Ab Initio Protein Folding

Homayoun Valafar

Department of Computer Science and Engineering, USC

From Sequence to Structure

- Does primary sequence lead to functional structure?
  - Isolate functional protein
  - Denature using urea or high temperature
  - Confirm loss of function
  - Reinstate folding conditions (remove urea, lower temp.)
  - Confirm gain of function
- In general protein sequence leads to functional structure
- Simulation of physical forces should allow computational folding of proteins
  - Levitt, M. and A. Warshel, *Computer simulation of protein folding. Nature, 1975. 253: p. 694-698.*

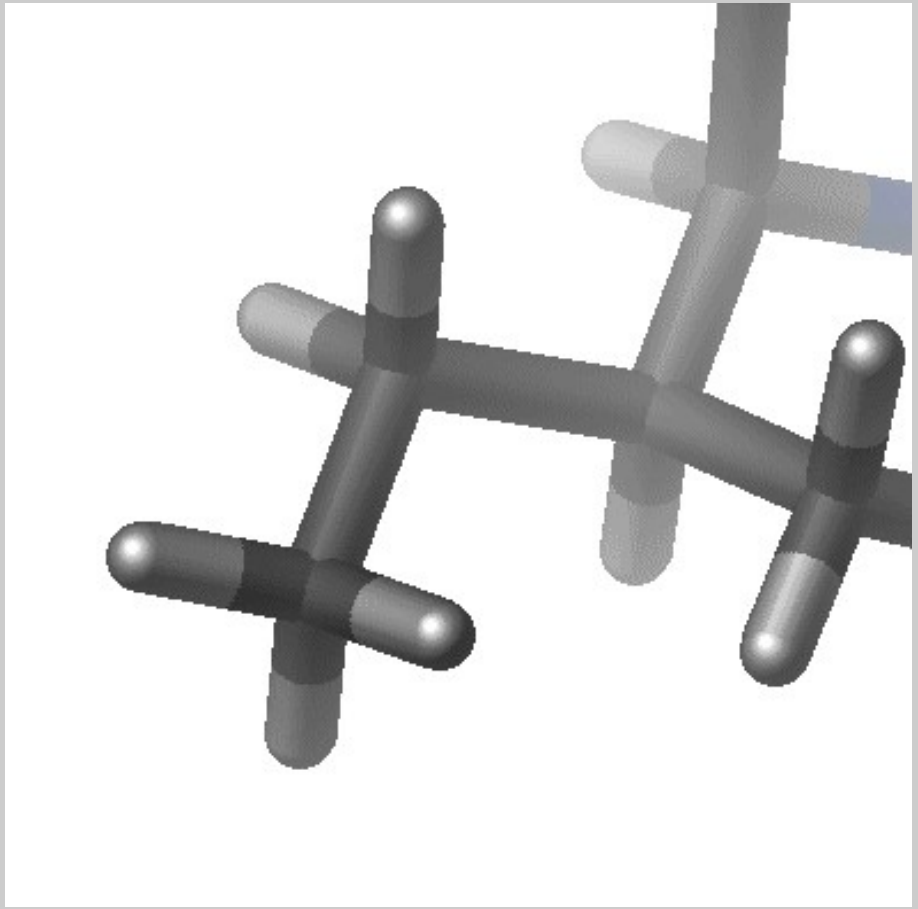
Total Potential Energy

- Mathematical expression of the potential function is necessary for simulation of protein fold
- $E_{Total} = E_{Empirical} + E_{Effective}$
  - $E_{Empirical}$: energy of the molecule as a function of the atomic coordinates
  - $E_{Effective}$: restraining energy terms that use experimental information
- Neglect $E_{Effective}$ term for true computational model
- Structure with the lowest total energy is the most stable structure
  - Justified by laws of Thermodynamics

Potential Energy of Bond Lengths

- The bond length a pair of atoms is known empirically
- Bond lengths should not exceed the expected values
- Defined by two atoms

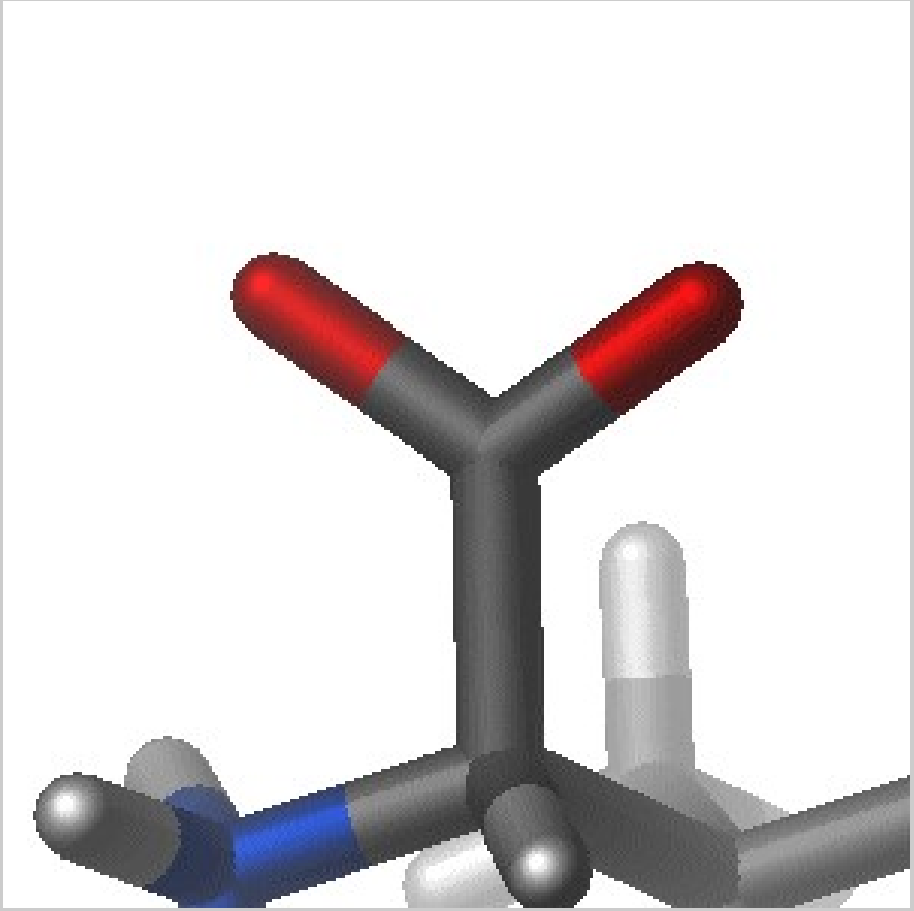
$$E_{Bond} = \sum_{bond} k_b (r - r_0)^2$$



Potential Energy of Bond Angles

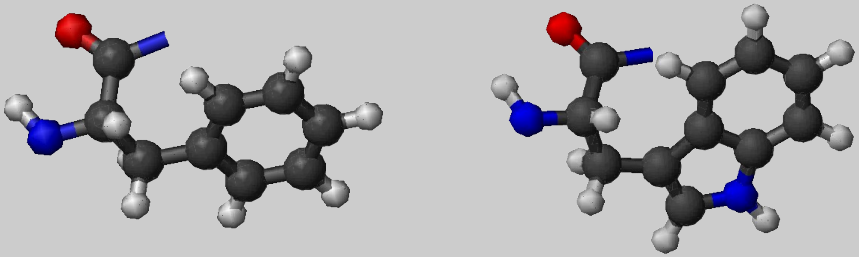
- Bond angles should not deviate from the known quantities
- Involves three atoms

$$E_{Angles} = \sum_{angles} k_\theta (\theta - \theta_0)^2$$



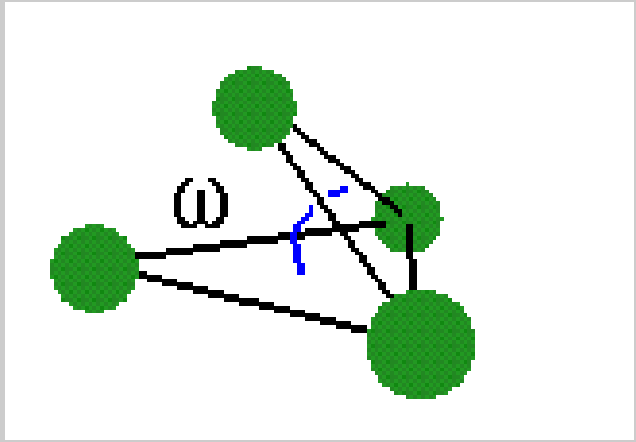
P.E. of Improper Dihedrals

- Improper dihedrals represent the planarity of a group of atoms
  - Peptide plane
  - Aromatic side chains: phenylalanine, tryptophan, tyrosine, histidine



- Four atoms are required for this measure

$$E_{Impr} = \sum_{impropers} k_i(\omega - \omega_0)^2$$



Empirical Energy Terms

- All energies defined in terms of atomic coordinates of two, three or four atoms
- Conformational Energy Terms:
  - E_{BOND} : describes the covalent bond energy over all covalent bonds
  - E_{ANGL} : describes the bond angle energy over all bond angles
  - E_{DIHE} : describes the dihedral angle energy over all dihedrals
  - E_{IMPR} : describes the improper angle energies (planarity and chirality)
- Nonbonded Energy Terms:
  - E_{VDW} : describes the energy of Van Der Waals terms
  - E_{ELEC} : describes the energy of electrostatic interactions

Other Potential Terms

- Hydrophobic and hydrophilic interaction.
  - Requires presence of water in the simulation.
  - Addition of water to the simulation is difficult.
  - Will require identification of cavities and calculation of movement of water molecules.
- Hydrogen bonds:
  - Also requires assessment of water accessibility.
  - Water interferes with formation of hydrogen bonds.
- Gas phase simulation
  - Absence of water.
  - Computationally much more convenient

Total Energy Term

- E_{Total} is the total potential energy of a conformation
- Force Field: A vector field representing the gradient of the total potential
- ω is referred to as the force constants

$$E_{Total} = \sum \left[w^p_{BOND} E_{BOND} + w^p_{ANGL} E_{ANGL} + w^p_{DIHE} E_{DIHE} + w^p_{IMPR} E_{IMPR} + w^p_{VDW} E_{VDW} + w^p_{ELEC} E_{ELEC} \right]$$

$$E_{BOND} = \sum_{bonds} k_b (r - r_0)^2$$

$$E_{ANGL} = \sum_{angles} k_\theta (\theta - \theta_0)^2$$

$$E_{DIHE} = \sum_{dihedrals} \sum_{i=1,m} \begin{cases} k_{\varphi_i}(1+\cos(n\varphi_i+\delta_i)) & n_i ¿ 0 \\ k_{\varphi_i}(\varphi_i-\delta_i)^2 & n_i = 0 \end{cases}$$

$$E_{IMPR} = \sum_{impropers} \sum_{i=1,m} \begin{cases} k_{\varphi_i}(1+\cos(n\varphi_i+\delta_i)) & n_i > 0 \\ k_{\varphi_i}(\varphi_i-\delta_i)^2 & n_i = 0 \end{cases}$$

$$E_{VDW} = \sum_{vdw} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}$$

$$E_{ELEC} = \sum_{i,j} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

Force Field

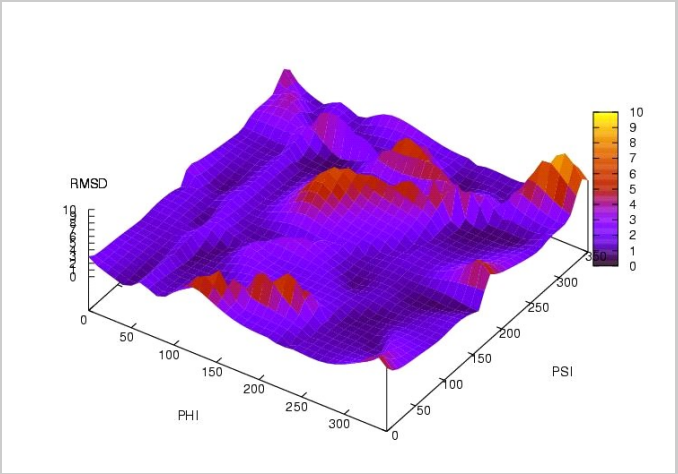
- Technically, the derivate of the potential energy
  - A vector field of forces
- Some currently existing force fields (forcefield):
  - Xplor-NIH
  - AMBER
  - CHARMm
  - MM2, MM3 and MM4
  - Sybyl
  - Etc.

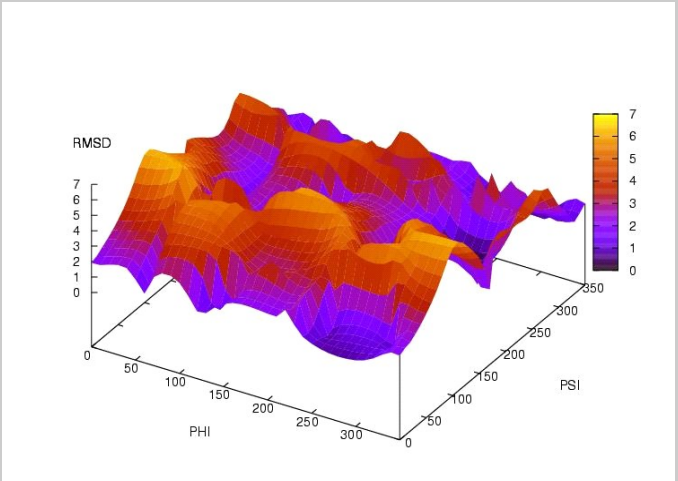
Minimization of Total Energy

- Theoretically, the structure with the minimum total energy is the structure of interest.
- A number of minimization algorithms can be utilized.
  - Gradient descent
  - Monte Carlo and Simulated Annealing
  - Newton's
  - Genetic Algorithm
  - Distributed Global Optimization
  - Branch and Bound
  - …

Complexity of The Problem

- Assuming a protein with 100 residues and in average 10 atoms per residue, what is the complexity of this problem?
- What are the variables of this problem? How many?
- How complex is the total energy landscape?
- How costly is each evaluation of the E_{Total} and its gradient?
- Beyond our computational capabilities.





Computational Complexity of Protein Folding

- For a protein of size N amino acids:
  - $df = 2 \times (N - 1)$
  - Each degree of freedom spans 0º-360º
  - Possible conformations at 10º resolution: $36^{2(N-1)}$
  - N = 100, 10^6 struct / sec ⇒ 4.4575E+291 millennia
  - NP class of problems.
  - N=11 ⇒ 32 millennia
  - N=11, 50 angles ⇒ 1 millennium