CISC 436/636 Computational Biology & Bioinformatics (Fall 2016)

Protein Structure Prediction

3-Dimensional Structure

Foundation

- Anfinsen Hypothesis: A native state of protein corresponds to a free energy minimum. (This hypothesis was by Anfinsen based on some experimental findings for smaller molecules, 1973)
- Levinthal paradox: how can a protein fold to a native conformation so rapidly when the number of possible conformations is so huge?
  - Experimental observation: proteins fold into native conformations in a few seconds
  - If a local move by diffusion takes 10^{-11} seconds, it would take 1057 seconds to try ~ 5^{100} possible conformations for a protein of 100 AAs.
- NP-completeness

Computational Methods for 3-D structures

- Comparative (find homologous proteins)
- Threading (recognize folds)
- *Ab initio* (Molecular dynamics)

Root Mean Square Deviation (RMSD)

$$\sqrt{\frac{\sum_{i=1,N}(\vec{x}_i - \vec{y}_i)^2}{N}}$$

Where x_i are the coordinates from molecule 1 and

y_i are the *equivalent** coordinates from molecule 2.

*Which atoms are *equivalent* is based on an **alignment**.

Credit: Chris Bystroff @ RPI

Comparative (homology based) modeling

- Identification of structurally conserved regions (using multiple alignment)
- Backbone construction
- Loop construction
- Side-chain restoration
- Structure verification and evaluation
- Structure refinement (energy minimization)

Least squares superposition

Problem: find the rotation matrix, $\underline{M}$, and a vector, v, that minimize the following quantity:

$$\sum_i \left| \underline{M}\vec{x}_i + \vec{v} - \vec{y}_i \right|^2$$

Where x_i are the coordinates from one molecule and

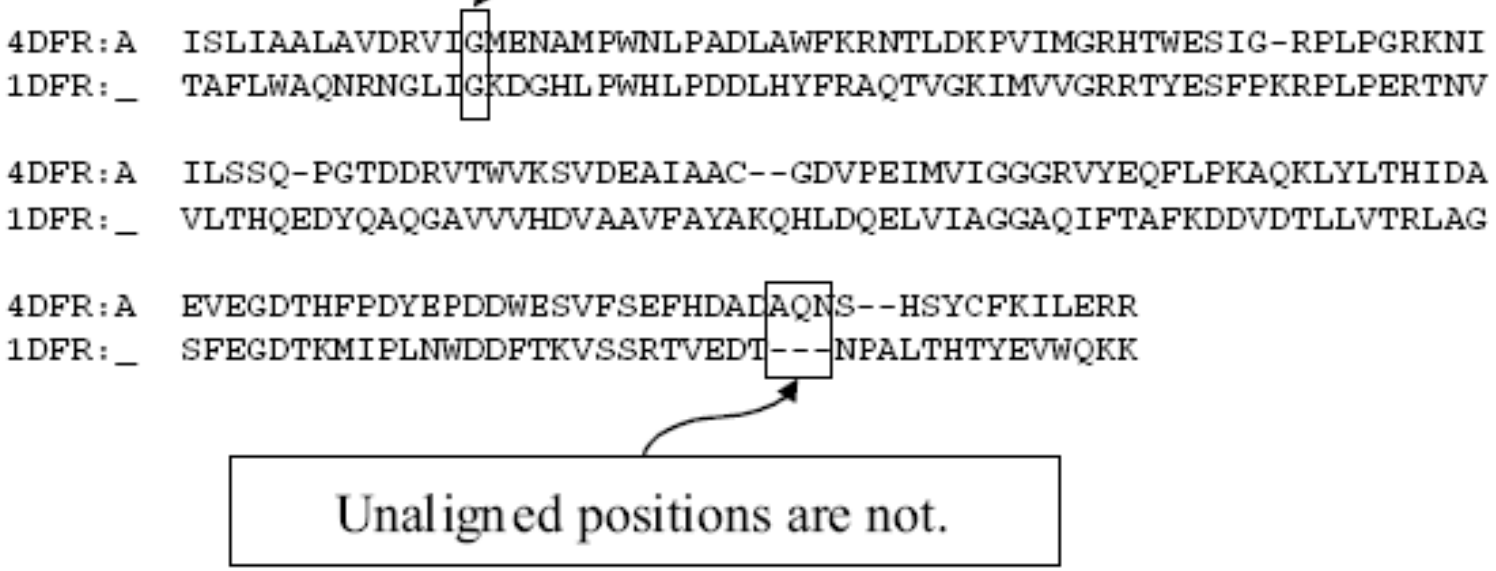
y_i are the *equivalent** coordinates from another molecule.

**equivalent* based on alignment

Credit: Chris Bystroff @ RPI

Mapping structural equivalence = aligning the sequence

Any position that is aligned is included in the sum of squares.



Credit: Chris Bystroff @ RPI

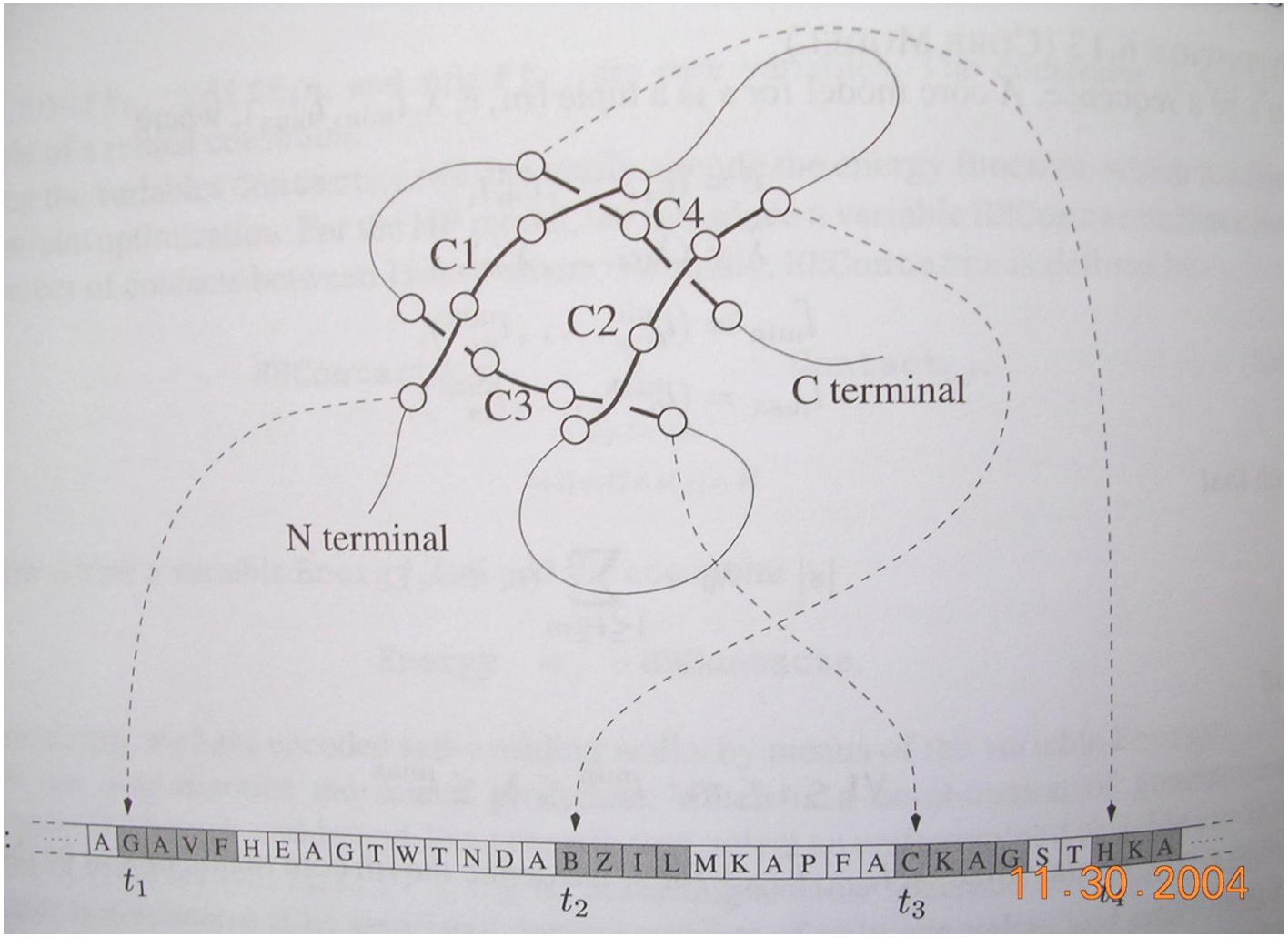
least-squares superimposed molecules

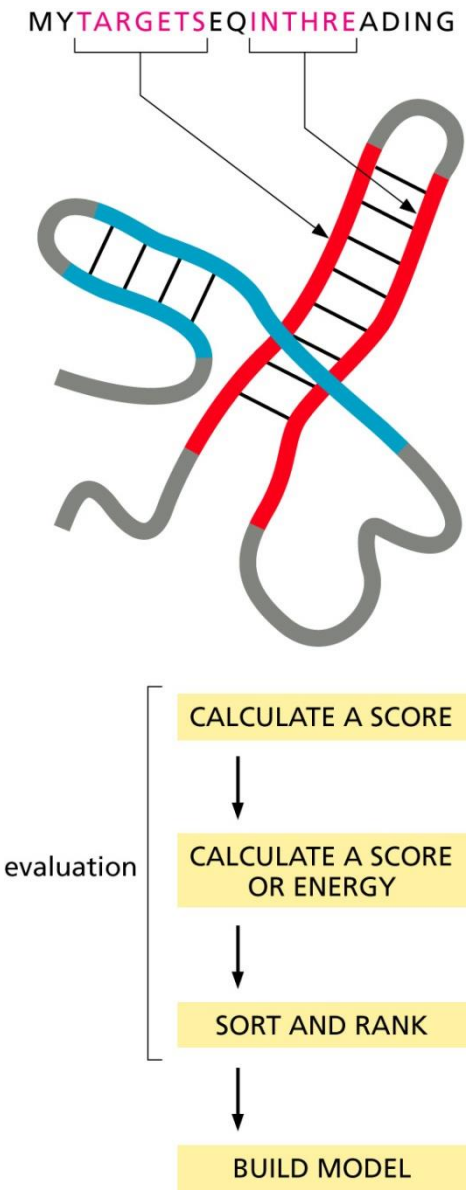
Credit: Chris Bystroff @ RPI

Protein Threading

- When two homologous proteins do not have high sequence similarity
- Structure of one protein, P1, is known
- The goal is to predict the structure for the other protein P2
- Identify core regions for P1
- Scoring a given alignment with a known structure (Contact potentials)

Threading: A schematic view





First, segments of sequence are structurally aligned (threaded) on to a fold and a score/energy is obtained for each alignment.

A dynamic programming technique is used to find the alignment that has the best score.

This is done for each fold in the fold library, and the results are ranked. The folds giving the best score are then selected for use in modeling the query sequence.

Lattice Models

Simplifications

- All residues have the same size
- Bond length is uniform
- Positions of residues are restricted to positions in a regular lattice (or grid).

HP model

- Energy function $B_{i,j}$ for a pair of residues w_i and w_j
  - = 0 when the two residues
    - do not have contact (i.e. not topological neighbors), or
    - one residue is polar/hydrophilic and the other is hydrophobic, or
    - both are polar/hydrophilic
  - = -1 when two residues
    - are topological neighbors, and
    - both are hydrophobic.

Note: Despite these simplifications, it is still NP complete to find an optimal conformation

- 2-dimensional and 3-dimensional lattice
- Two residues w_i and w_j are *connected neighbors* when j = i+1 or i-1.
- Two residues w_i and w_j are *topological neighbors* when $j \neq i+1$ or i-1, and

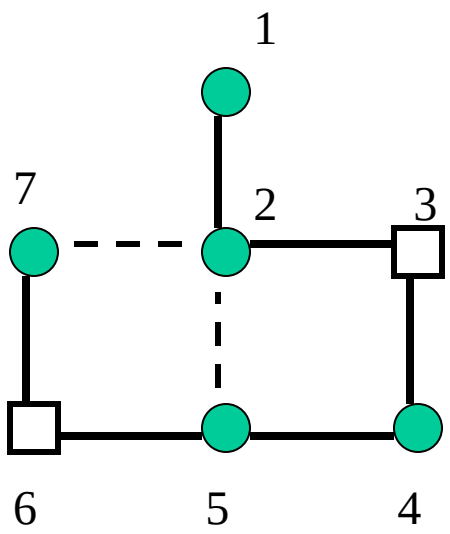
$$\| w_i - w_j \| = 1$$

- A native state is a conformation that has the minimum contact energy

$$E = \sum_{1 \leq i+1 < j \leq n} B_{i,j}\, \delta(w_i, w_j)$$

  where $\delta(w_i, w_j) = 1$ when w_i and w_j are *topological neighbor*, and =0 otherwise.
- The HP model approximates the hydrophobic force, which is not really a force, but rather an aggregate tendency for nonpolar residues to minimize their contact with the solvent.

- Example HP model



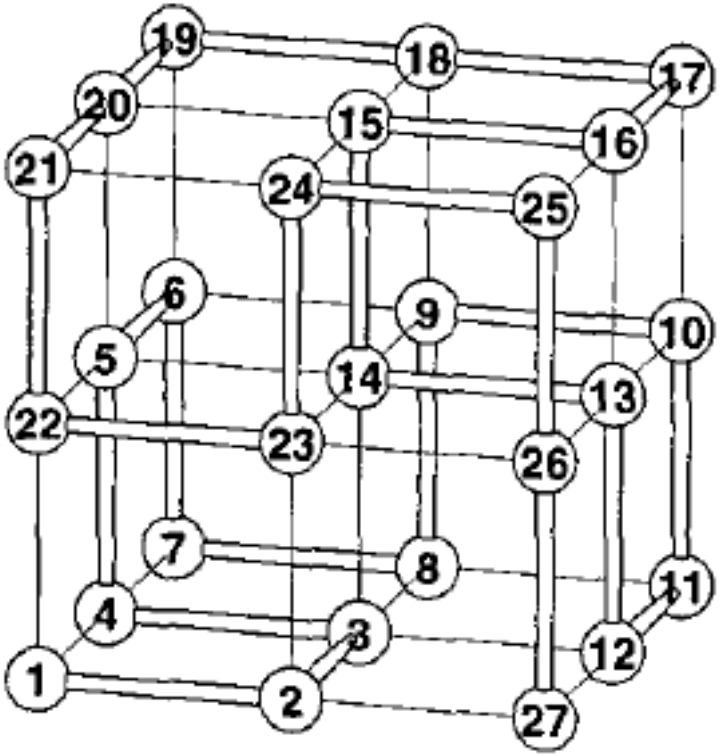
P

H

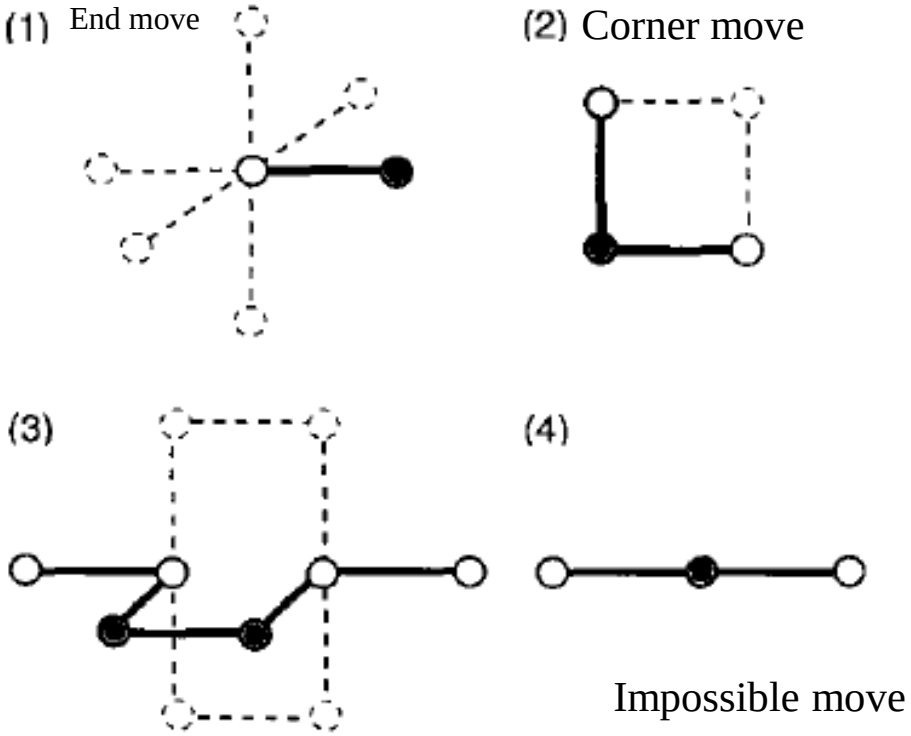
Peptide bond

Topological contact

- The HP model by SSK (Sali, Shakhnovich, and Karplus, 1994)
  - A 27-bead heteropolymer in a 3-d lattice.
  - Contact energy normally distributed
  - Possible conformation: $5^{26} \sim 10^{18}$
  - Compact conformation: in a 3x3x3 cube there are 103346 distinct conformations
  - Folding time t0 is short if energy gap is large
  - Solved Levinthal paradox


19
18
17
20
15
16
21
24
25
6
9
10
5
14
13
22
23
26
7
8
11
4
3
12
1
2
27

Local moves

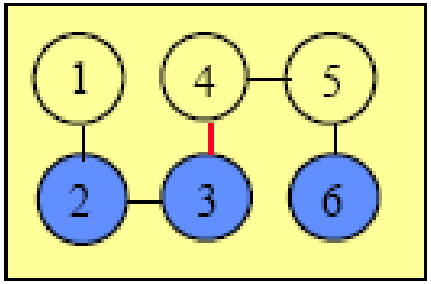


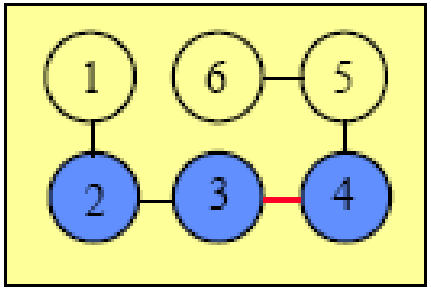
Crossover to create new conformations

Parents

10 00 01 00 10

10 00 00 01 11

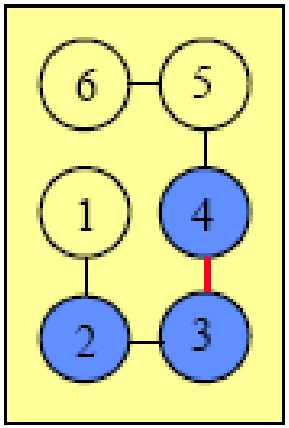


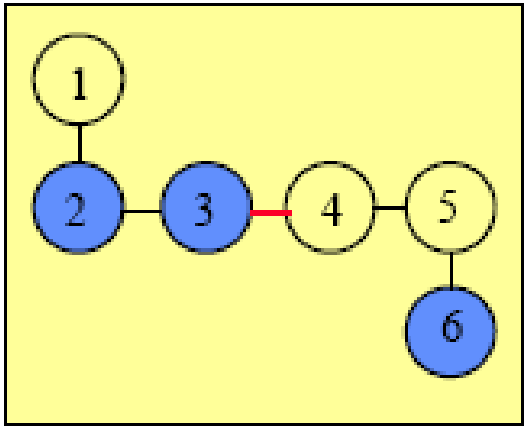


Children

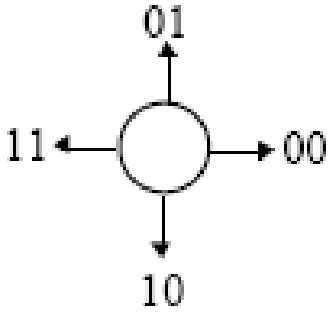
10 00 01 01 11

10 00 00 00 10





Encoding:



Credit: Iosif Vaisman @ GAU

Genetic algorithm

Input
- P, the population,
- r: the fraction of population to be replaced,
- f, a fitness,
- ft, the fitness_threshold,
- m: the rate for mutation.

Initialize population (randomly)

Evaluate: for each h in P, compute Fitness(h)

While [Max_h f(h)] < ft

do

1. Select
2. Crossover
3. Mutate
4. Update P with the new generation Ps
5. Evaluate: f(h) for all h $\in$ P

Return the h in P that has the best fitness

Unger-Moult hybrid genetic algorithm

```
initialize population P(t) of random coils
best = argmax {F(x) | x in P(t)}
repeat {
        pointwise mutation
        n = 0
        while (n < P) { // genetic algorithm part
                select 2 chromosomes m, f
                produce child c by crossover of m, f
                ave = average( F(m), F(f))
                if(F(c)  >= ave) {
                        place c in next generation; n++
                } else {   // Metropolis part
                        z = random(0,1)
                        if (z < e -(ave –F(c)/T  ) {
                            place c in next generation; n++
                        }
                }
         }// end while
         update best
         check if stopping criterion is met
         }
```

Ab initio approaches

Potential Energy Function

$$PEF(R) = \sum_{bonds} K_b\{b(R)-b_{eq}\}^2 + \sum_{angles} K_\theta\{\theta(R)-\theta_{eq}\}^2 +$$

$$\sum_{dihedrals} \frac{K_\psi}{2}\{1 + \cos[n\phi(R) - \gamma]\} +$$

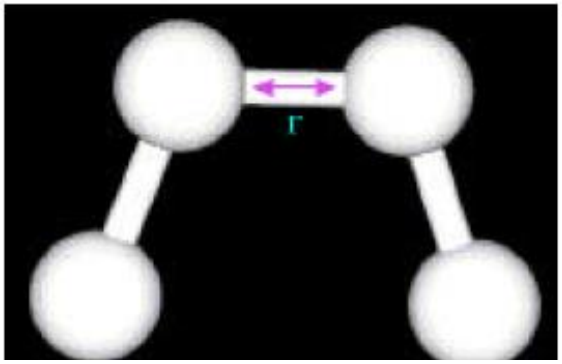
$$\sum_{\substack{non\text{-}bonded \\ atom\ pairs\ i,j}} \left[\frac{A_{ij}}{r_{ij}(R)^{12}} - \frac{B_{ij}}{r_{ij}(R)^{6}} + \frac{q_i q_j}{\varepsilon_r \varepsilon_0 r_{ij}(R)} \right] \quad (1)$$

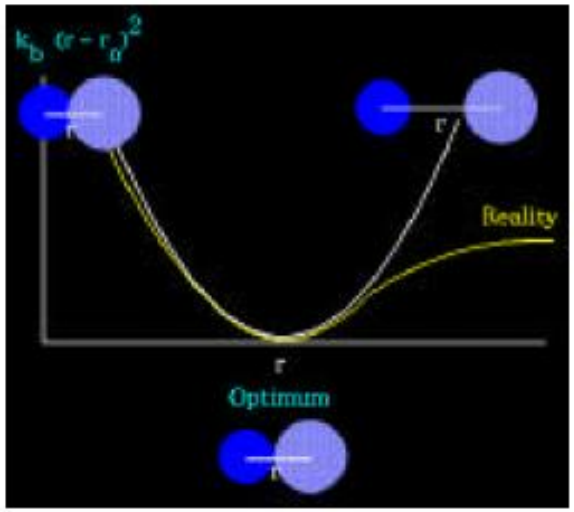
Forcefields: AMBER, CHARMM, CVF, ECEPP, GROMOS

Credit: Iosif Vaisman @ GMU

Bond length

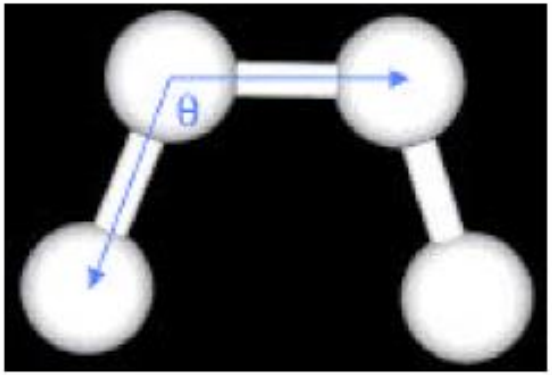
$$E = \sum_{bonds} k_b (r - r_o)^2$$

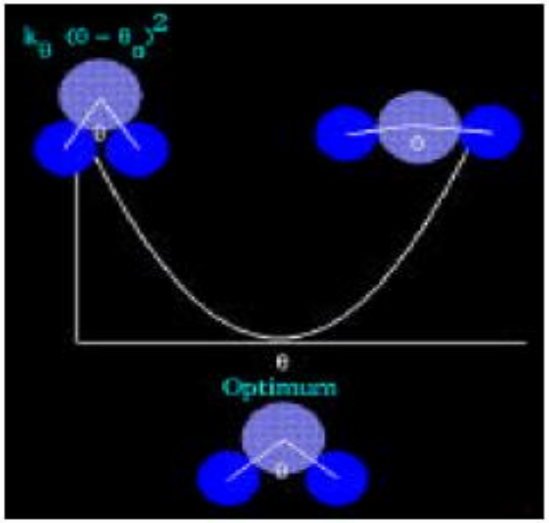




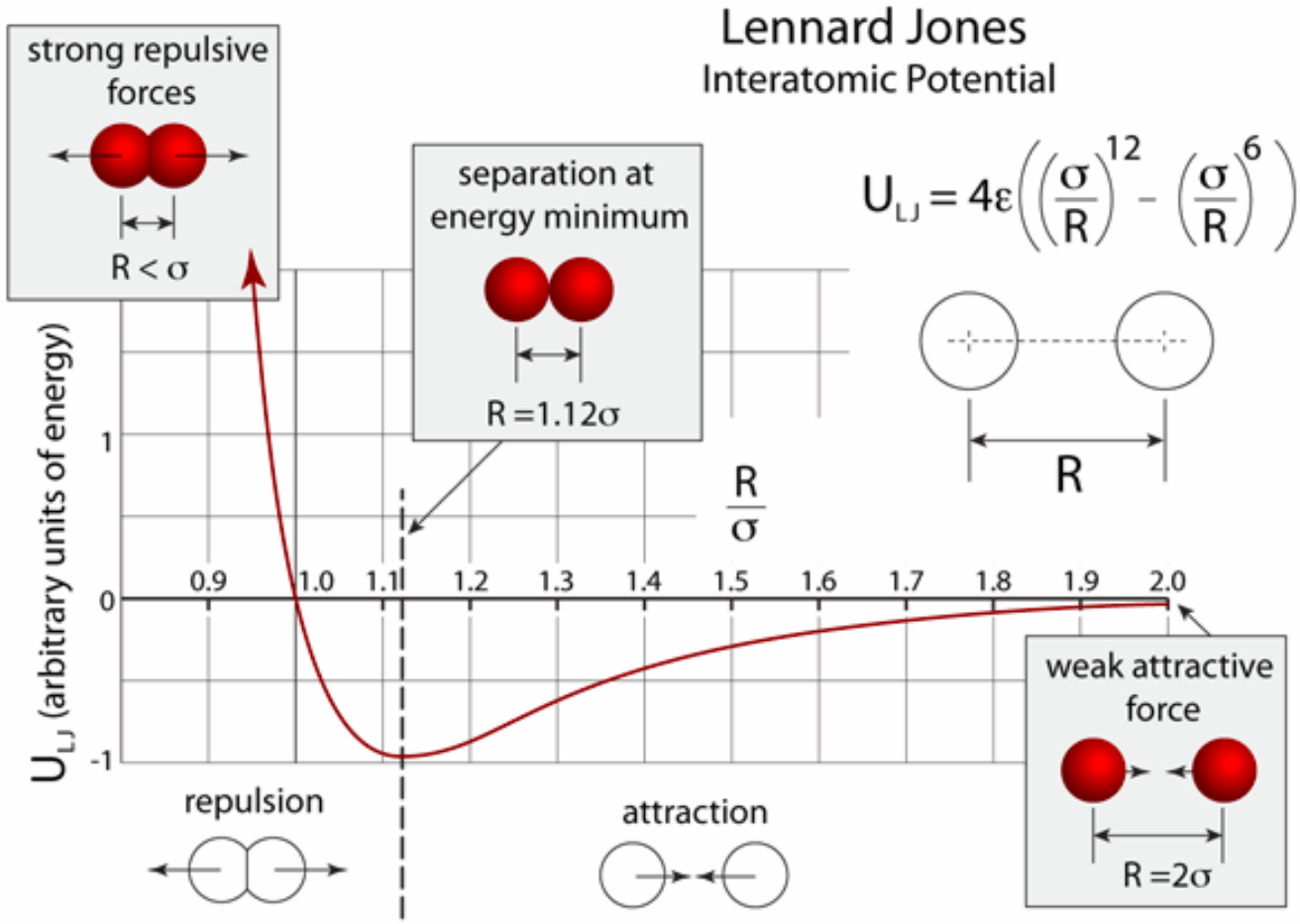
Bond angle

$$E = \sum_{angles} k_\theta (\theta - \theta_o)^2$$





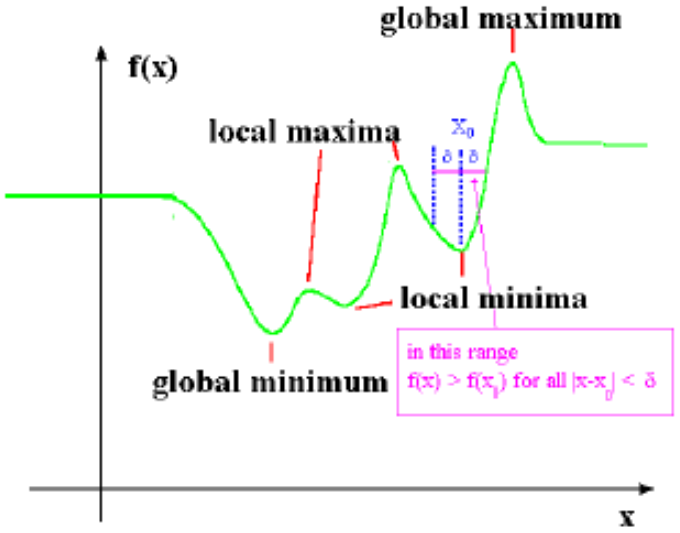
Credit: Iosif Vaisman @ GMU



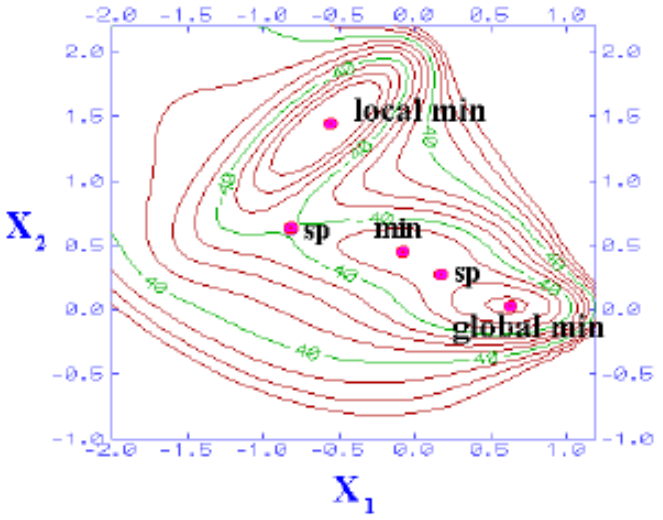
Lennard Jones Potential. The graph above plots the Lennard–Jones potential function, and indicates regions of attraction and repulsion. Atoms try to minimize their potential energy and at the lowest temperatures are sitting at the bottom of the potential curve. When the atomic separations are to the left of the minimum the atoms repel, otherwise they attract one another.

Credit: atomsinmotion.com

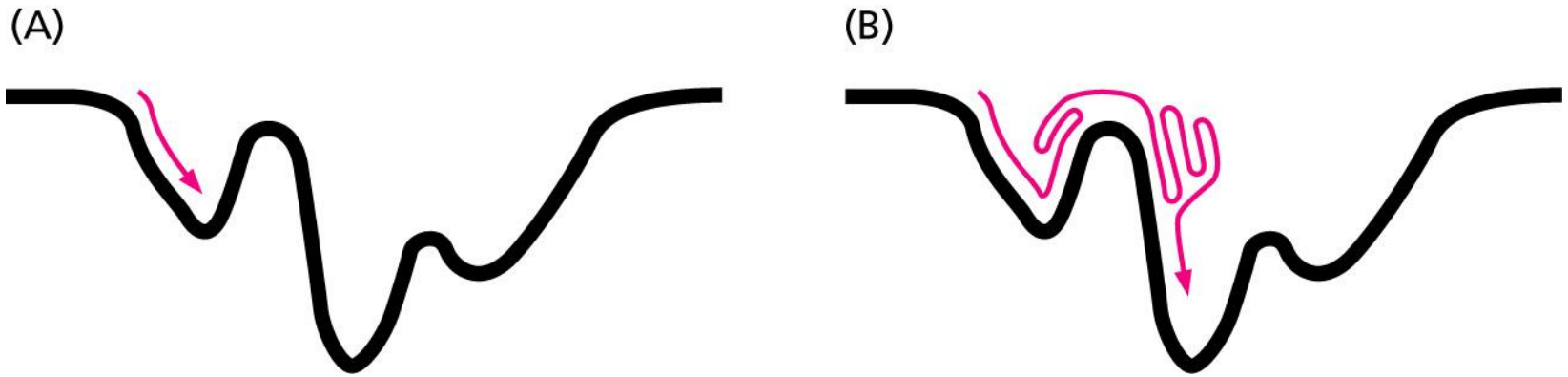
Energy Minimazation



Energy Minimazation



Credit: Iosif Vaisman @ GMU


(A)
(B)

Molecular Dynamics

$$F_i = m_i a_i$$

$$a_i = dv_i / dt$$
$$v_i = dr_i / dt$$

$$- dE / dr_i = F_i$$

$$- dE / dr_i = m_i \, d^2 r_i / dt^2$$

Credit: Iosif Vaisman @ GMU

Resources

- Homology modeling programs
  - http://www.expasy.ch/swissmod
- Threading:
  - http://www.ncbi.nlm.nih.gov/Structure/RESEARCH/threading.shtml