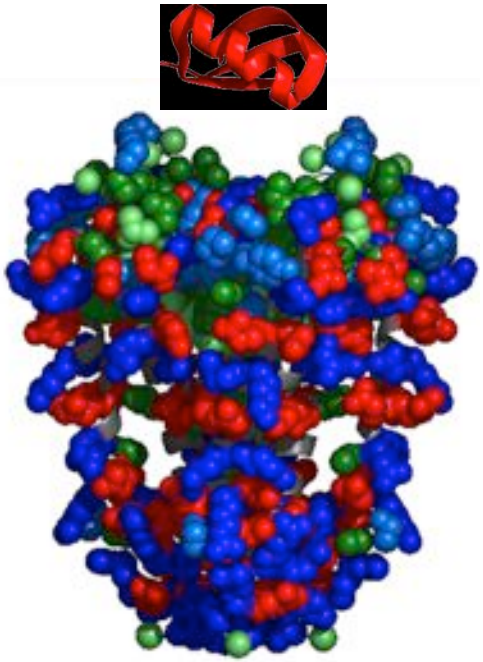
Calculations simultaneously consider solvation properties and inter-residue interactions of mutated residues

A. Slovic et al, *Proc. Natl. Acad Sci.*, 2004

Characterization of soluble variant

Express in soluble form

Properties shared with membrane soluble tKcsA



Structural criteria:

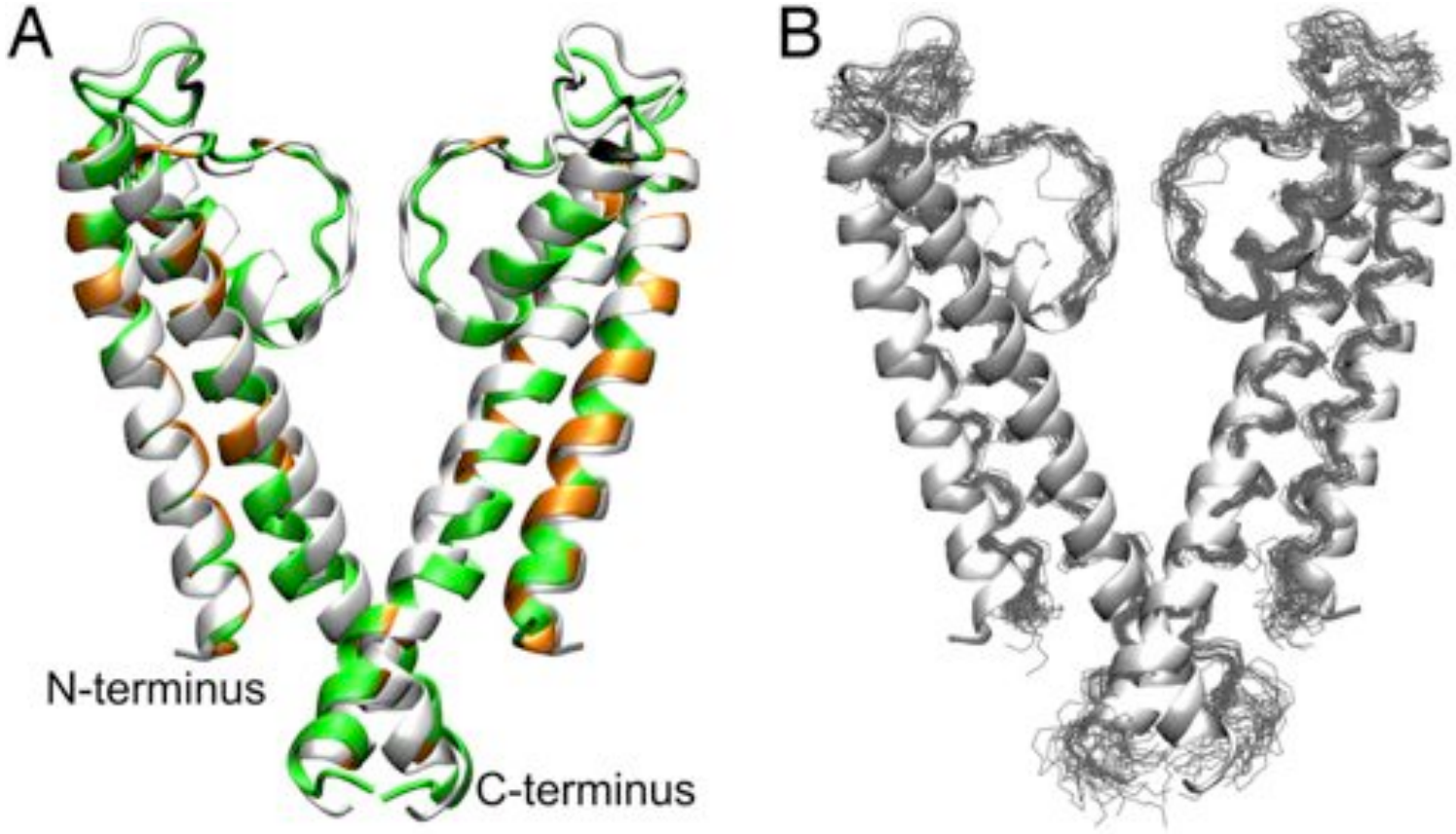
- Secondary structure
- Tetrameric

Functional criteria:

- Binds toxin specifically with same K_d
- Binds protein toxin stoichiometrically
- Binds small molecule channel blocker (TEA)

A. Slovic, H. Kono, J. Lear, J. G. Saven, and W. F. DeGrado, *Proc. Natl. Acad Sci., 101*: 1828-1833 (2004)

Solution structure of water soluble variant of K-channel (WSK-3)

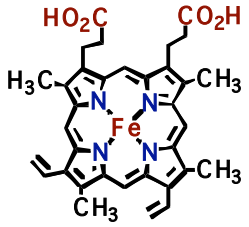


D. Ma, et al *PNAS* (2008), 105: 16537–16542

Designing protein complexes with nonbiological cofactors

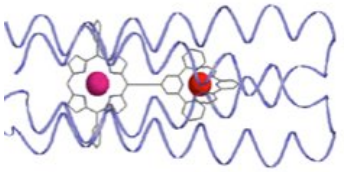
Groups of
Michael Therien, William DeGrado,
& J. Saven

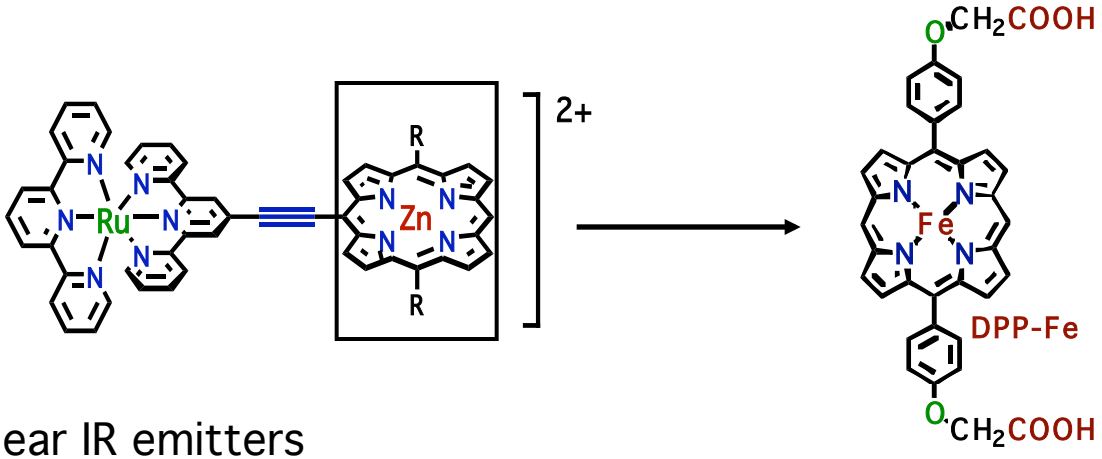
Protein complexes with nonbiological cofactors

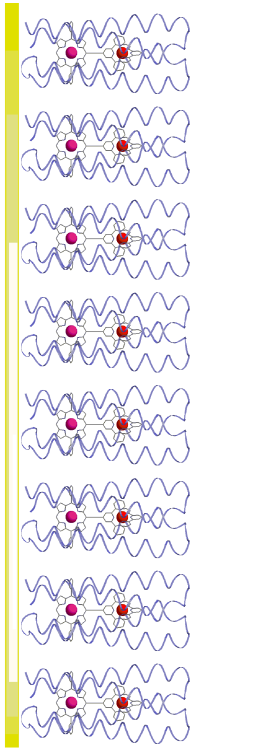


PPIX-Fe

- Cofactors confer function to proteins
  - e.g., Heme (oxygen binding, catalysis)
- New function and materials: proteins containing nonbiological cofactors
  - Controlled cofactor environment
  - Controlled protein assembly







- near IR emitters
- large molecular hyperpolarizability (NLO)
- long lived charge separated states
  (M. J. Therien)

Acknowledgments

J. G. Saven group (Penn)

Present
Chris Lanci
Chris MacDermaid
Jose Manuel Perez Aguilar

Former
Hidetoshi Kono
(Japan Atomic Energy Res. Inst.)
Andreas Lehmann
(Fox Chase Cancer Center)
Wei Wang (Colgate-Palmolive, Inc.)
Xi Yang (Standard & Poors)

William F. DeGrado group (Penn)

Michael J. Therien group (Penn & Duke)
Christopher Fry (Argonne Nat'l Lab)

Feng Gai group (Penn)

Ivan Dmochowski group (Penn)

Support
National Science Foundation
(MRSEC)
(NSEC)
Department of Energy
National Institutes of Health
University of Pennsylvania