# The State of Affairs in Protein Modeling

Andrew Gillette

Department of Mathematics,
Institute of Computational Engineering and Sciences
University of Texas at Austin, Austin, Texas 78712, USA
http://www.math.utexas.edu/users/agillette

# Intro and Disclaimer

These slides are based mainly on talks by Patrice Koehl (UC Davis) and Michael Levitt (Stanford) given at the IMA tutorial on the Mathematics of Proteins in January 2008 at the University of Minnesota. Images from those talks are credited.


*http://www.ima.umn.edu/2007-2008/T1.10-11.08*

# The Fundamental Relationship

*Sequence*

KKAVINGEQIRSISDLHQTLKK
WELALPEYYGENLDALWDCLTG
VEYPLVLEWRQFEQSKQLTENG
AESVLQVFREAKAEGCDITI

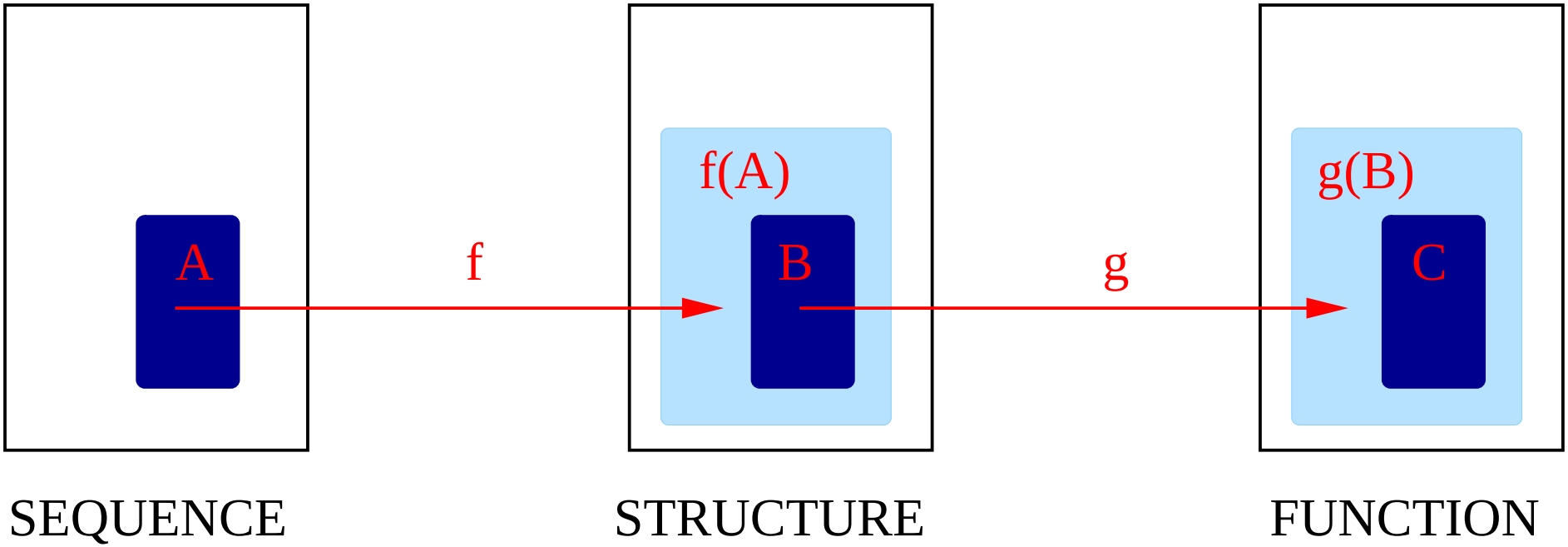
*Structure*

*Function*

*ligand*

image by Patrice Koehl

# Abstract Problem Statement

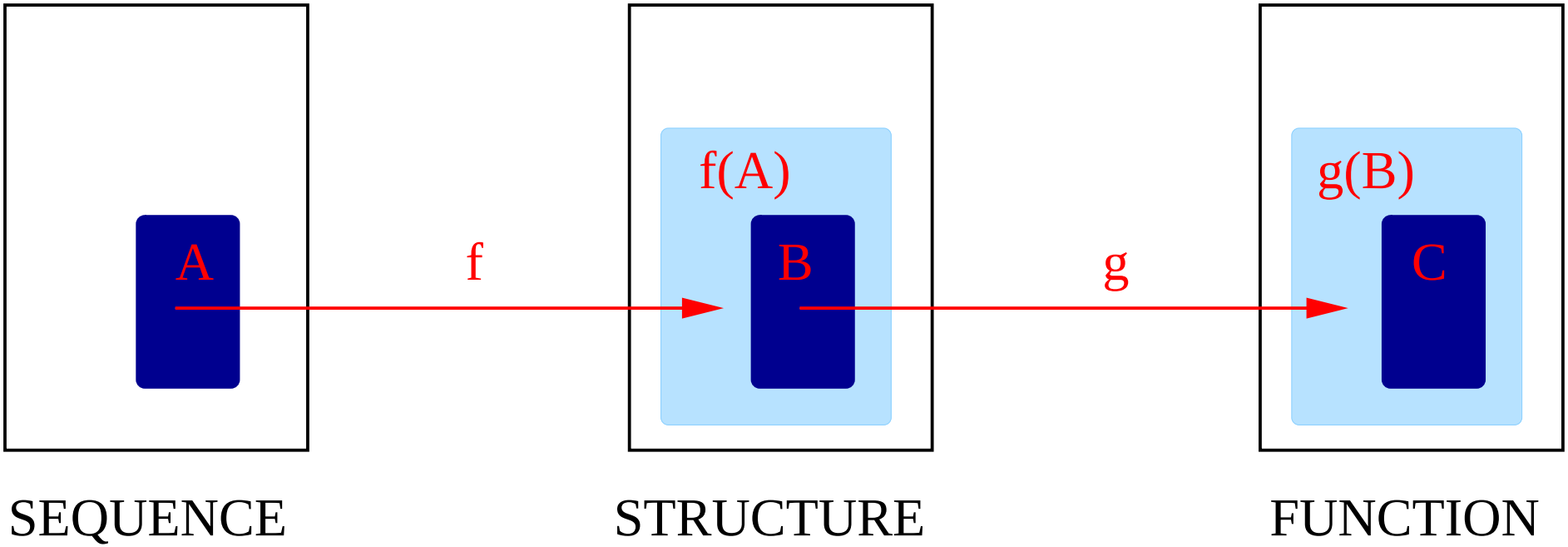



- white region = all possible elements
- light blue region = range of function over known elements
- dark blue region = known elements (i.e. information stored in PDB)

**Abstract Problem Statements:**

1. Expand light and dark blue regions to their maximal size.
2. Create measures on Sequence and Structure space.
3. Determine “formulae” for $f$ and $g$.

# Abstract Problem Statement

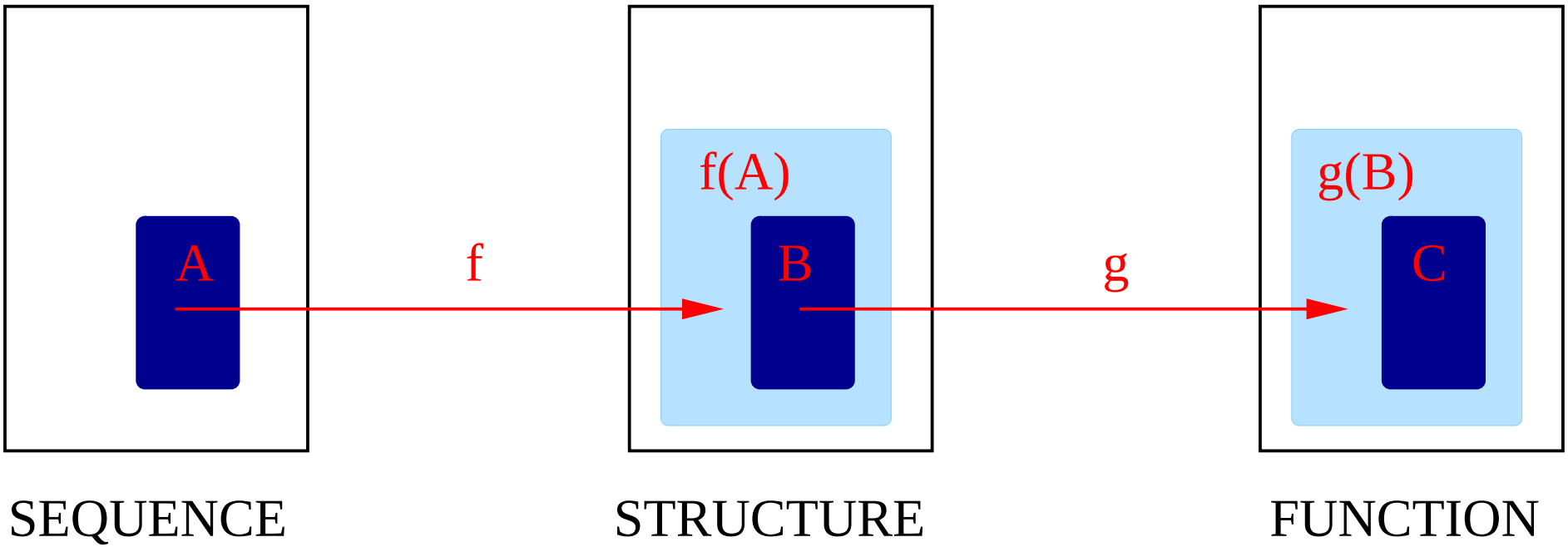



- white region = all possible elements
- light blue region = range of function over known elements
- dark blue region = known elements (i.e. information stored in PDB)

**Abstract Problem Statements:**

1. Expand light and dark blue regions to their maximal size.
2. Create measures on Sequence and Structure space.
3. Determine “formulae” for $f$ and $g$.

# Abstract Problem Statement



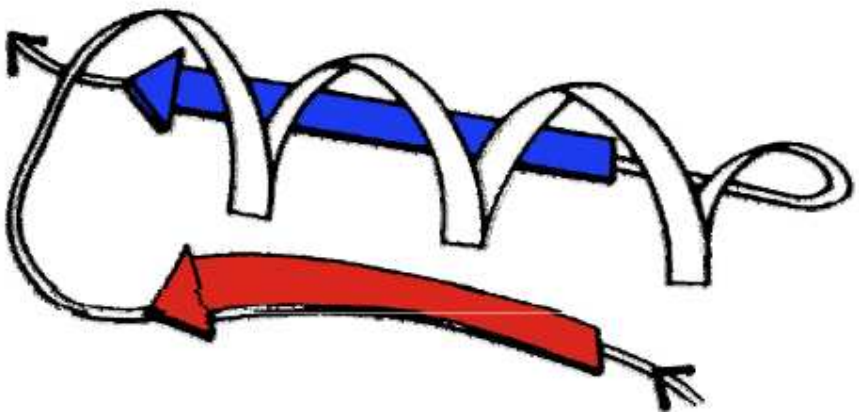

1. Expand light and dark blue regions to their maximal size.

- Protein Data Bank (PDB) currently has ≈48,000 proteins stored.
- Research in this area primarily belongs to biochemists.
- Many subtleties exist, for example. . .

# A Structural Subtlety

The *chirality* of a protein is relevant to its presence in Structure space.

common



very rare

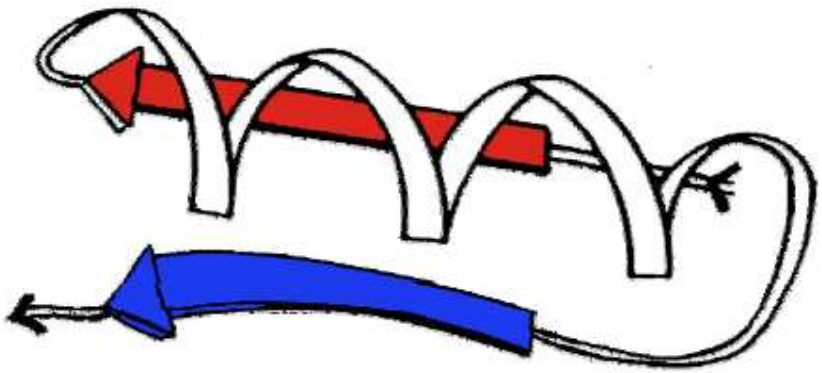
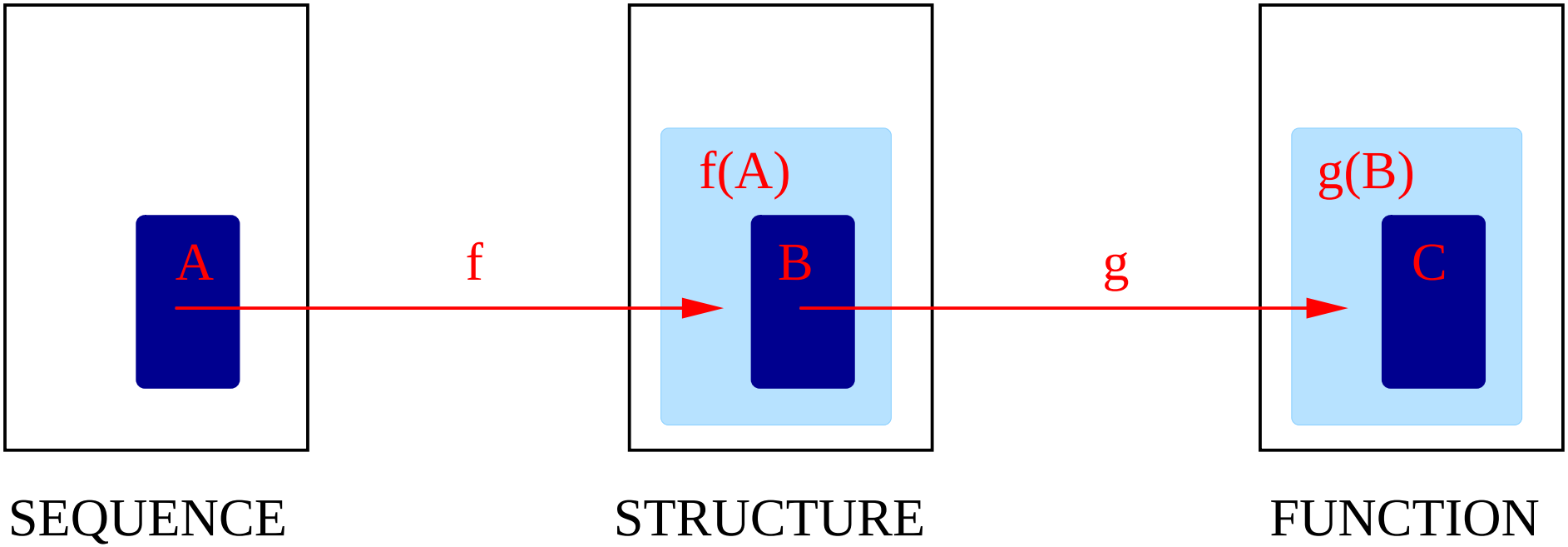


image by Michael Levitt

# Abstract Problem Statement




2. Create measures on Sequence and Structure space.

- A "good measure" on a space $S$ is a function $m : S \times S \to \mathbb{R}_{\geq 0}$ such that the output is small if and only if the two inputs have similar essential properties.
- On Sequence space: a combinatorial problem, similar to phylogenetics.
- On Structure space: Knot theoretical approaches by Isabel Darcy (U. Iowa), Jennifer Mann (UT Math), and others. Differential Geometry approaches by Koehl and others.

# CATH Structure Classification

Class

Architecture

Topology
(read: geometry)

Homology (biological)

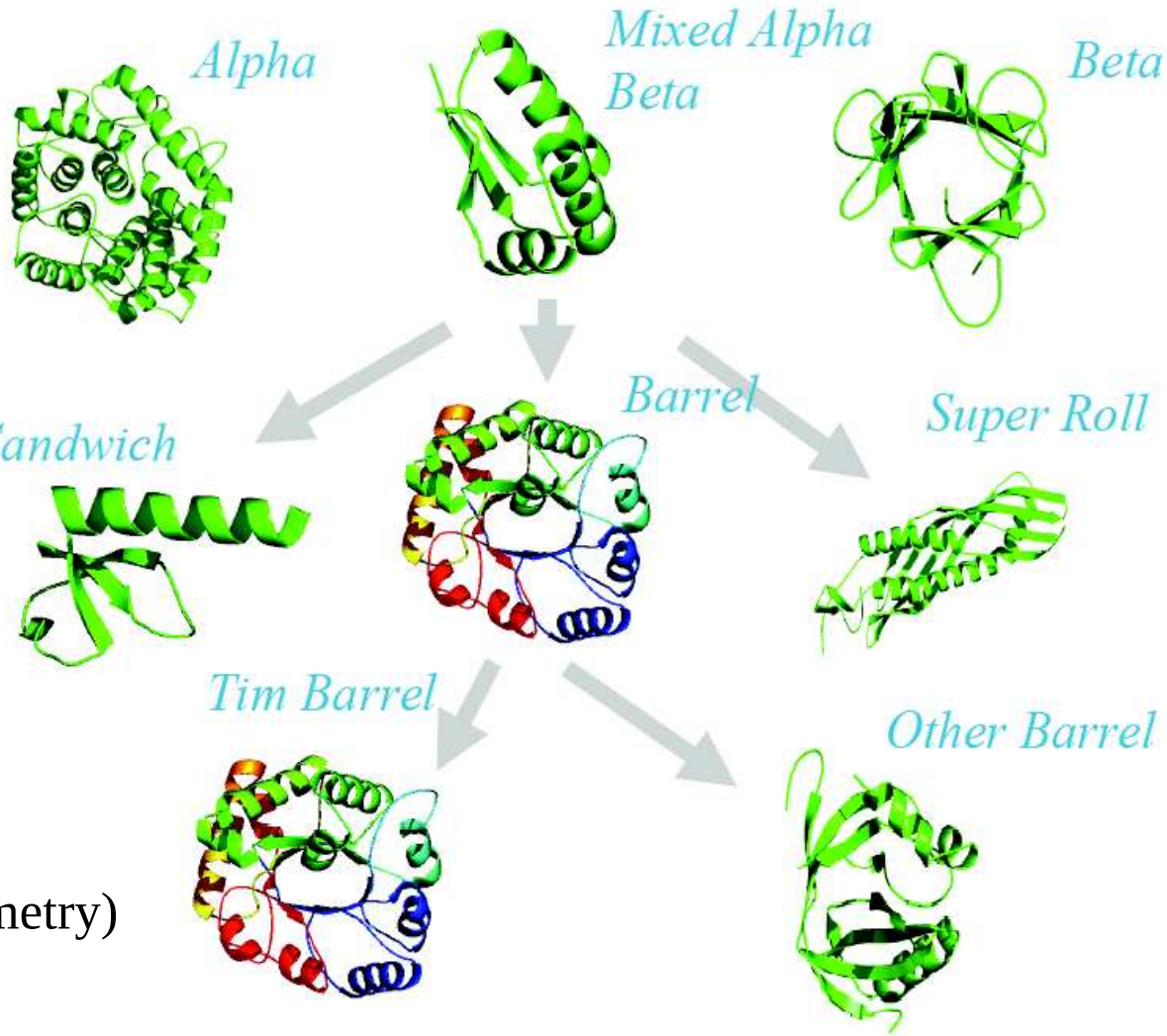
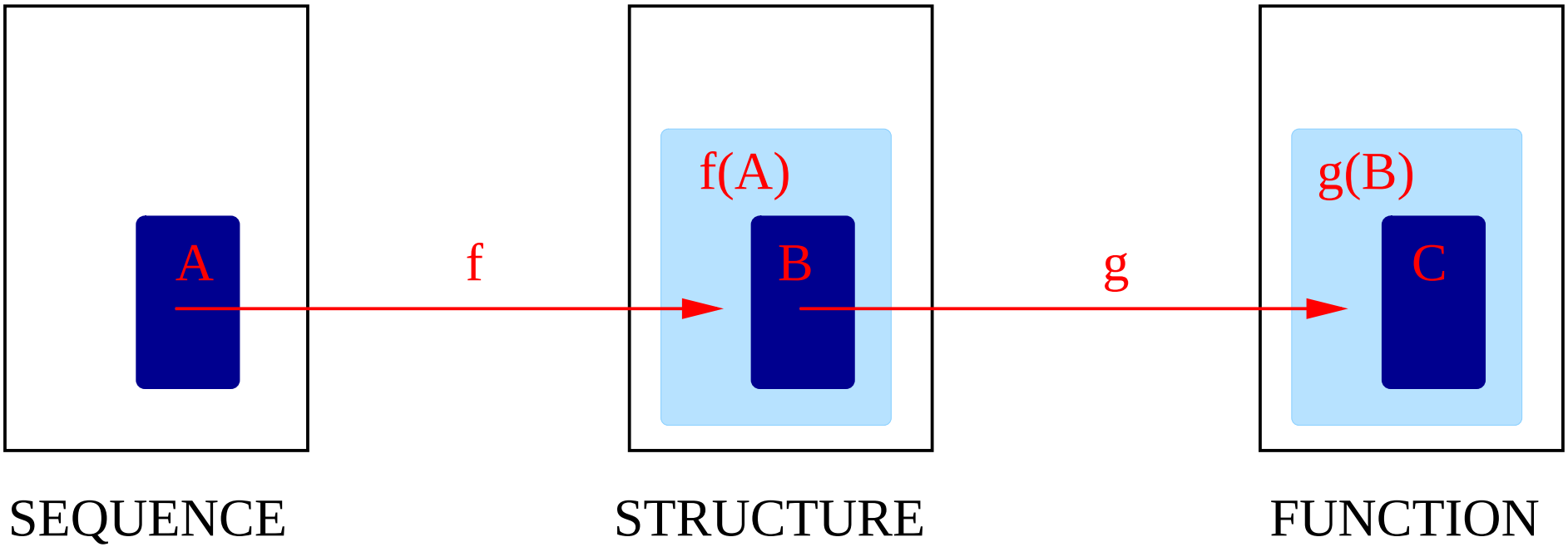



image by Patrice Koehl

# Abstract Problem Statement



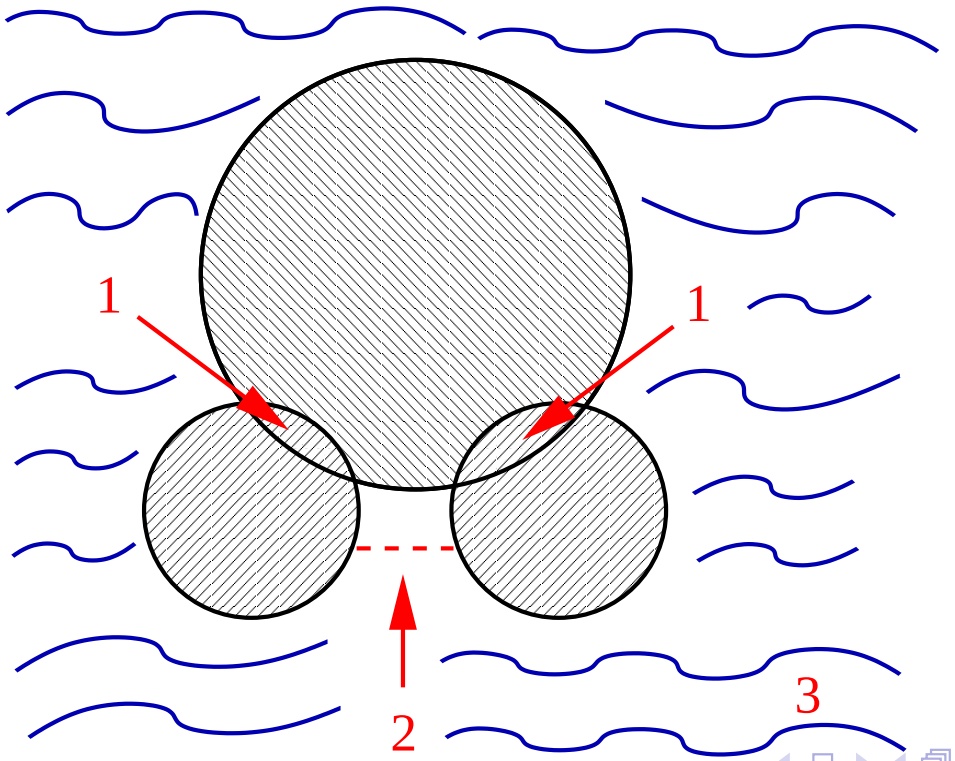

3. Determine "formulae" for $f$ and $g$.

- The function $f$ is not well-behaved: two proteins which are close in structural measure may be distant in sequence measure.
- A large body of work focuses on studying $g$, i.e. how structure dictates function.

# From Structure to Function

A protein's *function* is dependent upon the forces involved in forming and maintaining its *structure*. In order of decreasing importance, these forces are:

1. Bonded interactions (chemistry)
2. Non-bonded interactions (van Der Wal forces, electrostatics)
3. Solvent and environmental forces (*not* negligible)



# An explicit approach?

- Solvent molecules (usually water) surround the protein, move much more rapidly than the protein, and “outnumber” protein atoms $\approx$ 100 to 1.
- Although some explicit models have been attempted for molecular dynamics (see work of Levitt), force calculations are simplified by an implicit model.

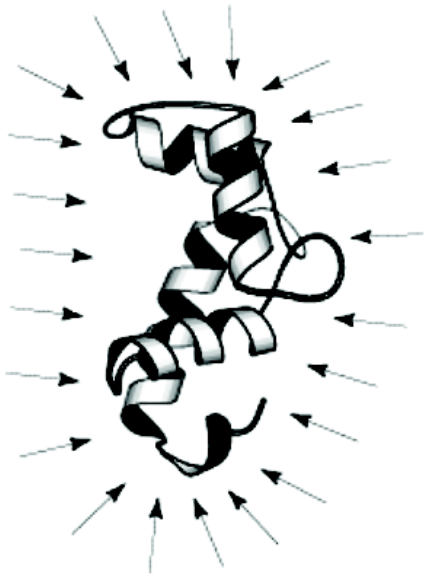


image by Patrice Koehl

# Potential of Mean Force

*A protein in solution occupies a conformation X with probability:*

$$P(X,Y) = \frac{e^{-\frac{U(X,Y)}{kT}}}{\iint e^{-\frac{U(X,Y)}{kT}} dXdY}$$

X: coordinates of the atoms of the protein

Y: coordinates of the atoms of the solvent

*The potential energy U can be decomposed as:*

$$U(X,Y) = U_P(X) + U_S(Y) + U_{PS}(X,Y)$$

$U_P(X)$: protein-protein interactions

$U_S(X)$: solvent-solvent interactions

$U_{PS}(X,Y)$: protein-solvent interactions

(slide by Patrice Koehl)

# Potential of Mean Force

*We study the protein's behavior, not the solvent:*

$$P_P(X) = \int P(X,Y)dY$$

*$P_P(X)$ is expressed as a function of X only through the* ***definition****:*

$$P_P(X) = \frac{e^{-\frac{W_T(X)}{kT}}}{\int e^{-\frac{W_T(X)}{kT}} dX}$$

*$W_T(X)$ is called the potential of mean force.*

(slide by Patrice Koehl)

# Potential of Mean Force

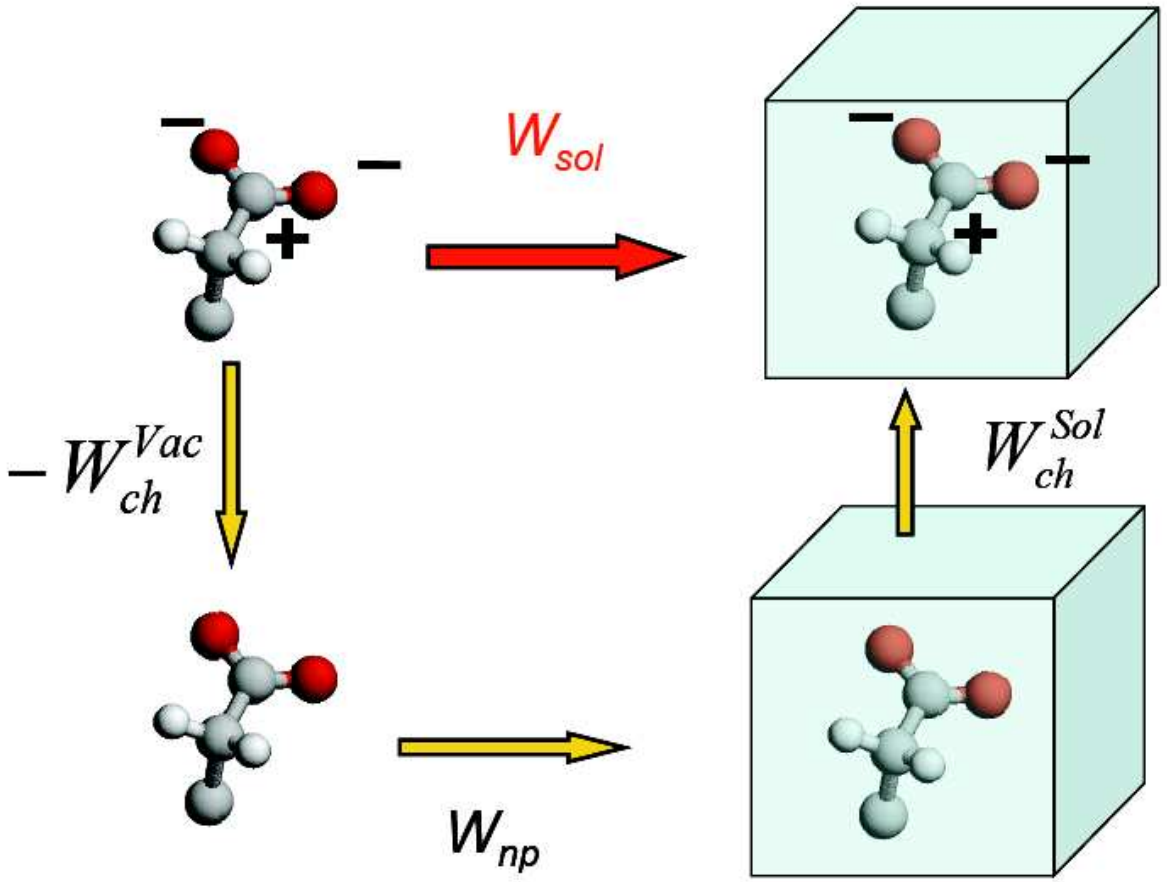
We calculate the Potential of Mean Force by the following:

$$W_T(X) = U_P(X) + W_{sol}(X)$$

- $U_P(X)$ accounts for the internal energy of protein
- $W_{sol}(X)$ accounts ***implicitly*** and ***exactly*** for the effect of the solvent on the protein

We estimate $W_{sol}(X)$ by an artificial thermodynamics path.

# Solvation of Free Energy

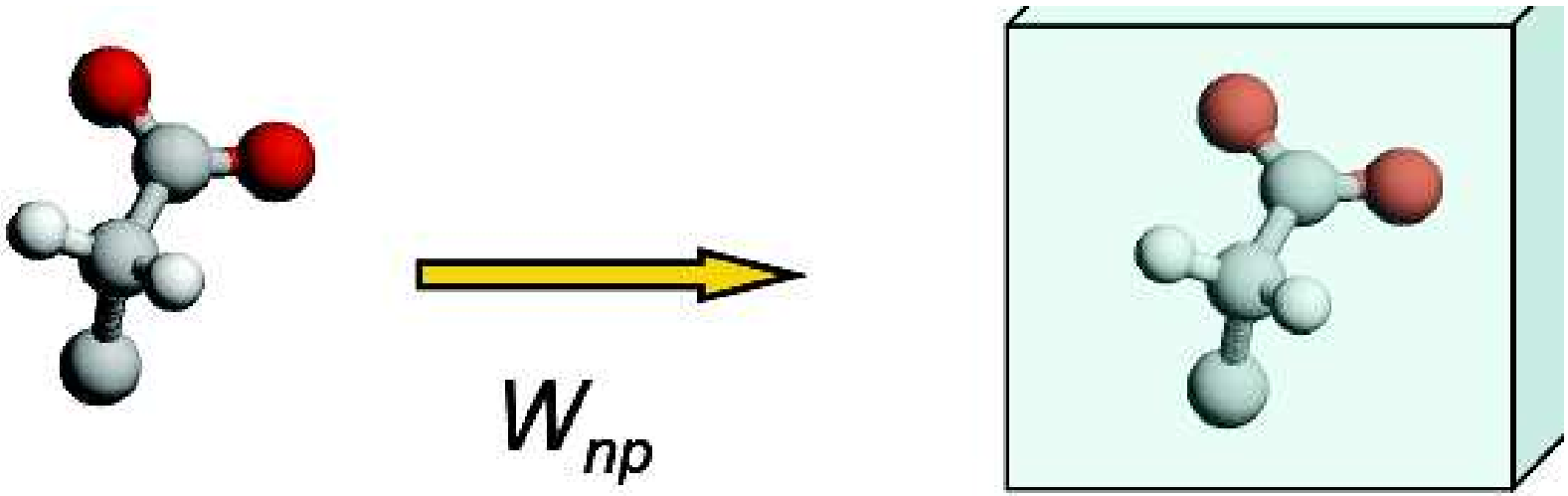



$$W_{sol} = W_{elec} + W_{np} = \left(W_{ch}^{sol} - W_{ch}^{vac}\right) + \left(W_{vdW} + W_{cav}\right)$$

image by Patrice Koehl

# Solvation of Free Energy

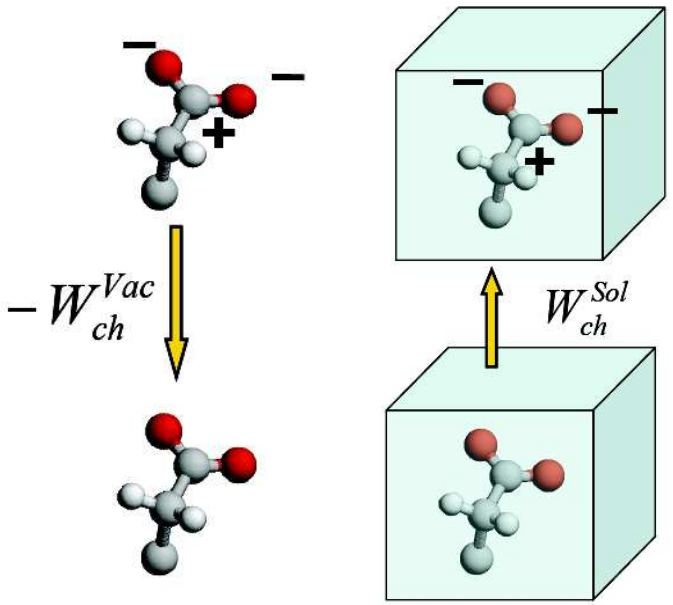
$W_{np}$ measures the purely physical effect of putting the protein in water.




This is a **geometrical** question and is tackled by members of our lab, as well as others.

# Solvation of Free Energy

$W_{ch}^{Vac}$ and $W_{ch}^{Sol}$ measure the electrostatic forces of the protein with or without the presence of water.




These forces are computed by means of the Poisson Boltzmann equation. Such computations rest upon some crucial assumptions.

# PB problems

Problems with the basic Poisson Boltzmann equation:

- **Dimensionless ions**: Number of ions present depends upon the *size* of ions in question.
- **Ion-ion interactions**: Not accounted for.
- **Ion-solvent interactions**: Not accounted for.
- **Non-uniform solvent concentration**: Water forms a hydration layer around protein.
- **Polarizaion**: Assumed to be a linear response to $E$ with constant proportion.

An alternative approach is to assume a *density* of dipoles with constant module $\rho_0$ yielding the **Generalized Poisson-Boltmann Langevin Equation**, however, . . .

# Generalized PB Langevin Equation

$$\frac{\beta}{4\pi}\bar{\nabla}\cdot\left(\varepsilon\bar{\nabla}\Phi(\bar{r})\right)+\beta\rho_f(\bar{r})=-\frac{2\beta\lambda_{ion}\sinh\left(\beta ez\Phi(\bar{r})\right)}{a^3D\left(\Phi(\bar{r})\right)}+\frac{\beta^2p_o^2\lambda_{dip}F_1(u)\bar{\nabla}\cdot\left(\bar{\nabla}\Phi(\bar{r})\right)}{a^3D\left(\Phi(\bar{r})\right)}$$
$$+\frac{\beta^4p_o^4\lambda_{dip}F_1'(u)\bar{\nabla}\Phi(\bar{r})\cdot\left(\bar{\nabla}\Phi(\bar{r})\cdot\bar{\nabla}\right)\bar{\nabla}\Phi(\bar{r})}{a^3D\left(\Phi(\bar{r})\right)u}$$
$$-\frac{2\beta^2p_o^2\lambda_{ion}\lambda_{dip}F_1(u)\left\|\bar{\nabla}\Phi(\bar{r})\right\|^2\beta ez\sinh\left(\beta ez\Phi(\bar{r})\right)}{a^3D\left(\Phi(\bar{r})\right)^2}$$
$$-\frac{\beta^4p_o^4\lambda_{dip}^2\left(F_1(u)\right)^2\bar{\nabla}\Phi(\bar{r})\cdot\left(\bar{\nabla}\Phi(\bar{r})\cdot\bar{\nabla}\right)\bar{\nabla}\Phi(\bar{r})}{a^3D\left(\Phi(\bar{r})\right)^2}$$

*with*

$$D\left(\Phi(\vec{r})\right)=1+2\lambda_{ion}\cos\left(\beta ez\Phi(\vec{r})\right)+\lambda_{dip}\frac{\sinh\left(\beta p_o\left\|\vec{\nabla}\Phi(\vec{r})\right\|\right)}{\beta p_o\left\|\vec{\nabla}\Phi(\vec{r})\right\|}$$

$$F_1(u)=\frac{1}{u}\frac{\partial}{\partial u}\left(\frac{\sinh(u)}{u}\right)=\frac{1}{u}\left(\frac{u\cosh(u)-\sinh(u)}{u^2}\right)$$

***and***

$$u=\beta p_0\left\|\vec{E}\right\|=\frac{p_0\left\|\vec{E}\right\|}{k_BT}$$

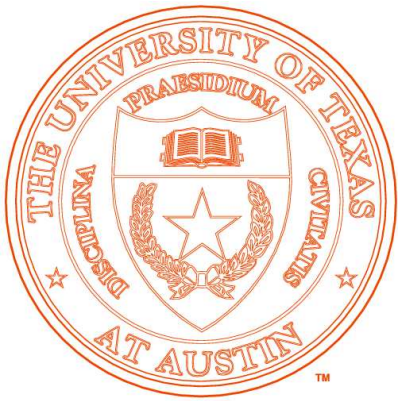
(slide by Patrice Koehl)

# References

PATRICE KOEHL *http://nook.cs.ucdavis.edu/ koehl/*

MICHAEL LEVITT *http://csb.stanford.edu/levitt/*

IMA TUTORIAL *http://www.ima.umn.edu/2007-2008/T1.10-11.08*




For a copy of these slides, please visit

*http://www.math.utexas.edu/users/agillette/*