7.91 Amy Keating

Comparing Protein Structures

Why?

detect evolutionary relationships
identify recurring motifs
detect structure/function relationships
predict function
assess predicted structures
classify structures - used for many purposes

Algorithms for detecting structure similarity

Dynamic Programming

- works on 1D strings - reduce problem to this
- can’t accommodate topological changes
- example: Secondary Structure Alignment Program (SSAP)

3D Comparison/Clustering

- identify secondary structure elements or fragments
- look for a similar arrangement of these between different structures
- allows for different topology, large insertions
- example: Vector Alignment Search Tool (VAST)

Distance Matrix

- identify contact patterns of groups that are close together
- compare these for different structures
- fast, insensitive to insertions
- example: Distance ALIgnment Tool (DALI)

Unit vector RMS

- map structure to sphere of vectors
- minimize the difference between spheres
- fast, insensitive to outliers
- example: Matching Molecular Models Obtained from Theory (MAMMOTH)

SSAP - Structure and Sequence Alignment Program

How about using dynamic programming? Any problems here?

Taylor & Orengo JMB (1989) 208, 1-22

SSAP - Structure and Sequence Alignment Program

How about using dynamic programming? Any problems here?

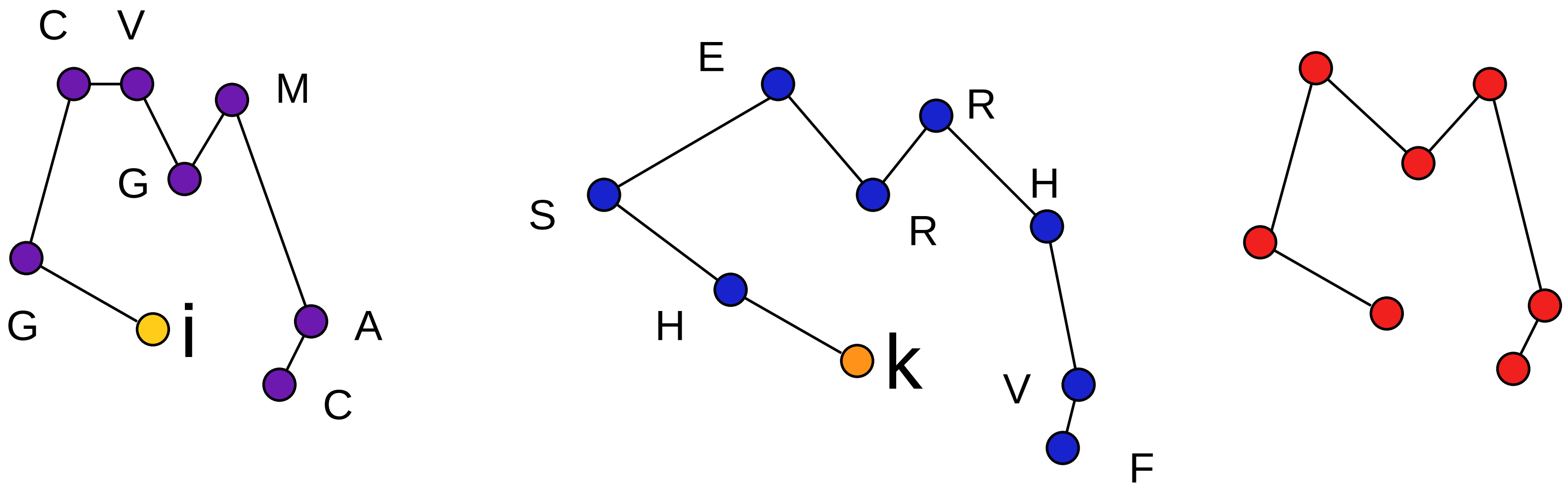
1. How will you evaluate if two positions are similar?
   Residue type
   expose to solvent
   secondary structure
   relationship to other atoms

2. Score that you give to an alignment of 2 residues depends on other residues
   ALIGNMENT depends on SUPERPOSITION but
   SUPERPOSITION depends on ALIGNMENT

Taylor, WR, and CA Orengo. "Protein Structure Alignment." *J Mol Biol.* 208, no. 1 (5 July 1989): 1-22.

SSAP - Structure and Sequence Alignment Program

For each pair of residues, (i,j), assume their equivalence. How similar are their environments wrt other residues?



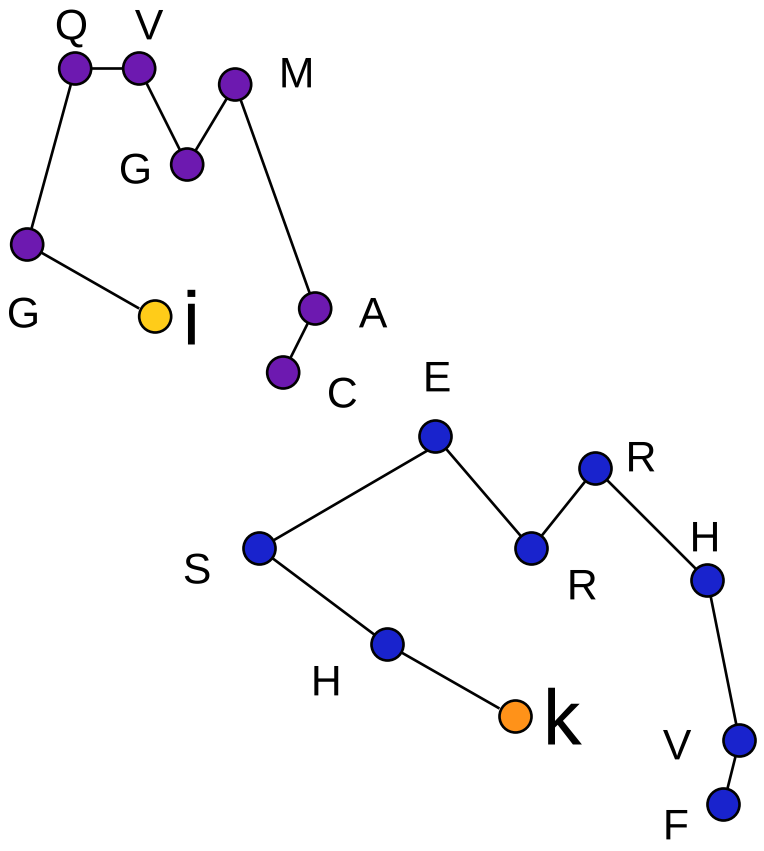
$s_{ik} = \Sigma a/(|d_{ij} - d_{kl}| + b)$; so s is large if d_{ij} and d_{kl} are similar.

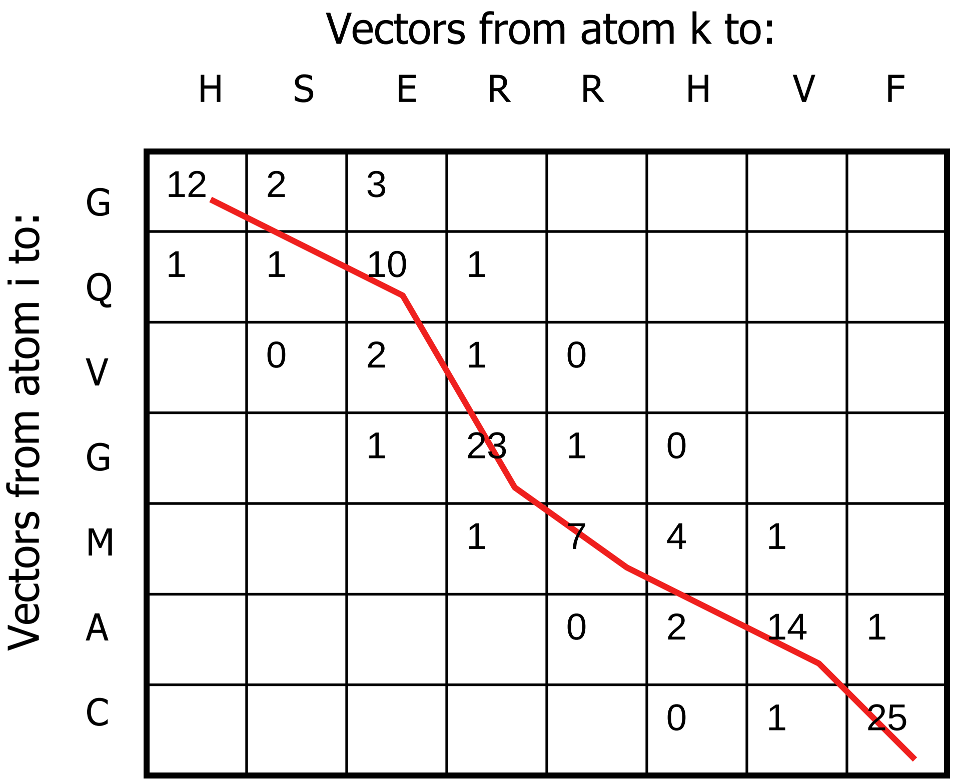
Which j and l should you compare with each other?

Images adapted from

Taylor, WR, and CA Orengo. "Protein Structure Alignment." *J Mol Biol.* 208, no. 1 (5 July 1989): 1-22.

Answer: use the j’s and l’s that give the **best** score





Vectors from atom k to:

Vectors from atom i to:

	H	S	E	R	R	H	V	F
G	12	2	3					
Q	1	1	10	1				
V		0	2	1	0			
G			1	23	1	0		
M				1	7	4	1	
A					0	2	14	1
C						0	1	25

NOTE: this gives an ALIGNMENT of how the residues of sequence A align with those of sequence B, when viewed from the perspective of i and k.

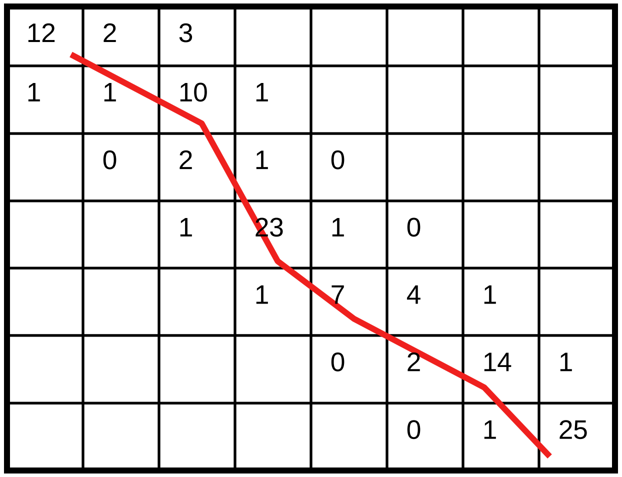
BUT, which i’s and k’s should you compare?

ALL OF THEM!

Then combine the results and take a consensus via another round of dynamic programming = “double dynamic programming”

Vectors from k = F

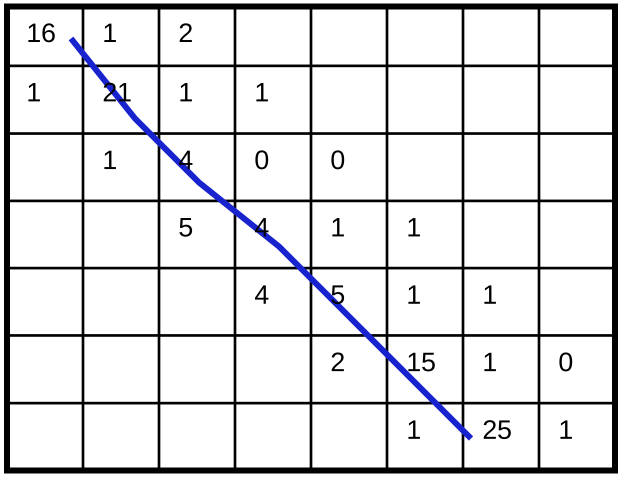
Vectors from i = C



12	2	3					
1	1	10	1				
	0	2	1	0			
		1	23	1	0		
			1	7	4	1	
				0	2	14	1
					0	1	25

Vectors from k = V

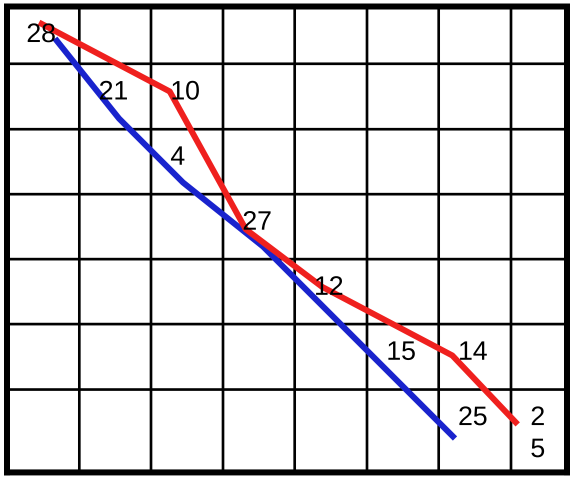
Vectors from i = C



16	1	2					
1	21	1	1				
	1	4	0	0			
		5	4	1	1		
			4	5	1	1	
				2	15	1	0
					1	25	1

Protein A

Protein B



28							
	21	10					
		4					
			27				
				12			
					15	14	
						25	2 5

Instead of using ***distances***, use ***vectors*** to include some directionality

$$s_{ij} = a/(|d_{ij} - d_{kl}| + b);$$

$$\Downarrow$$

$$s_{ij} = a/(|\mathbf{V}_{ij} - \mathbf{V}_{kl}| + b);$$

Can also include other information about residues i and k if desired
(e.g. sequence or environment information)

$$s_{ij} = (a + F(i,k)/(|\mathbf{V}_{ij} - \mathbf{V}_{kl}| + b);$$

It is important to assess whether detected similarities are SIGNIFICANT.

Various statistical criteria have been used.

General idea: How “surprising” is the discovery of a shared structure?

Structural Classification of Proteins

- Structure vs. structure comparisons (e.g. using DALI) reveal related groups of proteins
- Structurally-similar proteins with detectable sequence homology are assumed to be evolutionarily related
- Similarities between non-homologous proteins suggest convergent evolution to a favorable or useful fold
- A number of different groups have proposed classification schemes
  - SCOP (by hand)
  - CATH (uses SSAP)
  - FSSP (uses Dali)

Structural
Classification
Of
Proteins

statistics from August, 2003

7 CLASSES

(a,b,a/b,a+b…)

800 FOLDS

domain structures

1,294 SUPERFAMILIES

possible evolutionary relationship

2,327 FAMILIES

strong sequence homology

54,745 DOMAINS

Murzin, AG, SE Brenner, T Hubbard, and C Chothia. "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures." *J Mol Biol.* 247, no. 4 (7 April 1995): 536-40.

Structural
Classification
Of
Proteins

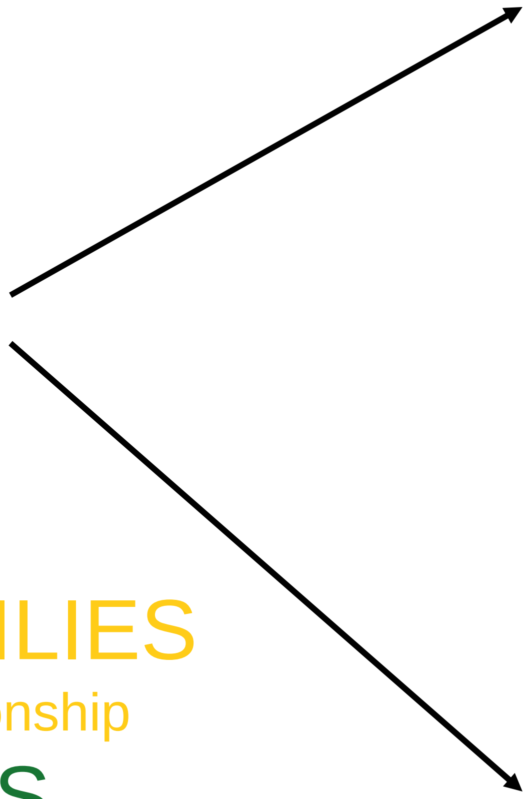
7 CLASSES
(a,b,a/b,a+b…)

800 FOLDS
domain structures

1,294 SUPERFAMILIES
possible evolutionary relationship

2,327 FAMILIES
strong sequence homology

54,745 DOMAINS



- all alpha
- all beta
- alpha/beta
- alpha + beta
- multi-domain
- membrane
- small
- *coiled-coil*
- *low-resolution*
- *peptide*
- *designed*

Structural
Classification
Of
Proteins

7 CLASSES

(a,b,a/b,a+b…)

800 FOLDS

domain structures

same secondary structure elements,
same order, same connectivity

1,294 SUPERFAMILIES

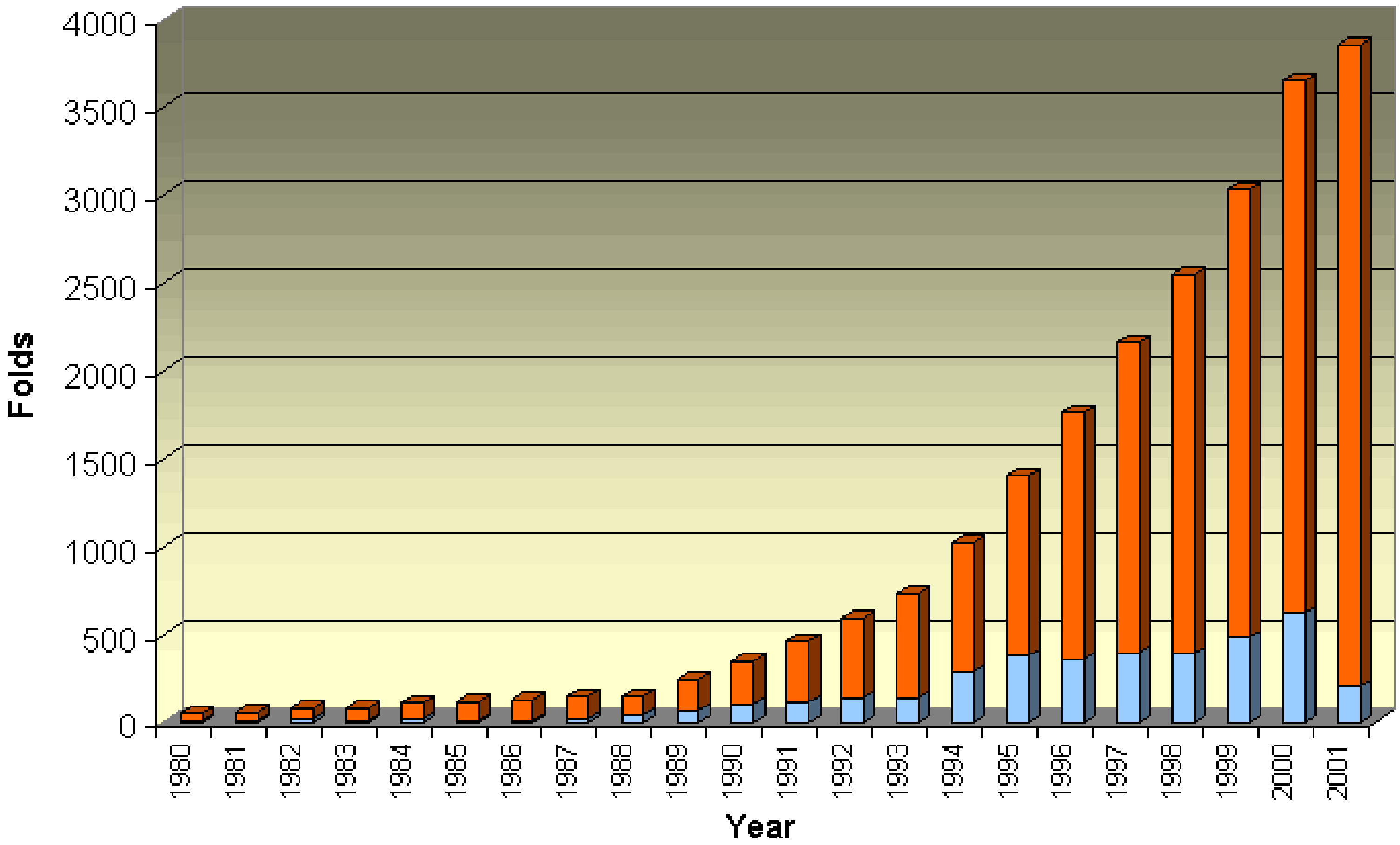
possible evolutionary relationship

2,327 FAMILIES

strong sequence homology

54,745 DOMAINS

PDB Growth in New Folds



structures submitted per year; new folds per year

(note that PDB criteria for a new fold differ from SCOP)

Structural
Classification
Of
Proteins

7 CLASSES

(a,b,a/b,a+b…)

800 FOLDS

domain structures

1,294 SUPERFAMILIES

possible evolutionary relationship

2,327 FAMILIES

strong sequence homology

54,745 DOMAINS

Low sequence identity, but probable evolutionary relationship (e.g. based on structure or function)

Structural
Classification
Of
Proteins

7 CLASSES
(a,b,a/b,a+b…)

800 FOLDS
domain structures

1,294 SUPERFAMILIES
possible evolutionary relationship

2,327 FAMILIES
strong sequence homology

54,745 DOMAINS

Clear evolutionary relationship;
often sequence identity > 30%

Structural

Classification

Of

Proteins

7 CLASSES

(a,b,a/b,a+b…)

800 FOLDS

domain structures

1,294 SUPERFAMILIES

possible evolutionary relationship

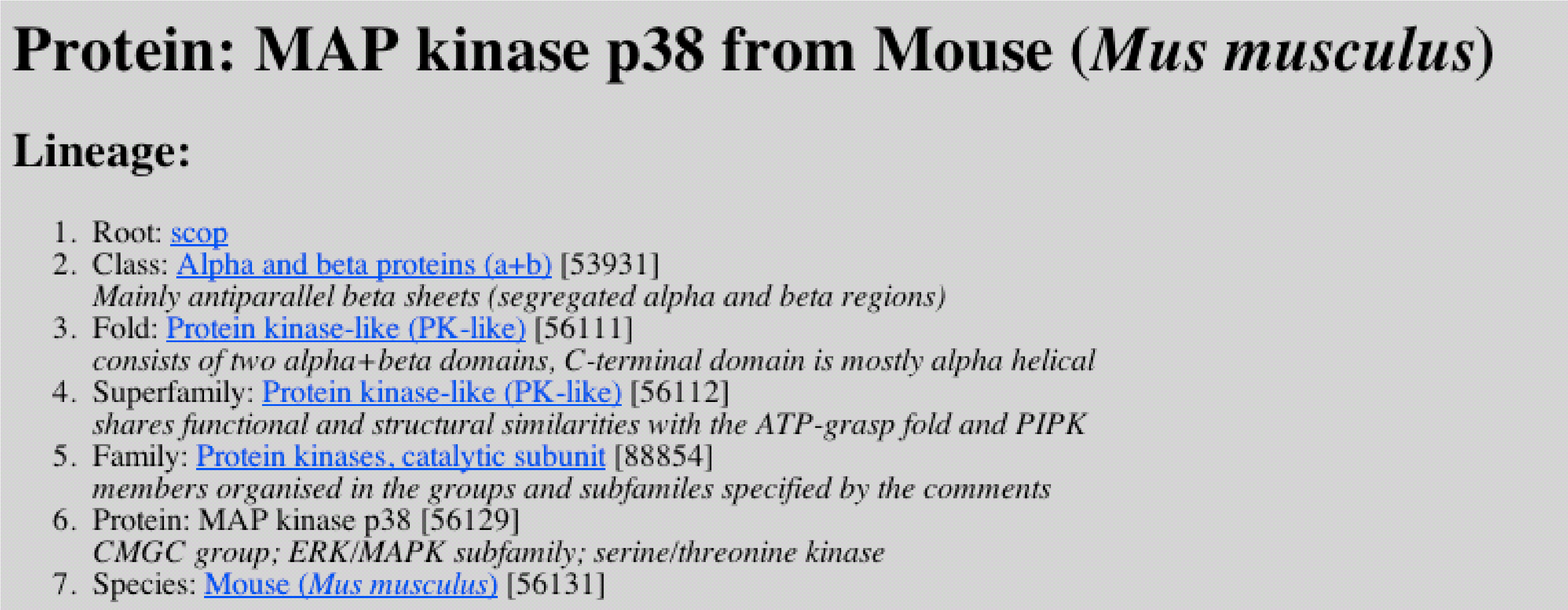
2,327 FAMILIES

strong sequence homology

54,745 DOMAINS

Autonomously-folding unit of compact structure

scop.mrc-lmb.cam.ac.uk/scop/index.html



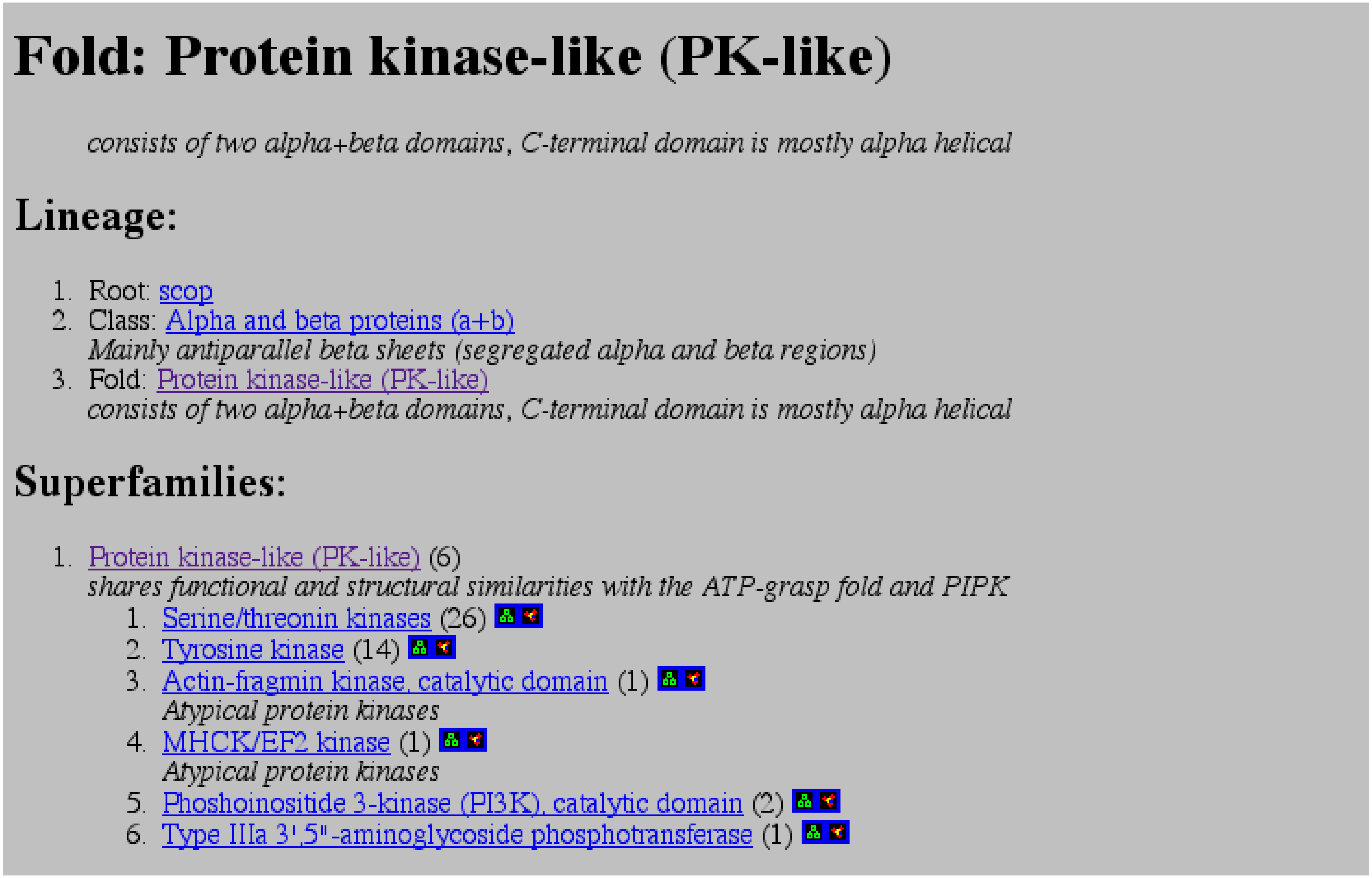
Protein: MAP kinase p38 from Mouse (*Mus musculus*)

Lineage:

1. Root: scop
2. Class: Alpha and beta proteins (a+b) [53931]
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
3. Fold: Protein kinase-like (PK-like) [56111]
   consists of two alpha+beta domains, C-terminal domain is mostly alpha helical
4. Superfamily: Protein kinase-like (PK-like) [56112]
   shares functional and structural similarities with the ATP-grasp fold and PIPK
5. Family: Protein kinases, catalytic subunit [88854]
   members organised in the groups and subfamiles specified by the comments
6. Protein: MAP kinase p38 [56129]
   CMGC group; ERK/MAPK subfamily; serine/threonine kinase
7. Species: Mouse (*Mus musculus*) [56131]

LCK kinase and p38 Map kinase in same family

Wasn’t true last year!



Fold: Protein kinase-like (PK-like)

consists of two alpha+beta domains, C-terminal domain is mostly alpha helical

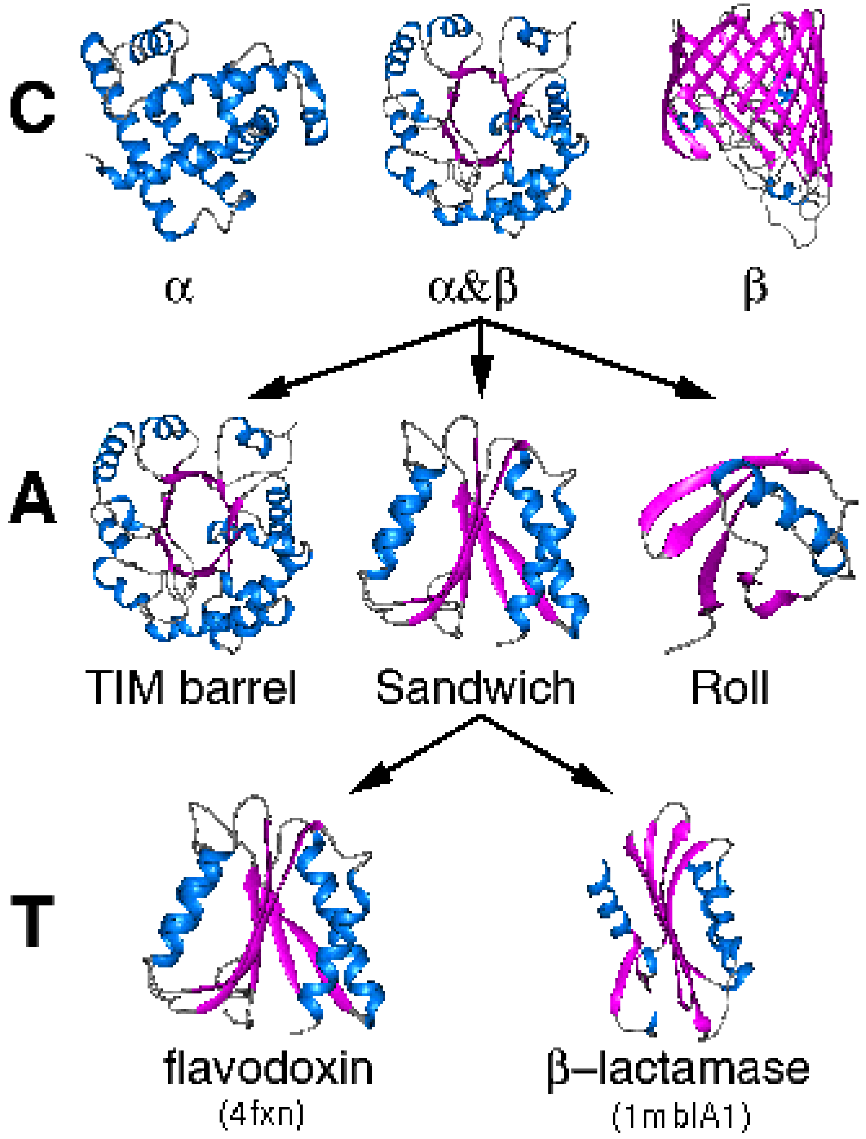
Lineage:

1. Root: scop
2. Class: Alpha and beta proteins (a+b)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
3. Fold: Protein kinase-like (PK-like)
   consists of two alpha+beta domains, C-terminal domain is mostly alpha helical

Superfamilies:

1. Protein kinase-like (PK-like) (6)
   shares functional and structural similarities with the ATP-grasp fold and PIPK
    1. Serine/threonin kinases (26)
    2. Tyrosine kinase (14)
    3. Actin-fragmin kinase, catalytic domain (1)
       Atypical protein kinases
    4. MHCK/EF2 kinase (1)
       Atypical protein kinases
    5. Phoshoinositide 3-kinase (PI3K), catalytic domain (2)
    6. Type IIIa 3',5"-aminoglycoside phosphotransferase (1)

CATH classification



CLASS

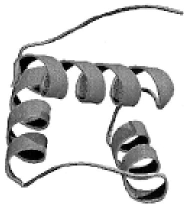
ARCHITECTURE

TOPOLOGY (fold)

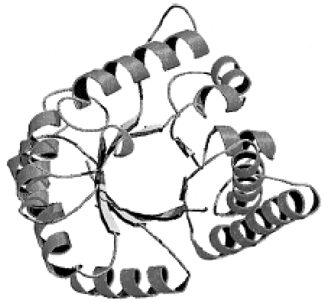
HOMOLOGY

Courtesy of Christine Orengo. Used with permission.

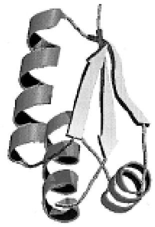
A few folds are highly-populated!



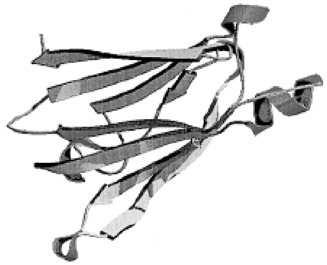
Arc repressor (1.10.10)



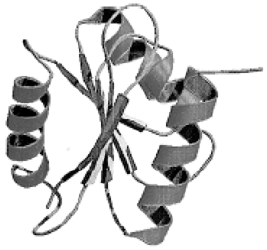
Tim barrel (3.20.20)



Alpha-beta plait (3.30.70)



Immunoglobulin-like (2.60.40)

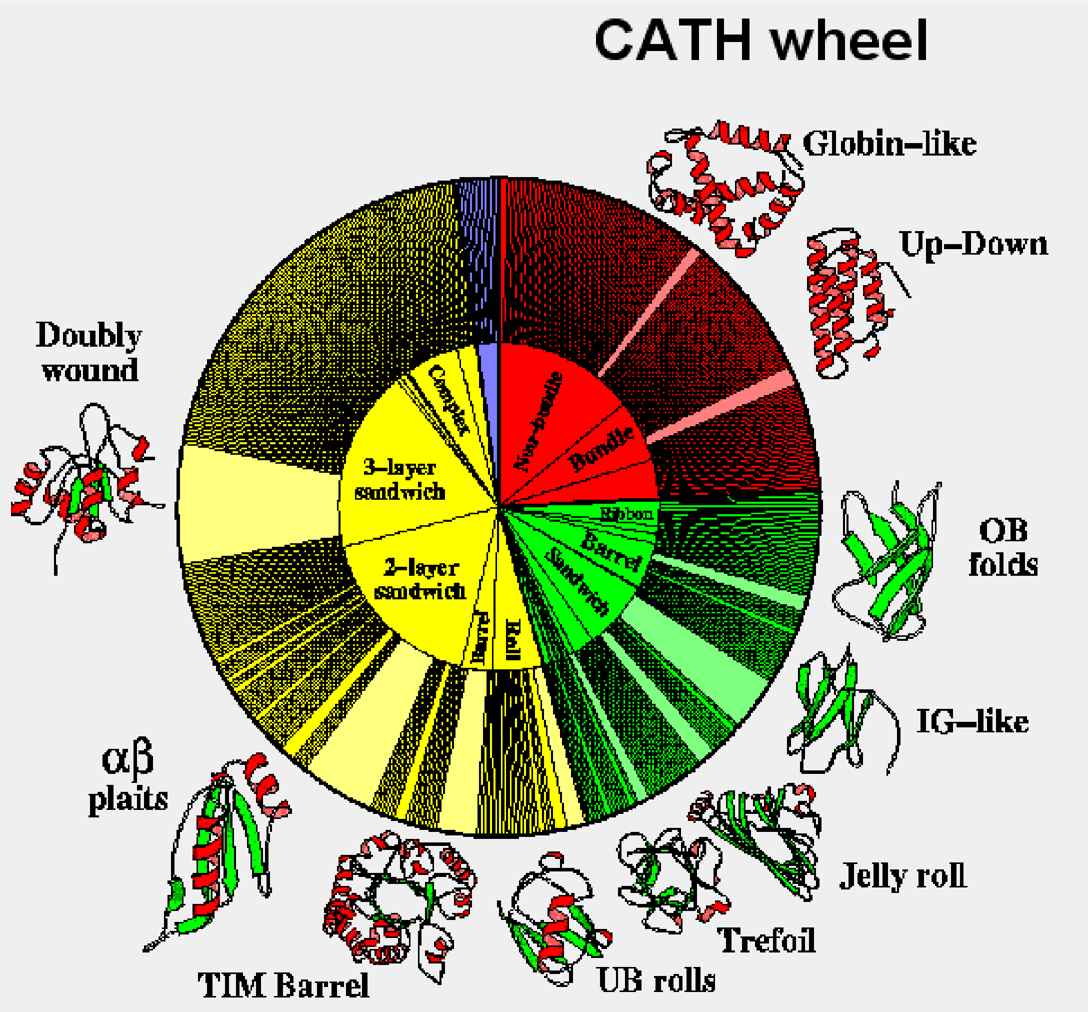


Rossmann (3.40.50)

Five folds in CATH contain 20% of all homologous superfamilies

Courtesy of Christine Orengo. Used with permission.

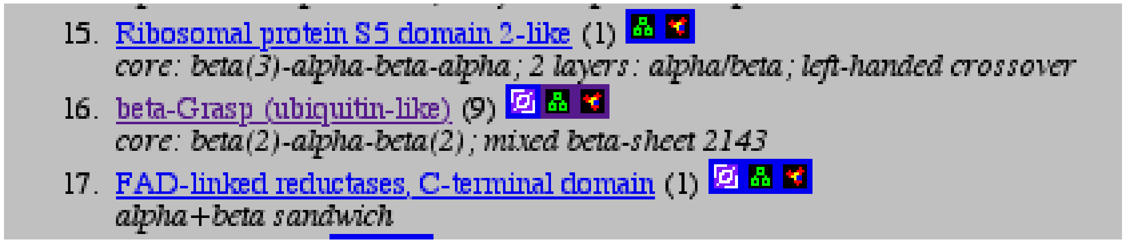
Some fold types are multi-functional



"superfolds" with > 3 functions

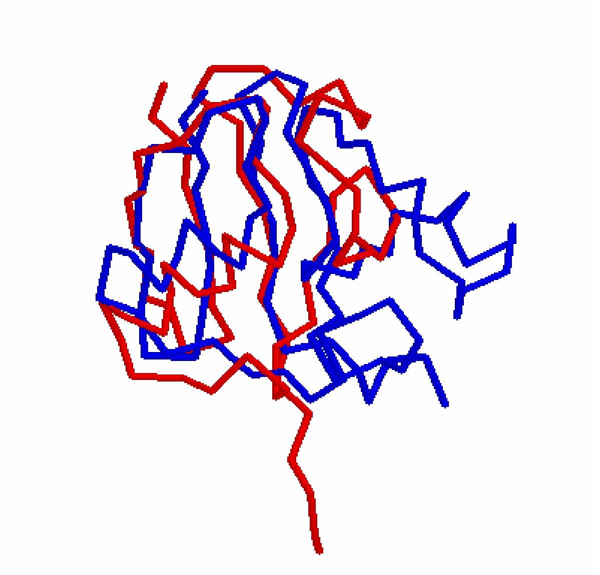
Courtesy of Christine Orengo. Used with permission.

SCOP entry:



15. Ribosomal protein S5 domain 2-like (1)
    core: beta(3)-alpha-beta-alpha; 2 layers: alpha/beta; left-handed crossover
16. beta-Grasp (ubiquitin-like) (9)
    core: beta(2)-alpha-beta(2); mixed beta-sheet 2143
17. FAD-linked reductases, C-terminal domain (1)
    alpha+beta sandwich

Use RASMOL to view the structures for ubiquitin and ferredoxin…



11% sequence identity

DALI superposition

Ubiquitin [MEDLINE: 91274342], PUB00000768, PUB00005320 is a protein of seventy six amino acid residues, found in all eukaryotic cells and whose sequence is extremely well conserved from protozoan to vertebrates. It plays a key role in a variety of cellular processes, such as **ATP-dependent selective degradation of cellular proteins, maintenance of chromatin structure, regulation of gene expression, stress response and ribosome biogenesis**. Ubiquitin is a globular protein, the last four C-terminal residues (Leu-Arg-Gly-Gly) extending from the compact structure to form a 'tail', important for its function. The
latter is mediated by the covalent conjugation of ubiquitin to target proteins, by an isopeptide linkage between the C-terminal glycine and the epsilon amino group of lysine residues in the target proteins.

The **ferredoxins** are **iron-sulfur proteins that transfer electrons in a widevariety of metabolic reactions**. They have a cofactor which binds a 2FE-2S cluster. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron sulfur cluster(s) and according to sequence similarities IPR000564.

Pfam annotations

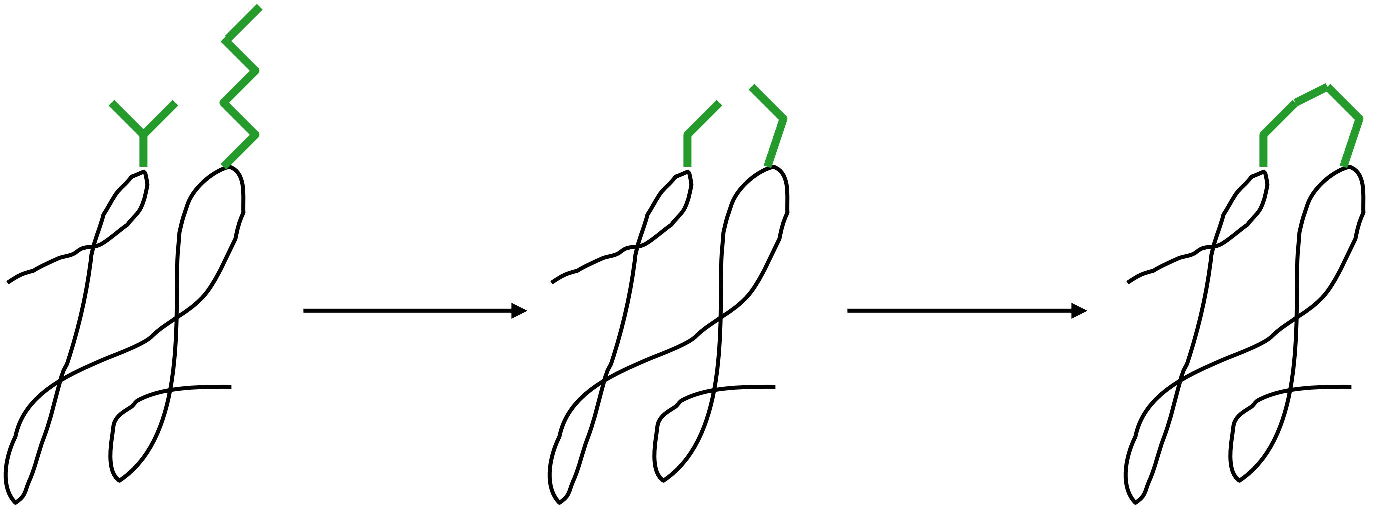
Molecular Modeling: Methods & Applications

Acknowledgement: The following materials were prepared by or with the help of Professor Bruce Tidor.

How do we use computational methods
to **analyze**, **predict**, or **design**
protein sequences and structures?

Theme:
Methods based on ***physics*** *vs. methods*
based on our ***accumulated empirical***
knowledge *of protein properties*

Example: Design of Disulfide-Stabilized Proteins



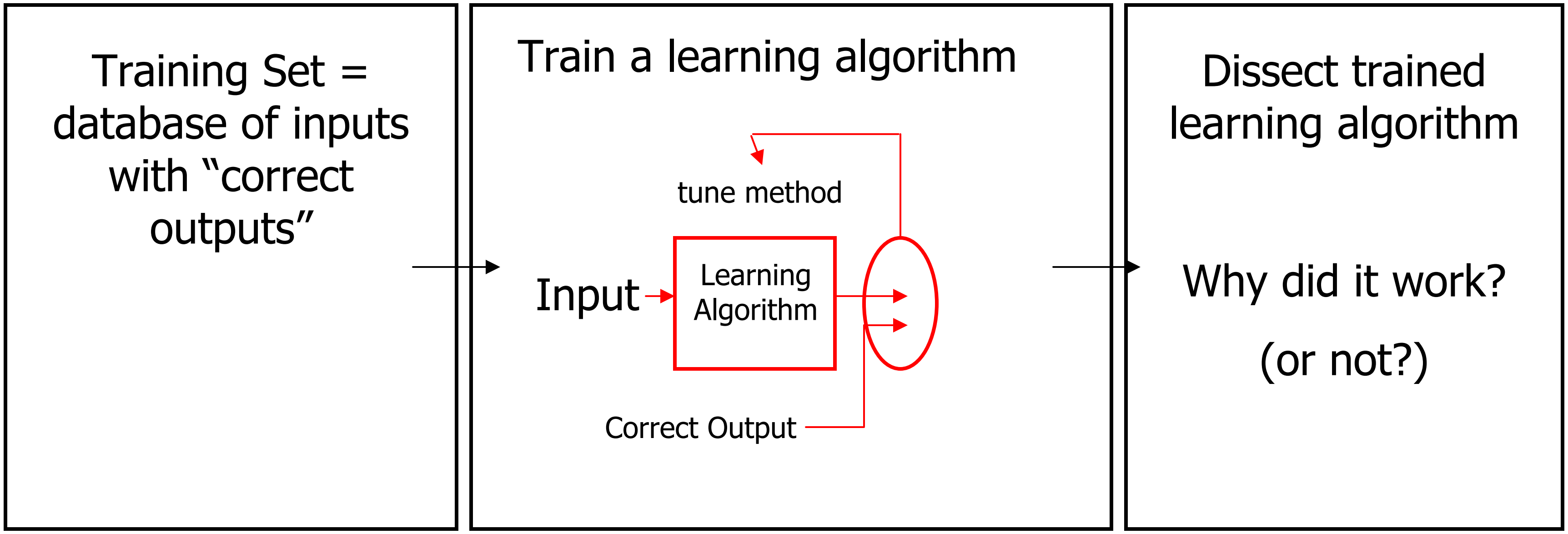
2 wild-type residues

2 Cys mutations

1 disulfide bond

Approach 1: learn from sequence

If you only have a protein sequence, can you identify isolated Cys residues versus those that are involved in disulfide bonds?



Training Set = database of inputs with "correct outputs"

Train a learning algorithm

tune method

Input

Learning Algorithm

Correct Output

Dissect trained learning algorithm

Why did it work?

(or not?)

Muskal, SM, SR Holbrook, and SH Kim. "Prediction of The Disulfide-bonding State of Cysteine in Proteins." *Protein Eng.* 3, no. 8 (August 1990): 667-72.

Results for Approach 1

- Input: Protein sequence flanking Cys residues (±5)
- Learning algorithm: Neural network
- Predictive success: ~80%
- Implies that Cys-bond formation is largely influenced by local sequence
- Analysis of trained network weights
  - Hydrophilic local sequence increases propensity for disulfide bonded structure
  - Hydrophobic local sequence increases propensity for isolated sulfhydryl
  - Shows interesting difference between Phe and Trp vs. Tyr
- Drawback: don’t learn which Cys residues are paired!

Approach 2: Database Driven

- Start with database of known disulfide bond geometries from the PDB
- For target protein structure, search over all pairs of residues
  - Try all disulfide bond geometries from database for compatibility with this pair of positions
  - Record any compatible disulfides
- Report successful pairs of residues
- Result: successful introduction of S-S bond into l repressor -> more stable protein, still binds DNA

Pabo, CO, and EG Suchanek. "Computer-aided Model-building Strategies for Protein Design." *Biochemistry* 25, no. 20 (7 October 1986): 5987-91.

Approach 3: Energy-Function Based

- For our protein structure, search over all pairs of residues
  - Build a model of the Cβ and Sγ atoms and determine if these are compatible with a disulfide bond in this geometry
  - If so, build lowest energy disulfide between this pair of residues
  - Evaluate energy of this disulfide with some energy function
- Report successful pairs of residues
- Succeeds in predicting the geometry of many known disulfide bonds

Hazes, B, and BW Dijkstra. "Model Building of Disulfide Bonds in Proteins with known Three-dimensional Structure." *Protein Eng.* 2, no. 2 (July 1988): 119-25.

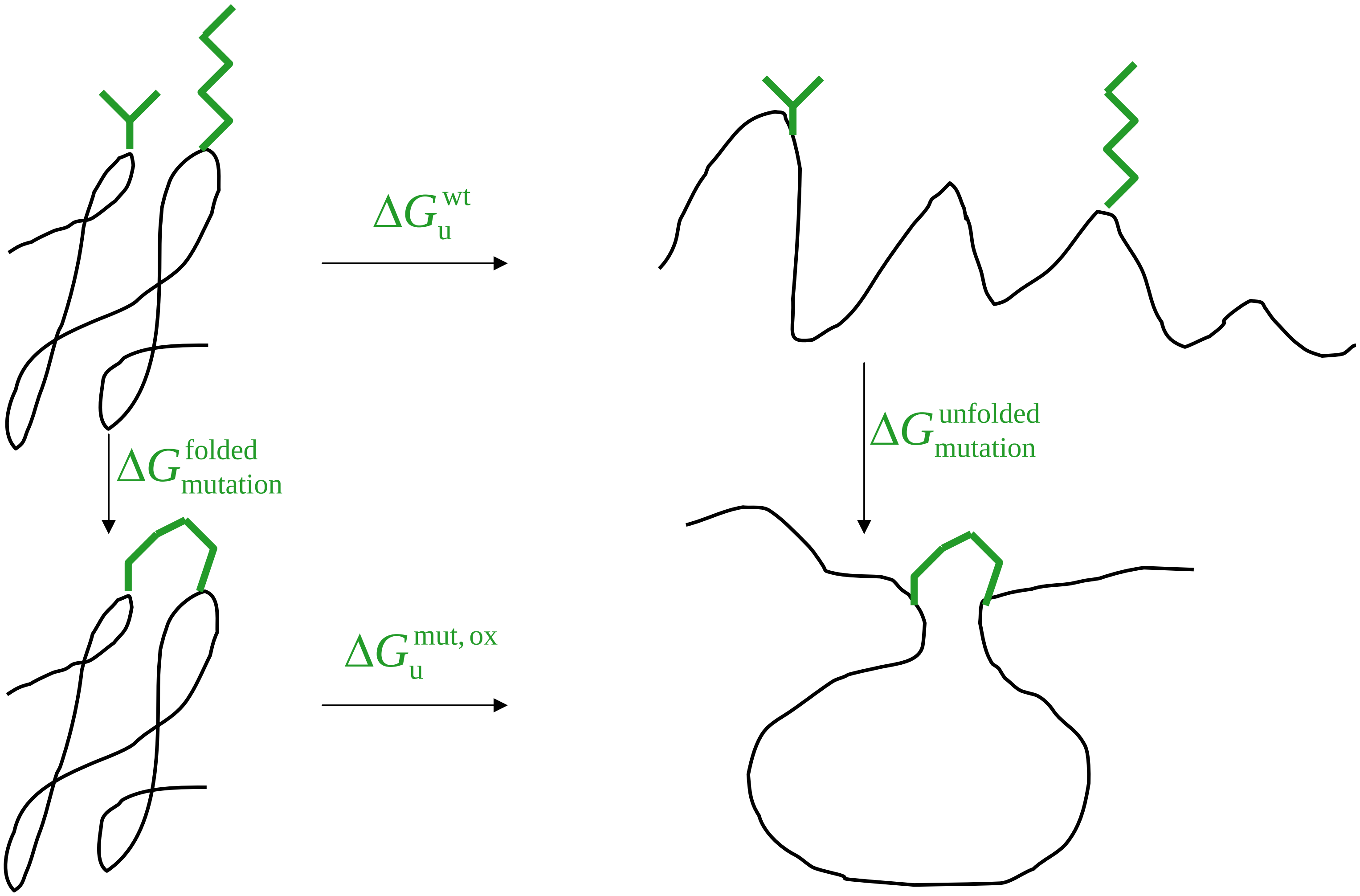
Pros and Cons of the Different Approaches

- Machine-learning methods often don’t provide a clear understanding of why they worked
- There are obvious structural constraints on disulfide bonds, and sequence-based methods may not be able to capture these
- Structure data isn’t always available, so sequence-based methods can be valuable
- Databases of known disulfides may be incomplete
- Disulfides might not be transferable to a different context
- When using a database, you don’t need to have an accurate description of the physics
- Methods based on first principles can identify things never seen before
- Our ability to model proteins from first principles is limited
  Does the model include structural relaxation?

And one more caveat…

How do disulfide bonds stabilize proteins?

What if you want to compute how much the disulfide bond stabilizes the protein?


$\Delta G_{\mathrm{u}}^{\mathrm{wt}}$
$\Delta G_{\mathrm{mutation}}^{\mathrm{folded}}$
$\Delta G_{\mathrm{mutation}}^{\mathrm{unfolded}}$
$\Delta G_{\mathrm{u}}^{\mathrm{mut,\,ox}}$

$$\Delta\Delta G_{\mathrm{u}} = \Delta G_{\mathrm{u}}^{\mathrm{mut,\,ox}} - \Delta G_{\mathrm{u}}^{\mathrm{wt}} = \Delta G_{\mathrm{mutation}}^{\mathrm{unfolded}} - \Delta G_{\mathrm{mutation}}^{\mathrm{folded}}$$

Energy-based modeling of protein structure and function

- CONFORMATIONAL ANALYSIS - what are the low-energy structures a protein can adopt?
- DYNAMICS - how do proteins move?
- THERMODYNAMICS - can compute quantities that characterize the system (e.g. enthalpy, entropy, heat capacity, free energy differences)
- ENERGY COMPONENTS - which atoms or which forces contribute the most to protein stability?
- REACTIVITY - what are the mechanisms and rates of reactions? Typically requires quantum mechanics.

For a molecular simulation or model you need:

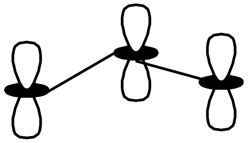
1. A representation of the protein

2. An energy function

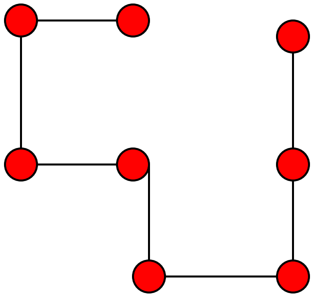
3. A search algorithm or optimizer

Levels of Representation

Electrons:



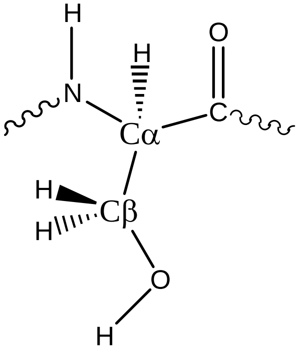
Residues:

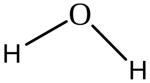


on or off a lattice

Atoms:

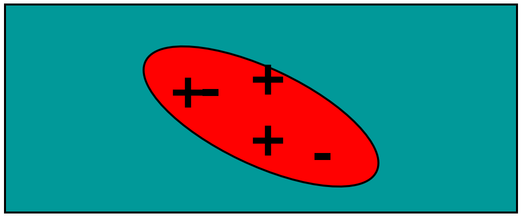
All-atom





protein, DNA, solvent, ligands, ions

Continuum:



Solvent as a high dielectric, protein as a low-dielectric “glob” with charges inside

Quantum mechanics describes the energy of a molecule in terms of a *wavefunction* describing the location and motion of nuclei and electrons in the molecule

$$H\Psi(\mathbf{r},\mathbf{R}) = E\bullet\Psi(\mathbf{r},\mathbf{R})$$

$\Psi(\mathbf{r},\mathbf{R}) = \Psi(\mathbf{r})\bullet\Psi(\mathbf{R})$ *Born-Oppenheimer*

This can only be solved exactly for a small number of systems - even the helium atom is too complex for an exact solution!

It is *much* too expensive to compute the energies of proteins and DNA using quantum methods. Instead, we use empirical approximations that capture the important effects. For the most part, this is ok for the description of biological macromolecules at room temperatures.

NOTE: once we ignore the electronic part of the wavefunction we can no longer compute the energy of bonds breaking and forming.

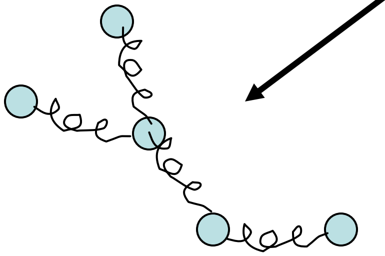
Potential Energy Using Molecular Mechanics

Goal: Describe potential energy of any conformation of molecule

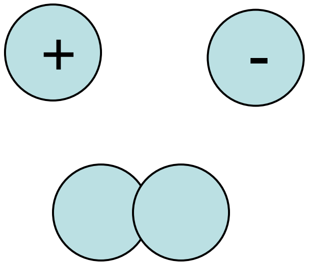
Use molecular mechanics: based on physics, but uses simplified “ball and spring” model. Think **Newton**, not **Schroedinger**!

Model is EMPIRICALLY adjusted to capture quantum effects that give rise to bonding.

$$\boxed{U(\vec{R}^{3N}) = U_{\text{Covalent}} + U_{\text{Non-covalent}}}$$



bonds become “springs”



Covalent Potential Energy Terms

$$U_{\text{Covalent}} = U_{\text{bond}} + U_{\text{bond angle}} + U_{\text{improper dihedral}} + U_{\text{torsion}}$$

$$U_{\text{bond}} = \sum_{\text{bonds}} \frac{1}{2} k_b (b - b_0)^2$$

$$U_{\text{bond angle}} = \sum_{\text{bond angles}} \frac{1}{2} k_\theta (\theta - \theta_0)^2$$

$$U_{\text{improper dihedral}} = \sum_{\text{improper dihedrals}} \frac{1}{2} k_\Phi (\Phi - \Phi_0)^2$$

$$U_{\text{torsion}} = \sum_{\text{torsions}} \frac{1}{2} k_\phi [1 + \cos(n\phi - \delta)]$$

Brooks et al., *J. Comput. Chem*. 4: 187-217 (1983)

Key to Symbols: covalent terms

k_b, k_θ, and k_Φ are harmonic force constants for bond, bond angle, and improper dihedral terms, respectively.

b_0, θ_0, and Φ_0 are equilibrium bond lengths, bond angles, and improper dihedrals, respectively.

b, θ, and Φ are actual values for bond lengths, bond angles, and improper dihedrals, respectively, in this particular structure.

k_ϕ is the barrier height for an individual torsion, n is its "periodicity" (2-fold, 3-fold, etc.), δ is the position of the maximum, and ϕ is the value of this torsion in this particular structure.

$$U_{\text{bond}} = \sum_{\text{bonds}} \frac{1}{2} k_b (b - b_0)^2$$

$$U_{\text{bond angle}} = \sum_{\text{bond angles}} \frac{1}{2} k_\theta (\theta - \theta_0)^2$$

$$U_{\text{improper dihedral}} = \sum_{\text{improper dihedrals}} \frac{1}{2} k_\Phi (\Phi - \Phi_0)^2$$

$$U_{\text{torsion}} = \sum_{\text{torsions}} \frac{1}{2} k_\phi [1 + \cos(n\phi - \delta)]$$

Non-Covalent Potential Energy Terms

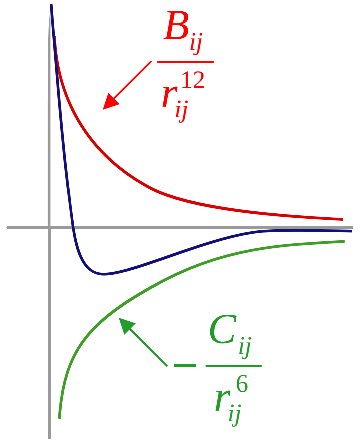
$$U_{\text{Non-covalent}} = U_{\text{vdW}} + U_{\text{elec}}$$

$$U_{\text{vdW}} = \sum_{i \succ j} \left(\frac{B_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}} \right)$$

“accurate”

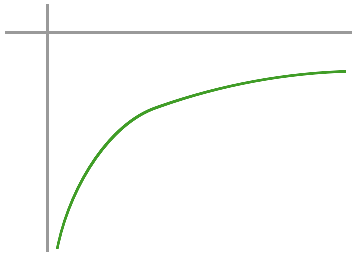
approximate

Lennard-Jones potential



$$U_{\text{elec}} = \sum_{i \succ j} \frac{q_i q_j}{\varepsilon r_{ij}}$$

Coulomb’s law



Key to Symbols: non-covalent

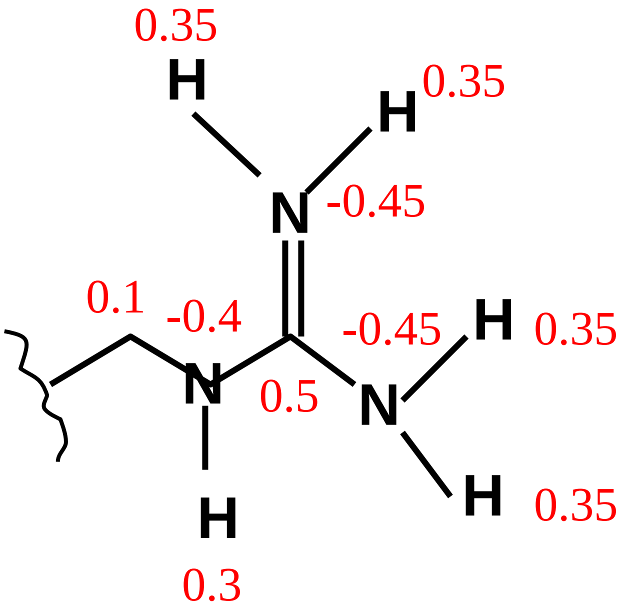
$$U_{\mathrm{vdW}} = \sum_{i \succ j} \left(\frac{B_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}} \right)$$

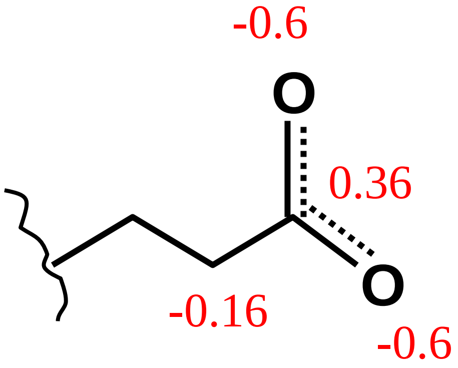
r_{ij} is the distance between atom *i* and atom *j*, B_{ij} and C_{ij} are parameters describing the vdW function

$$U_{\mathrm{elec}} = \sum_{i \succ j} \frac{q_i q_j}{\varepsilon r_{ij}}$$

q_i & q_j are the partial atomic charges on atoms *i* & *j*, and ε is the effective dielectric constant.

Partial atomic charges are used in Coulomb's Law





These charges come from higher-level quantum calculations.

Parameterization of the Potential

$$k_b,\ b_0,\ k_\theta,\ \theta_0,\ k_\Phi,\ \Phi_0,\ k_\phi,\ n,\ \delta,\ q_i,\ B_{ij},\ C_{ij}$$

- Must develop set of **transferable** parameters
- Parameters obtained from fits to both experimental and theoretical data
  - Much of data is from small molecules
  - Crystal structures (lengths & angles, non-bonded coeffs.)
  - Vibrational spectroscopy & ab initio QM calculations (*q*'s, *K*'s)
  - Calorimetric & thermodynamic measurements (*q*'s, *K*'s)
- Test parameters in context of entire protein

- *Overriding assumption: Parameters for fragments of proteins are appropriate for that fragment in different contexts.*

"Missing Terms" in the Potential Function

- No hydrogen-bond term
  - Treated as part of electrostatics

- No hydrophobic term
  - Is resultant from all other forces

Adding either of these would result in an imbalance in the potential due to double-counting.

What about the solvent?

The preceding energy function will give you the energy in the GAS PHASE. Not so useful for studying biology…

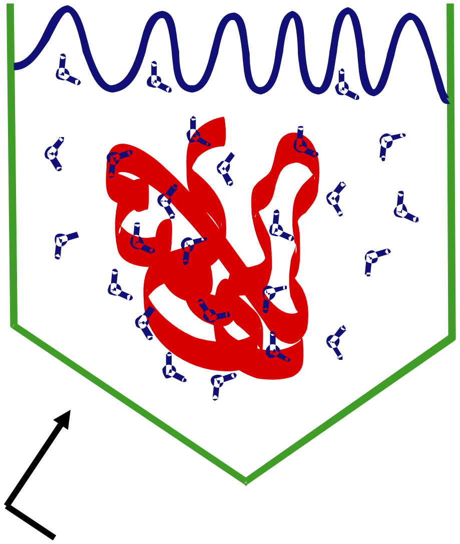
Aqueous solvent is troublesome for two reasons:

1. There are LOTS of solvent molecules
2. The water has a strong influence on the electrostatic interactions

Calculations that provide an accurate description of proteins or DNA in solvent are computationally demanding.

Alternative Electrostatic Treatments

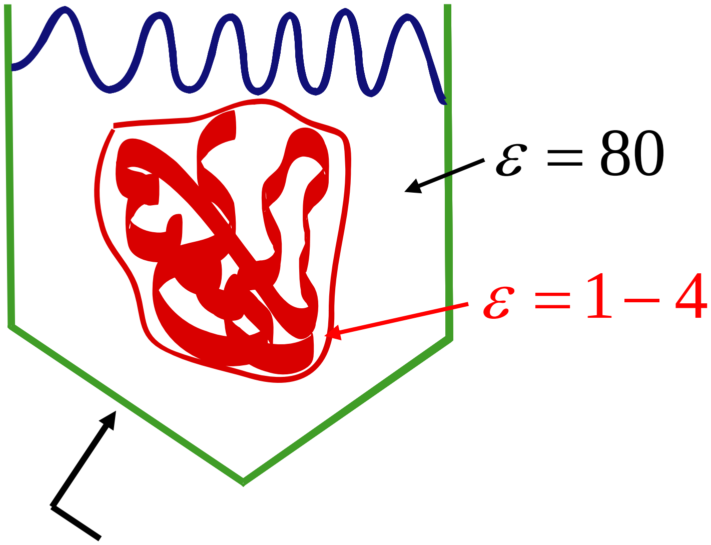
Microscopic treatment



Coulomb's Law OK

- Must include all solvent atoms in sum

Macroscopic treatment



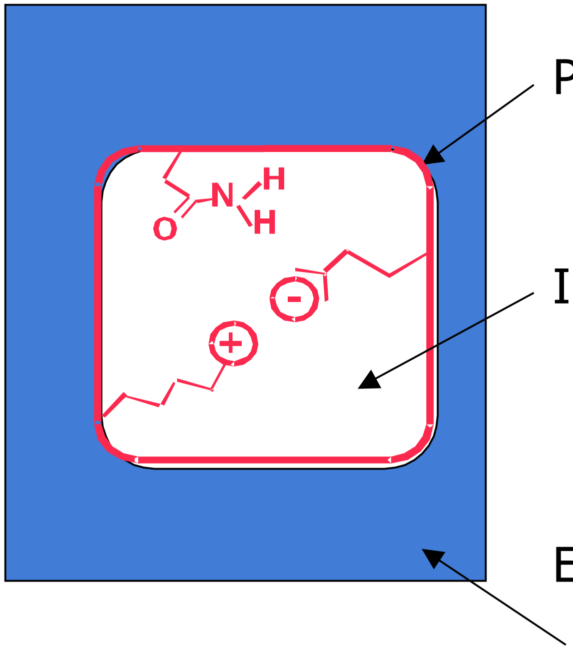
Coulomb's Law Not OK

- But can use Poisson–Boltzmann equation

Simpler Representation ⇒ More Complex Physics

Slide courtesy of B. Tidor.

Continuum Electrostatics



Protein Boundary

- Defined by contact surface with water probe

Interior of Protein

- Atoms represented as fixed point charges
- Low dielectric constant (usually 1, 2, 3, or 4)

Exterior of Protein

- No explicit solvent atoms
- Solvent water represented by high dielectric constant (80)
- Ionic strength treated with Debye-Hückel-type model

<u>Numerically solve the Poisson-Boltzmann equation on a grid</u>

$$\nabla\varepsilon(\vec{r})\nabla\phi(\vec{r}) - \overline{\kappa}^{2}(\vec{r})\sinh[\phi(\vec{r})] = -4\pi\rho(\vec{r})$$

E(r) = dielectric, *f*(r) = electrostatic potential, *r*(r) = charge density,
k is related to the ionic strength

Slide courtesy of B. Tidor.

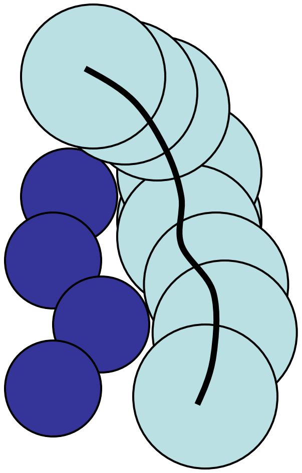
Empirical solvation models (crude!)

1. Solvent-accessible surface area model

polar atoms are rewarded for exposure to solvent

hydrophobic atoms are penalized

"roll" solvent over surface to get area



$$E_{\text{solvation}} = \sum_{atoms\_i} s \bullet SA_i$$

This model doesn't account for the fact that water screens (weakens) electrostatic interactions. Often used in combination with:

2. Distance-dependent dielectric model

$$U_{\text{elec}} = \sum_{i \succ j} \frac{q_i q_j}{\varepsilon(r) r_{ij}}$$

Properties of Potential

$$U(\vec{R}^{3N}) = \underbrace{U_{\text{bond}} + U_{\text{bond angle}} + U_{\text{improper dihedral}} + U_{\text{torsion}}}_{\text{scales as N (number of atoms)}} + \underbrace{U_{\text{vdW}} + U_{\text{elec}}}_{\text{scales as } N^2}$$

- Often implement some type of cutoff function to smoothly turn off non-covalent interactions beyond some distance

IMPORTANT: parameterized only to give ***differences*** in energy for conformations - does not give energy of folding or free energy of formation! Must formulate your problem (with an appropriate reference state) so that you are considering energy differences.