Exploring the energy surface

Evgenii B. Rudnyi and Jan G. Korvink
IMTEK
Albert Ludwig University
Freiburg, Germany

Preliminaries

Learning Goals

- Minimizing Energy
- Conformational Analysis
- Global Optimization
- Structure of Proteins
- Protein Folding
- Docking

References

- Leach, A.R., *Molecular modelling: principles and applications.*

On-line resources

- *Conformational Energy Searching*, cmm.info.nih.gov/modeling/guide_documents/conformation_document.html
- Catherine A. Royer, *PROTEINS*, www.biophysics.org/btol/protein.html

Minimizing Energy

$U(\boldsymbol{R})$ - Potential Energy Surface

- Quantum Chemistry - better but computationally intensive.
- Molecular Mechanics - faster but quality depends on the used empirical force field.

Potential surface contains:

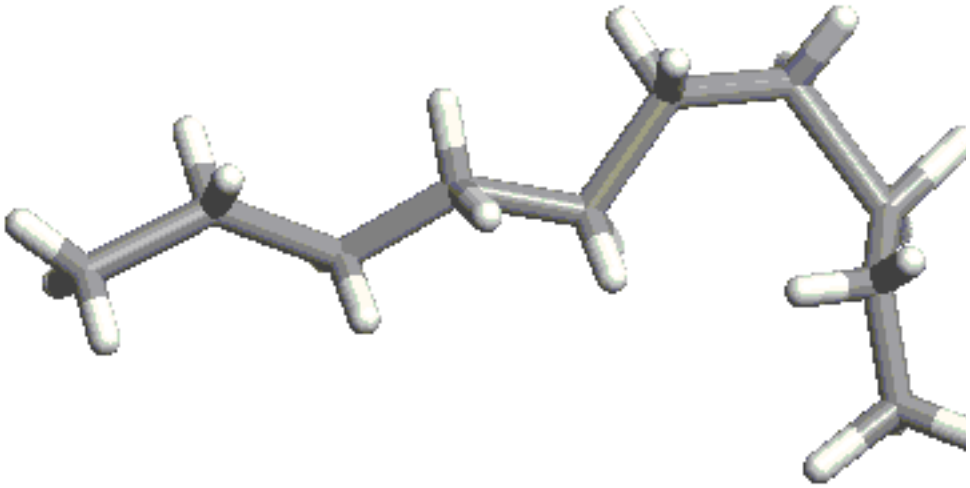
- Minima - equilibrium geometries at O K.
  - For a given temperature, one need to minimize free energy.
  - Molecular equilibrium properties may be quite different for a gas phase state and within a solution.
- Saddles points - a transition state for a chemical reaction.

Why to minimize?

- To find an initial state for molecular dynamics or Monte Carlo simulations.
- To study properties of the individual molecule.

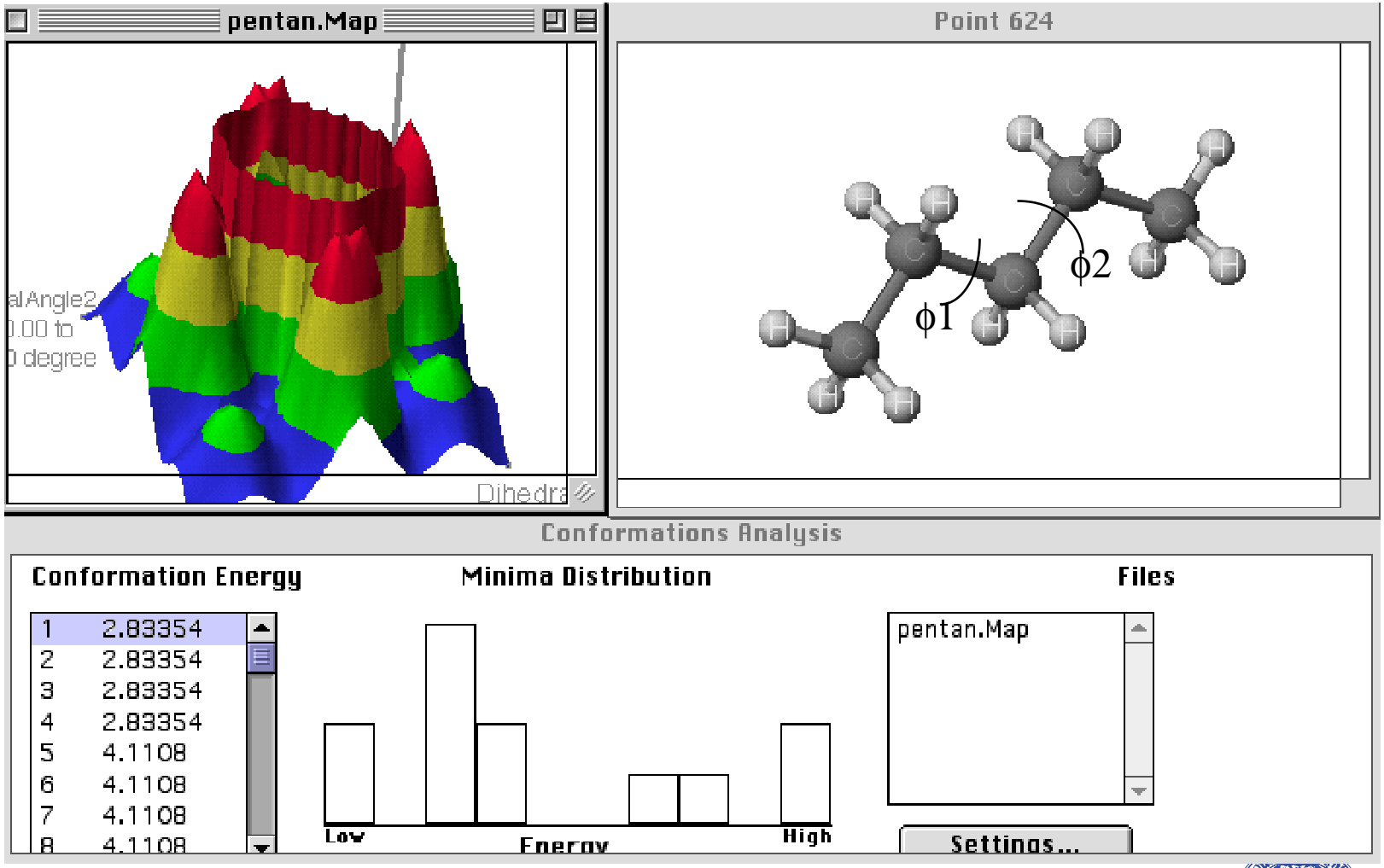
Conformational Analysis

- Demo of Conformational Search
- Systematic Methods
- Random Search Methods
- Example: $C_{17}H_{34}$
- Molecular Fitting



Conformational Analysis: Demo

- Energy map of pentane (made with Cache, www.cachesoftware.com)



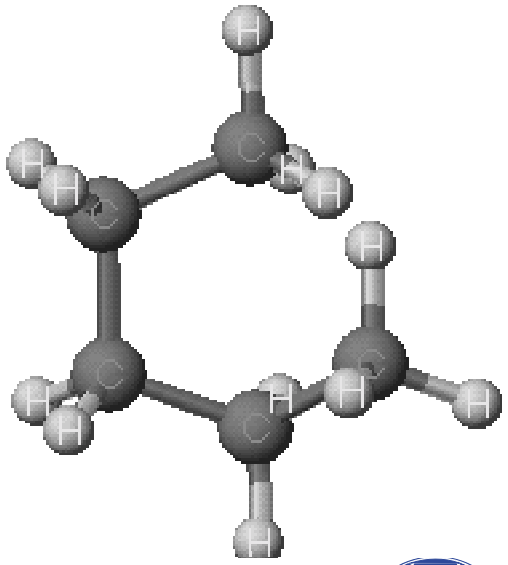
Conformational Analysis

Systematic methods

- Systematically vary dihedral angles and perform minimi-zation at any variation as an initial guess.
- Combinatorial explosion: Θ_i is the dihedral increment for bond *i*:
  - number of initial structures

$$\prod_{i=1}^{N} \frac{360}{\Theta_i},$$

  - five bonds and $\Theta_i = 30^o$ - 248832 structures.
- Could be used to problems up to 10-15 bonds if minimization is eliminated for high energy configurataion.



Conformational Analysis

Random search methods

- Choose an initial configuration randomly: then minimize.
  - Random changes to torsion angles of rotatable bonds.
  - Add a random amount to Cartesian coordinates of all atoms.
  - Distance geometry matrix:
    - distances between all pairs of atoms - N(N-1)/2 distances represented by an NxN symmetric matrix,
    - take into account that interatomic distances are interrelated: hence it is possible to estimate upper and lower bounds.
- Apply random changes to:
  - the configuration found previously,
  - a configuration taken at random from all previous.
- Simulation methods - molecular dynamics at high temperature.

Conformational Analysis

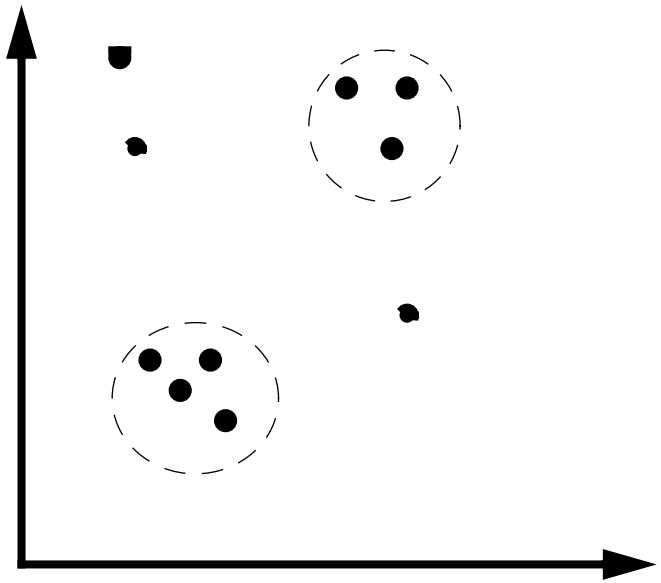
Example: $C_{17}H_{34}$

- Cycloheptadecane.
- Total unique conformers found after 30 days of processing (in 1990).
  - Within 3 kcal / mole of the global minimum.
    - Systematic search: 211.
    - Random Cartesian search: 222.
    - Random dihedral search: 249.
    - Distance geometry: 176.
    - Molecular dynamics: 169.
- Grand total 262 conformers.

Conformational Analysis

Molecular Fitting

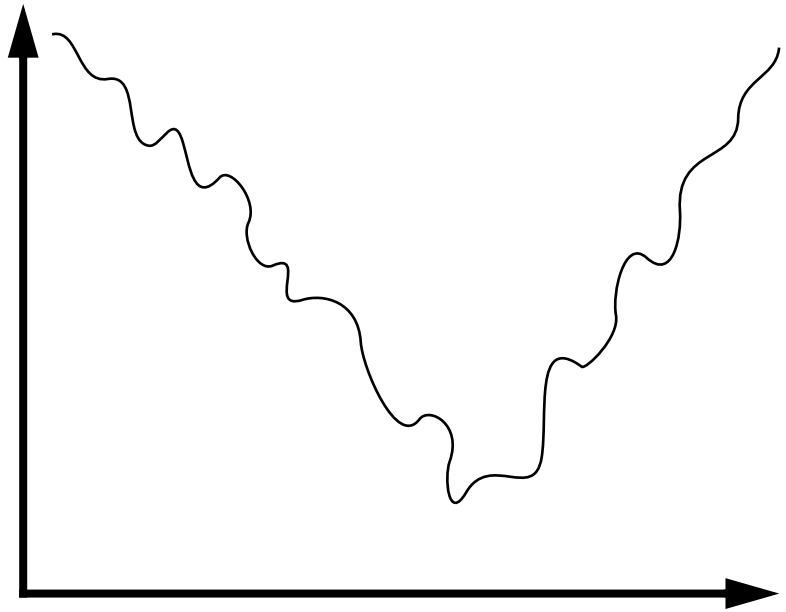
- Quantitative difference between two structures:



- RMSD, root mean square distance $\sqrt{\Sigma_i d_i^2}$.
- Clustering algorithms.
- Principal components analysis.

Global Optimization

- Contest: Alkane Global Minima
- Evolutionary Algorithms
- Simulated Annealing
- Branch and Bound

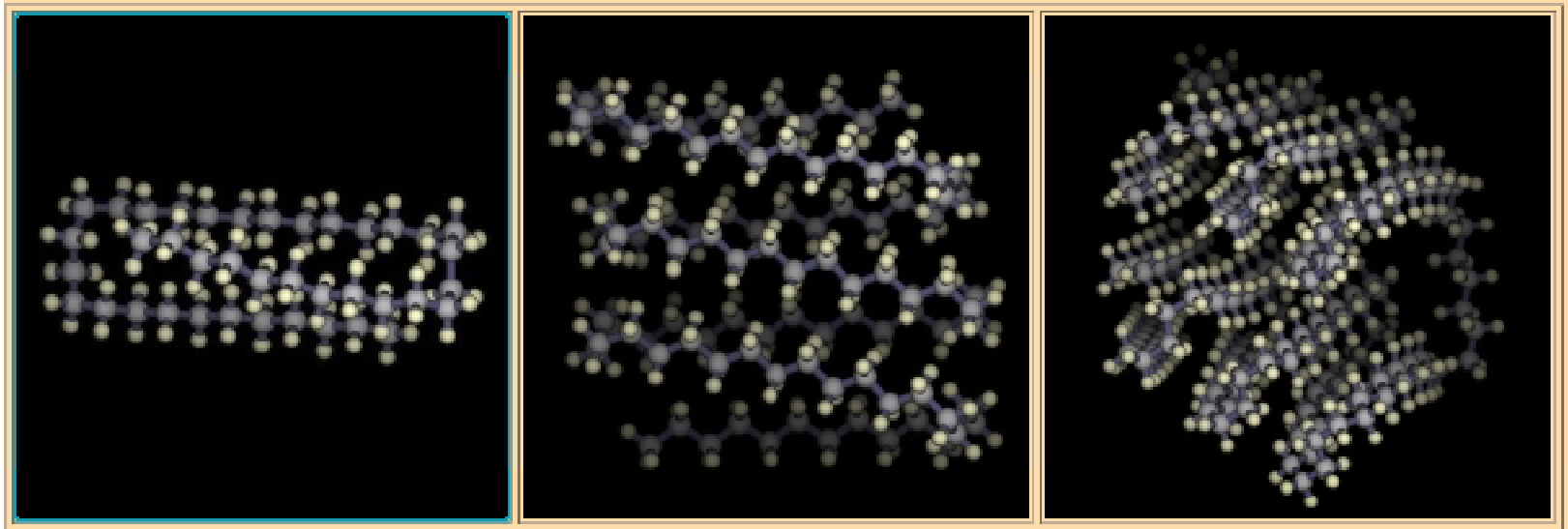


- List of Internet resources at solon.cma.univie.ac.at/~neum/glopt.html

Contest: Alkane Global Minima

- www.ch.cam.ac.uk/MMRG/alkanes/comp.html

Lowest energy structures known for unbranched alkanes



$C_{39}H_{80}$	$C_{100}H_{202}$	$C_{200}H_{402}$
58.98 kJ / mol	36.06 kJ / mol	-59.78 kJ / mol

Global Optimization

Evolutionary algorithms

- Each state is modeled by “chromosome” (linear string of bits).
- New state is generated by “crossover” and “mutation”:
  - Crossover: 00000000 and 11111111 gives 00011111 and 11100000.
  - Mutation: inverting of bit with some low probability.
- Pseudo-code

(www.cs.sandia.gov/opt/survey/ea.html):

- Initialize the population.
- Evaluate initial population.
- Repeat:
  - Perform competitive selection.
  - Apply genetic operators to generate new solutions.
  - Evaluate solutions in the population.
  - Until some convergence criteria is satisfied.

Global Optimization

Simulated annealing

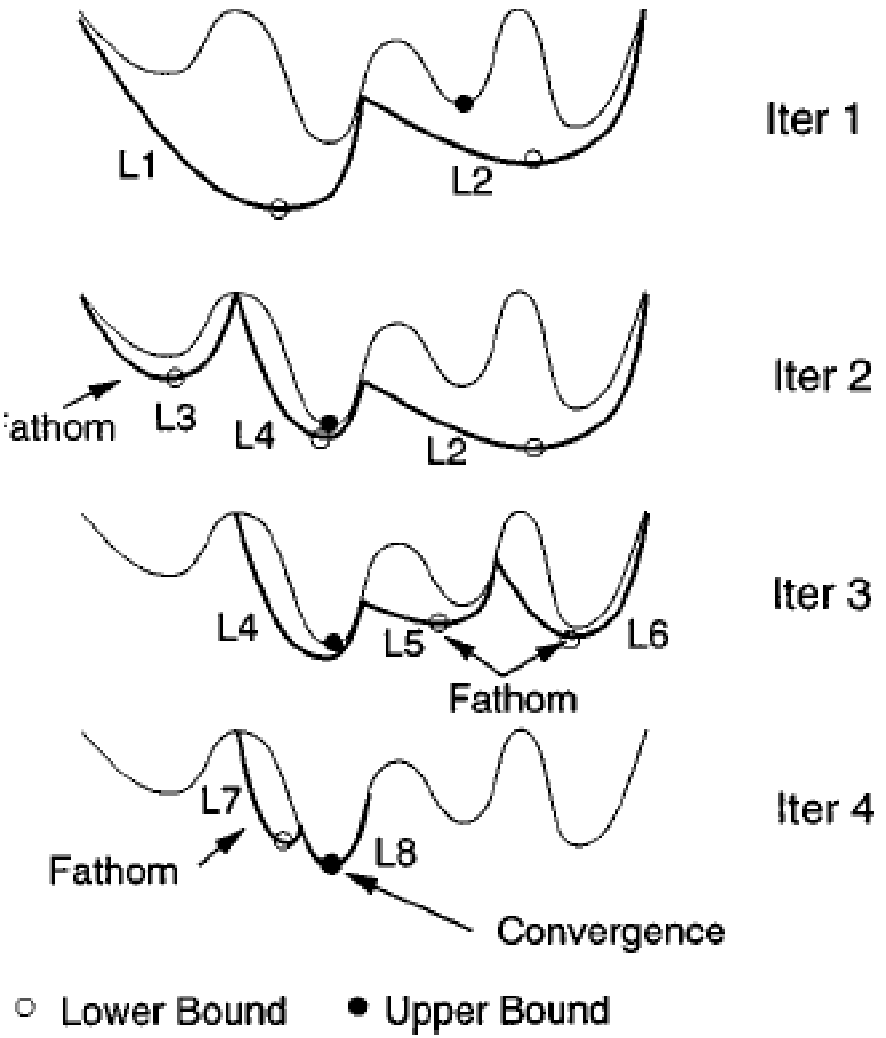
- Generalization of Monte Carlo method.
  - The goal function is an equivalent of energy.
  - Allow moves to increase energy with probability $e^{-\Delta E/T}$.
  - Start with high temperature, equilibrate, then decrease temperature, and go on.
  - Success and performance highly depends on:
- scaling energy and temperature,
- the schedule to reduce temperature,
- how moves are made.

Global Opt: Branch and Bound

- Upper bound - local minimization.
- Lower bound - by replacing f(x) with a convex function over $[x^L, x^U]$

$$L(x) = f(x) + \sum \alpha_i (x_i^L - x_i)(x_i^U - x_i)$$

- f(x) should be twice differentiable - the hessian is required to estimate α_i.
- Branch - generalized bisection (partition) of the domain.
- Example: J. L. Klepeis and C. A. Floudas, J. Chem. Phys. 1999, 110, 7491



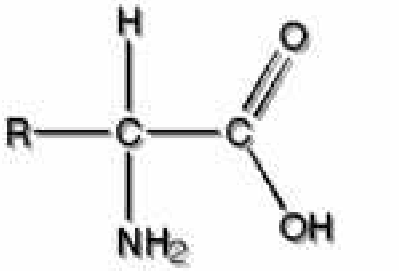
Structure of Proteins

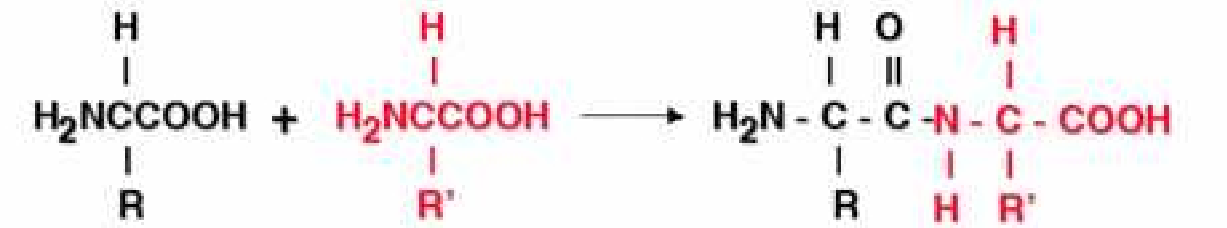
- Amino Acids, Polypeptides and Proteins
- Secondary Structure: α–helix, β–strand
- Tertiary and Quarternary Structures
- Swiss PDB Demo: enzyme lysozyme in complex with the trisaccharide inhibitor tri-(N-acetylglucosamine) or tri-NAG
  - with Deep View (Swiss PDB Viewer) www.expasy.ch/spdbv/
  - Tutorial: www.usm.maine.edu/~rhodes/SPVTut/index.html

Structure of Proteins

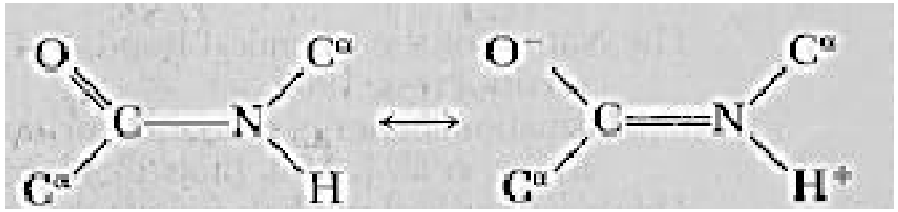
Amino Acids and Polypeptides

- 20 amino acids.
- R - side chain, hydrophobic and hydrophilic.
- Peptide - amino acid residues connected by peptide bond.
- Protein is a big polypeptide (N from 50 to 5000), the primary structure is a sequence of amino acid residues.
- Peptide bond is difficult to rotate.





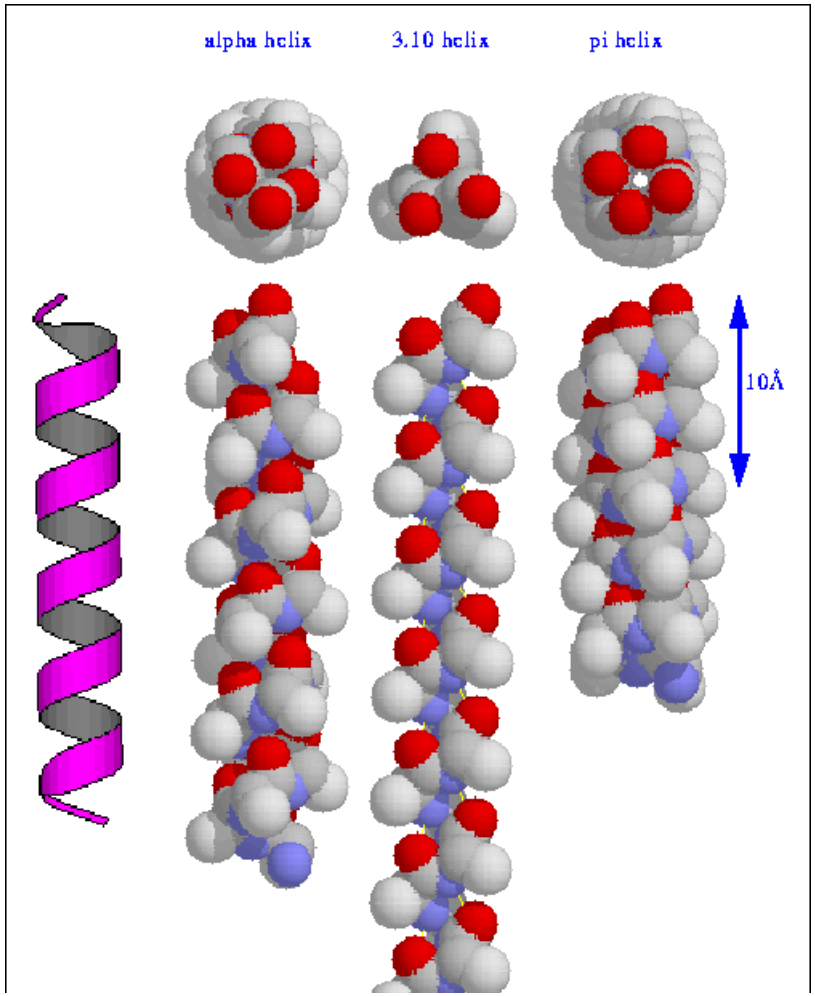
http://ull.chemistry.uakron.edu/genobc/Chapter_19/



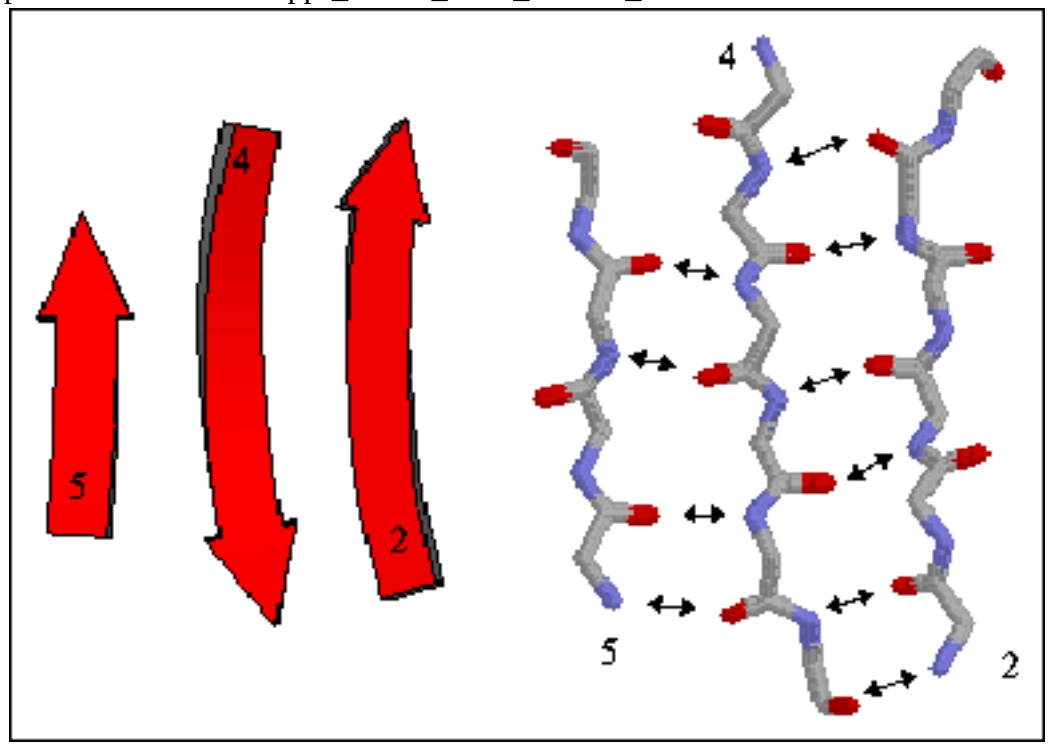
Structure of Proteins

Secondary Structure

- helix and strand (sheets).



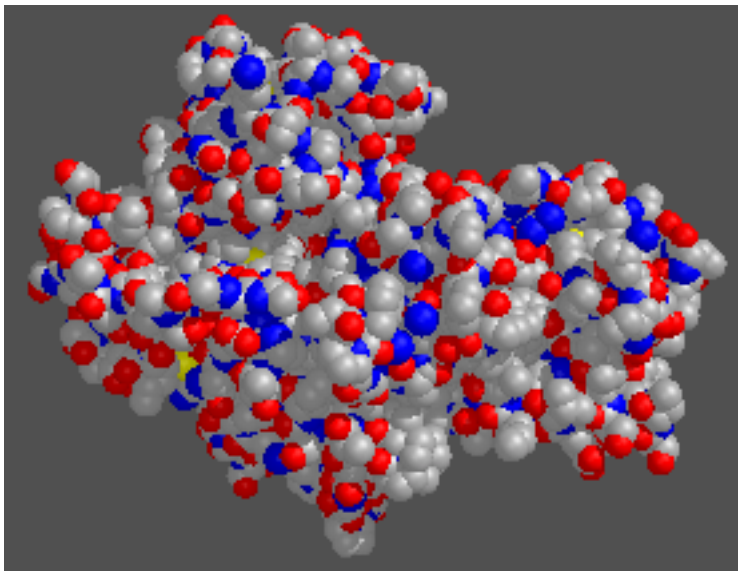
http://broccoli.mfn.ki.se/pps_course_96/ss_960723_1.html



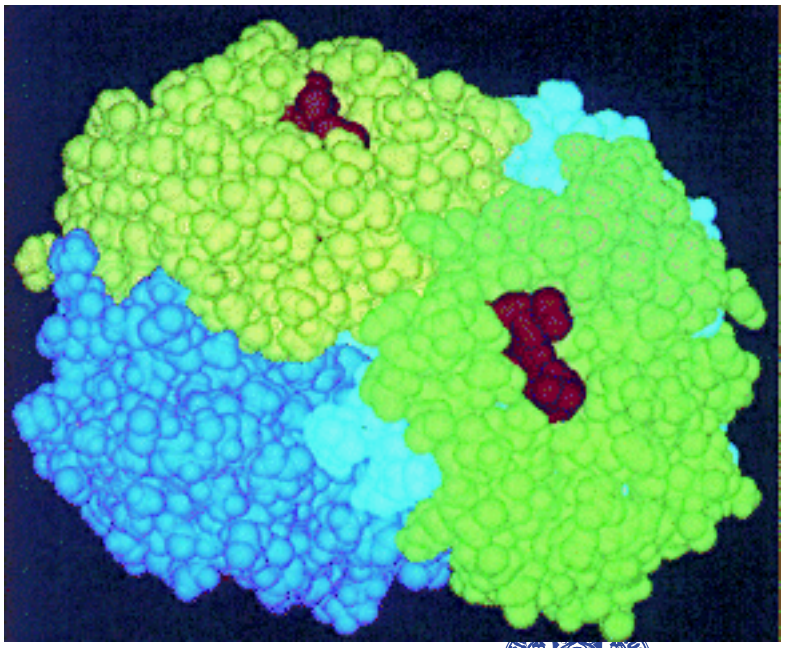
Structure of Proteins

Tertiary and Quarternary Structure

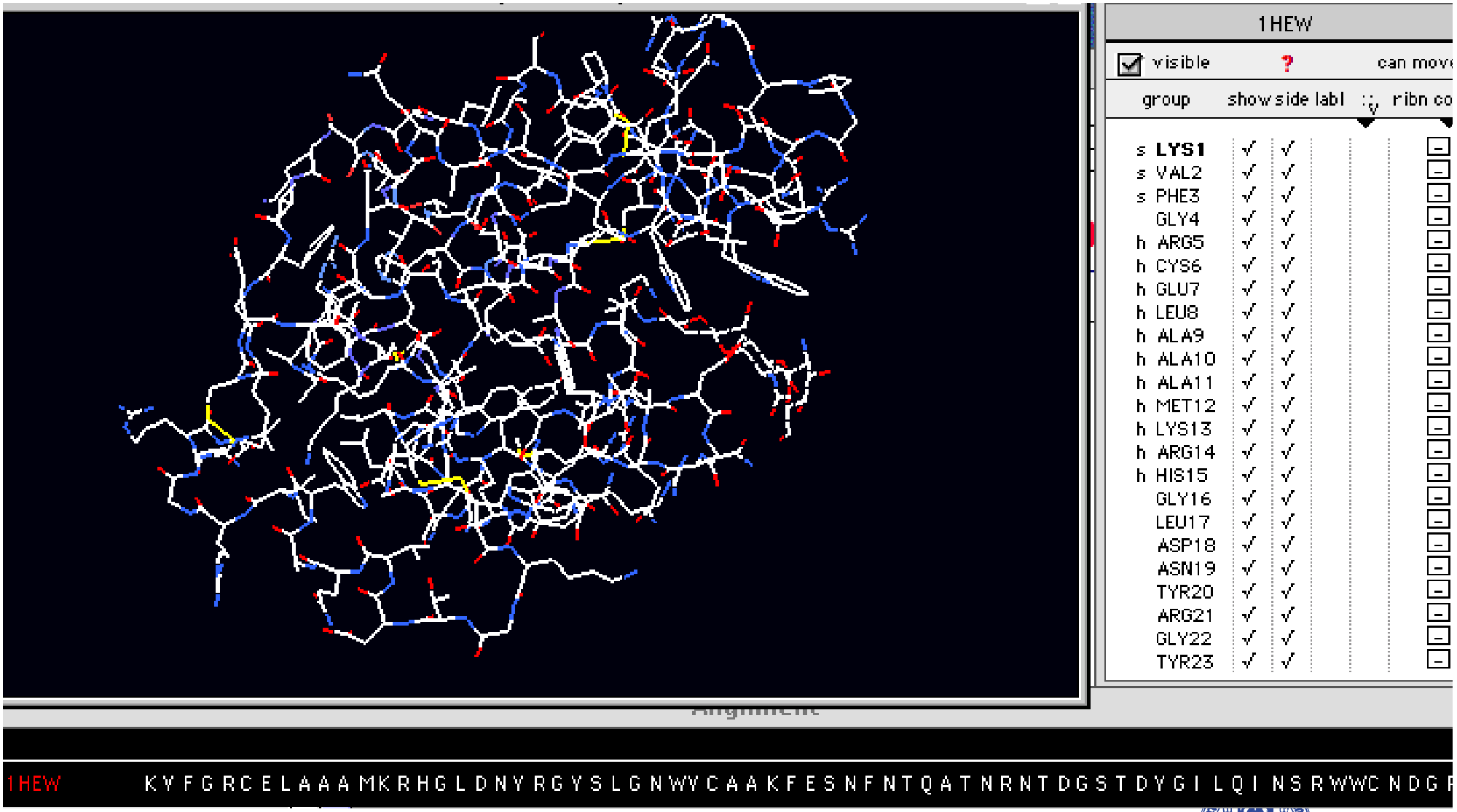
- Tertiary structure (globule) - example: hexokinase.
- Quarternary structure (several polypeptide chains) example: hemoglobin, a protein with four polypeptides-- two alpha-globins, and two beta-globins.



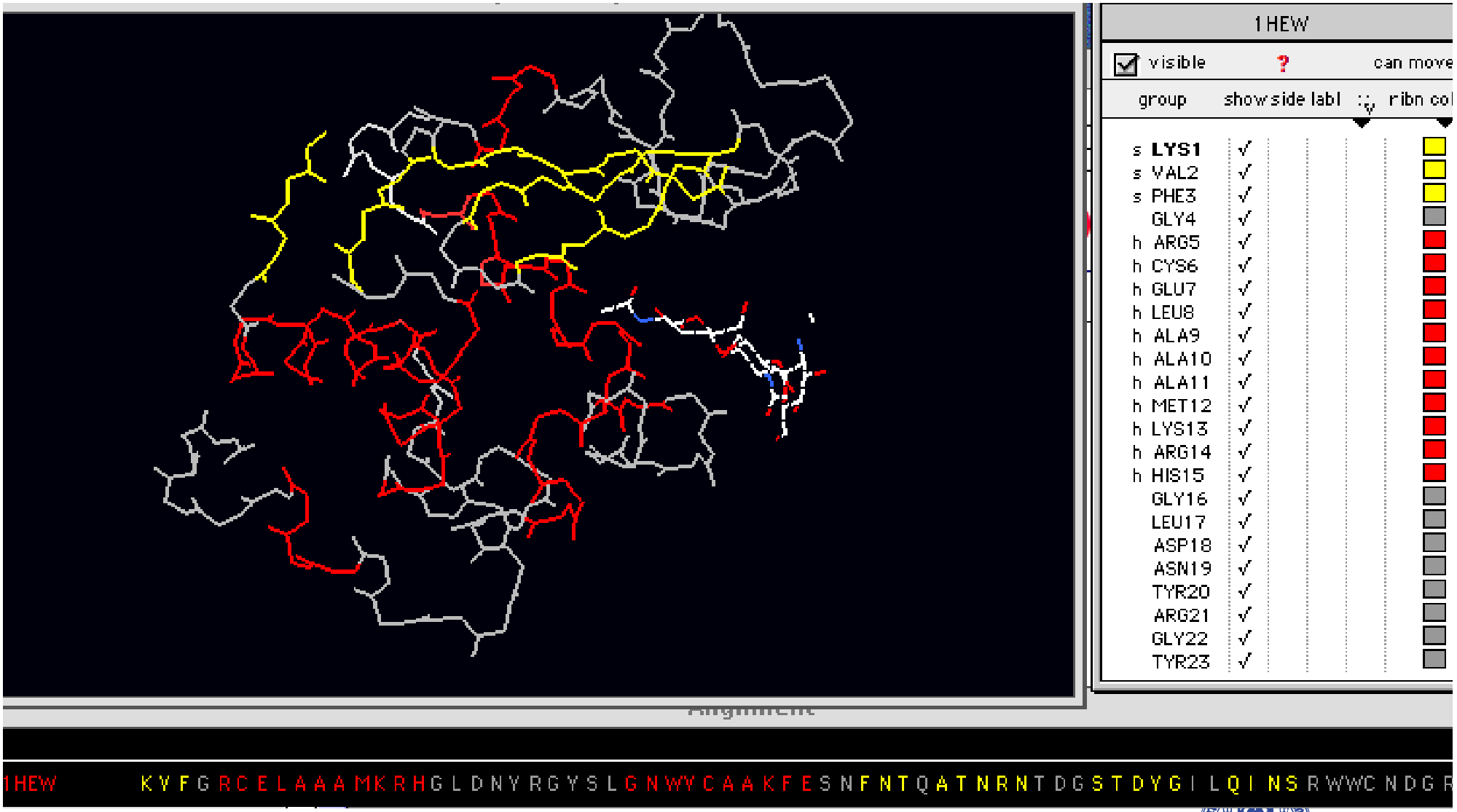
http://esg-www.mit.edu:8001/esgbio/lm/proteins/structure/structure.html



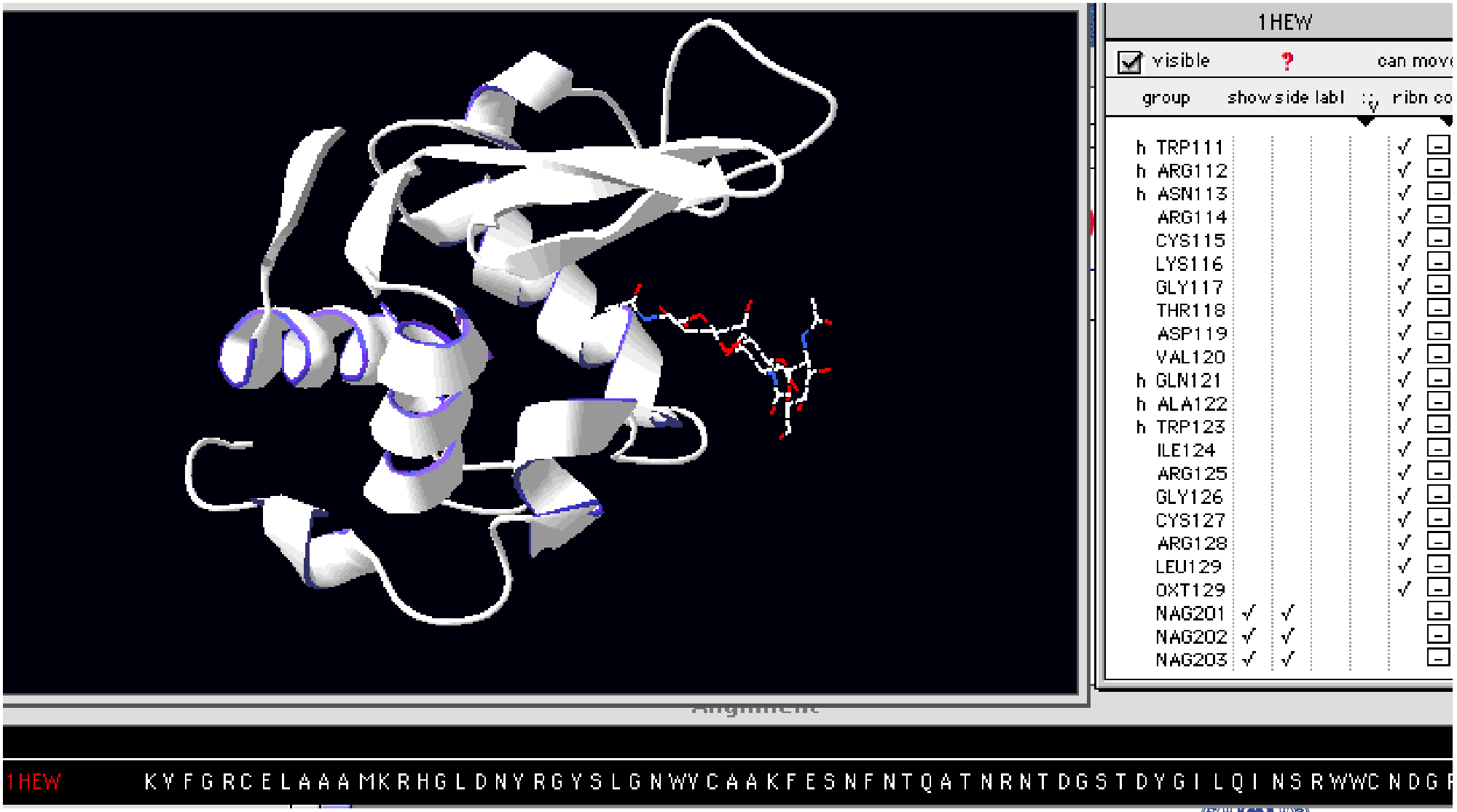
Swiss PDB Demo 1



Swiss PDB Demo 2

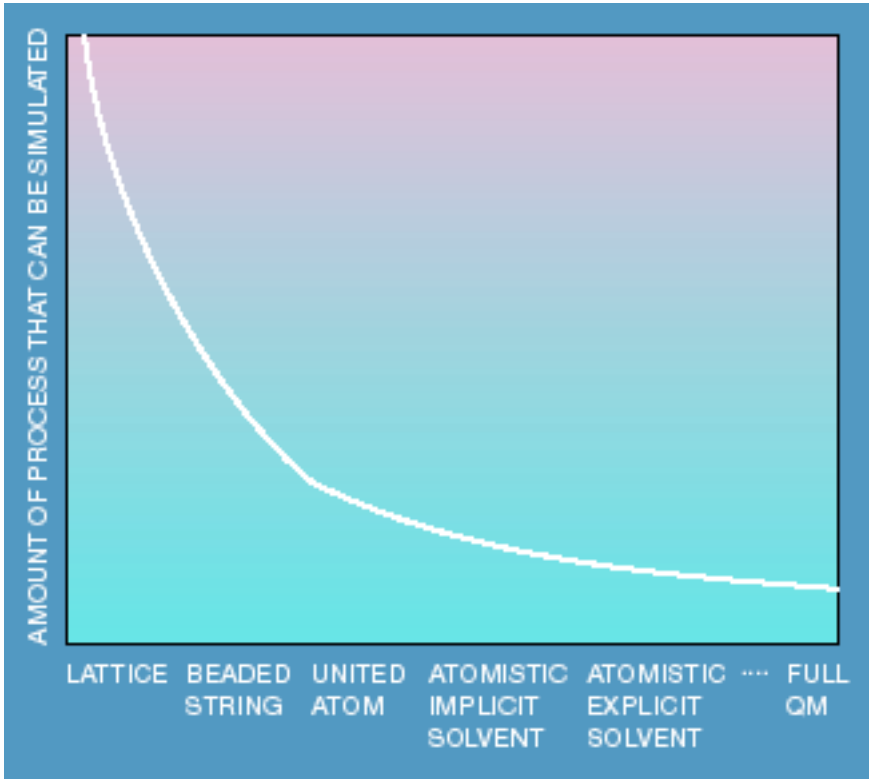


Swiss PDB Demo 3



Protein Folding

- Levinthal and Anfinsen
- Current View
- Lattice Models - a Conceptual Solution
- Brute Force Minimization
- Brute Force Molecular Dynamics
- DNA
- Sequencing and Bioinformatics
- Restrained Molecular Dynamics - a Tool to Clear the Protein Structure



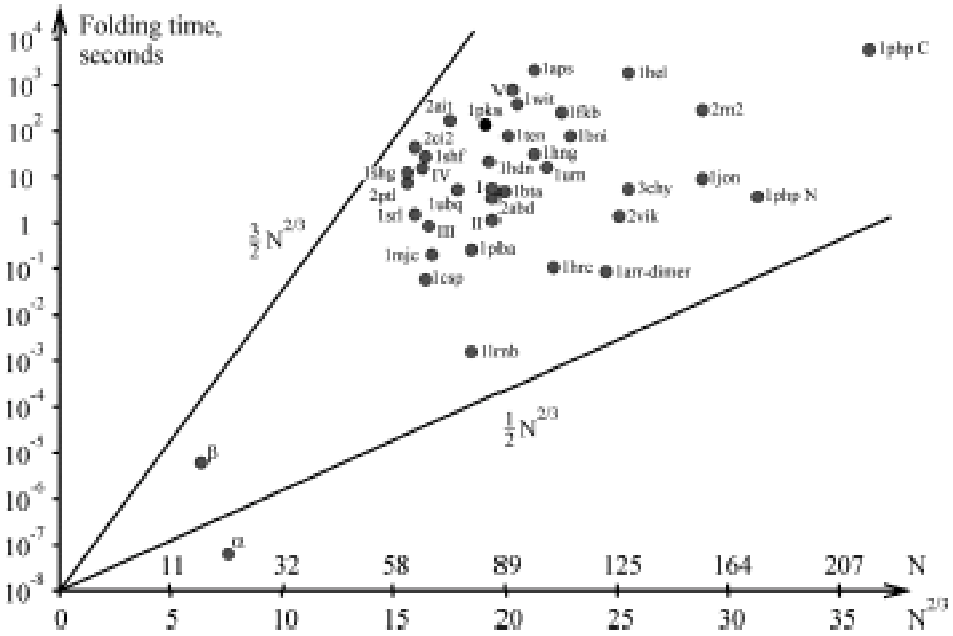
Blue Gene. IBM Systems Journal, v. 40, N. 2, 2001, p. 310.

Protein Folding

The Problem

- Anfinsen, experimenter, 1961, unfolded and folded enzyme ribonuclease A *(in vitro)*.
- Levinthal, theoretician, 1968, protein folding is impossible:
  - 100 amino acids chain has 10^{48} conformations. If a change between two conformations would take 10^{-11} sec, than it takes 10^{29} years to explore all the conformations.

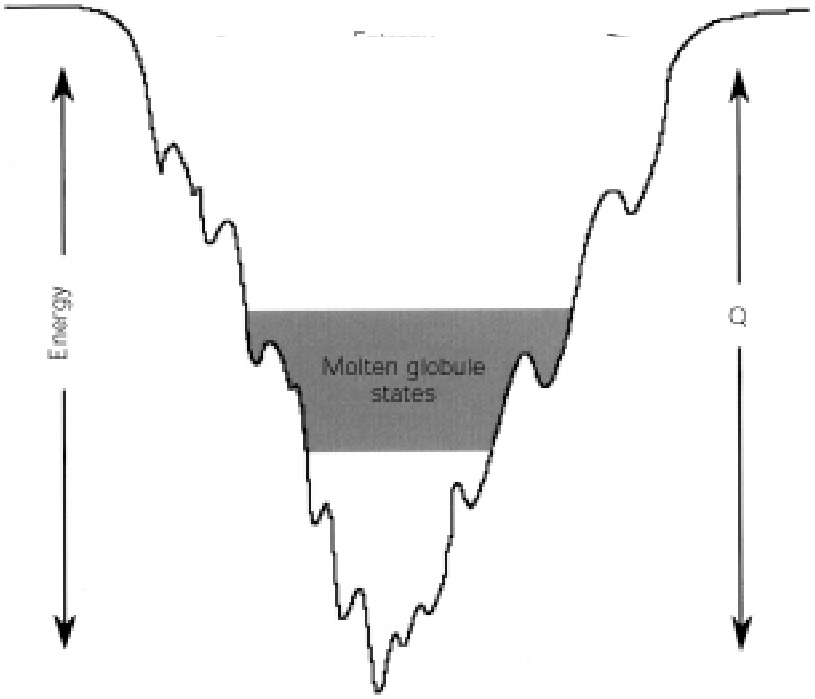
Galzitskaya, FEBS Letters, 2001, 489, 113

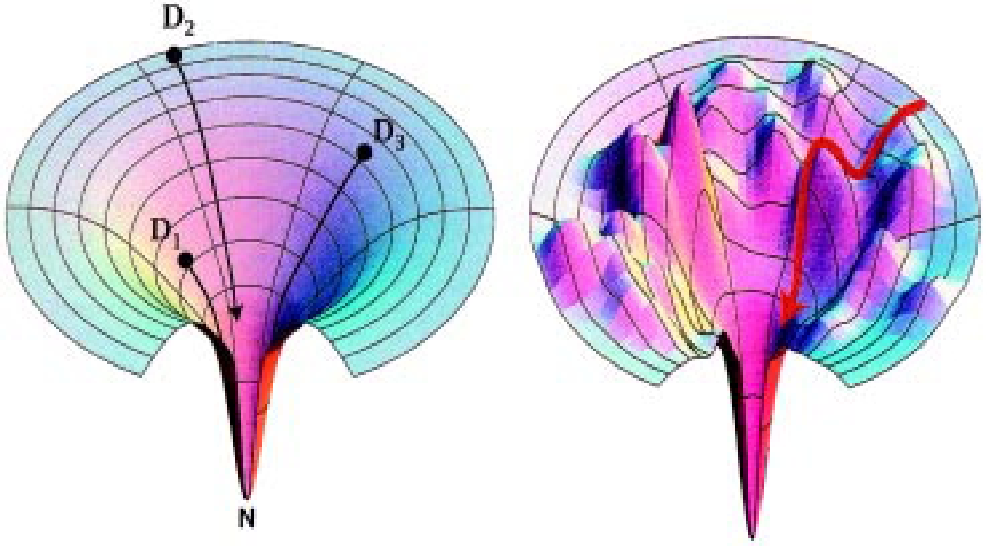


Protein Folding

Current View

- From a lot of experimental and computational works.
- 1987, discovery of chaperons - helpers to fold some proteins *in vivo*.
- Misfolding of proteins leads to some diseases.





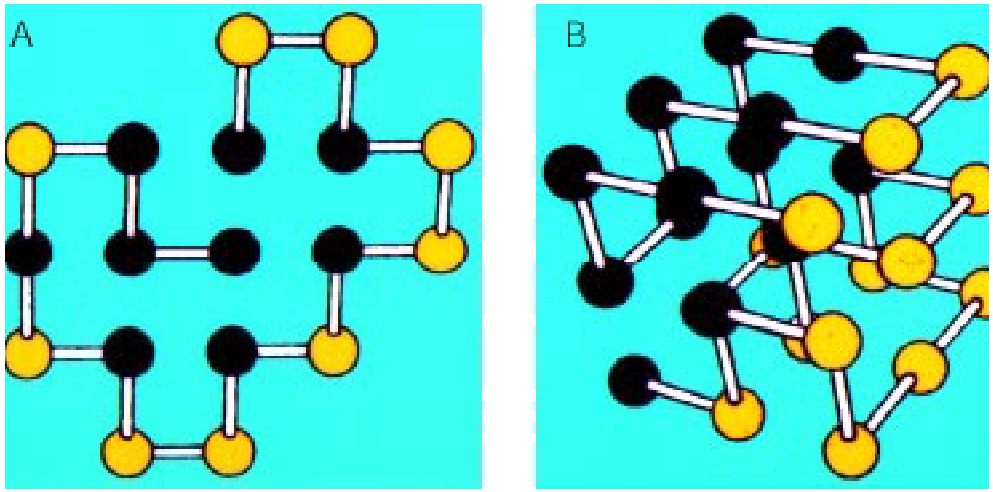
from Yon, Brazilian J. Med. Biol. Res., 2001, 34, 419

Protein Folding

Lattice models

- Protein is modeled by string of beads, hydrophobic and hydrophilic.
- The interaction between beads provide the energy function for the Monte Carlo simulation,
- Typical size 3x3x3, 27 residue.
- 200 randomly generated sequences.
- 20 found their native state very easily.
- 146 never found the native state.

on, Brazilian J. Med. Biol. Res., 2001, 34, 419



Brute Force Minimization

- Typical test problem is met-enkephalin: 5 residue benchmark - up to 24 unknowns, about 10^{11} local minima.
- J. L. Klepeis and C. A. Floudas, *Free energy calculations for peptides via deterministic global optimization*. J. Chem. Phys. 1999, v. 110, p. 7491.

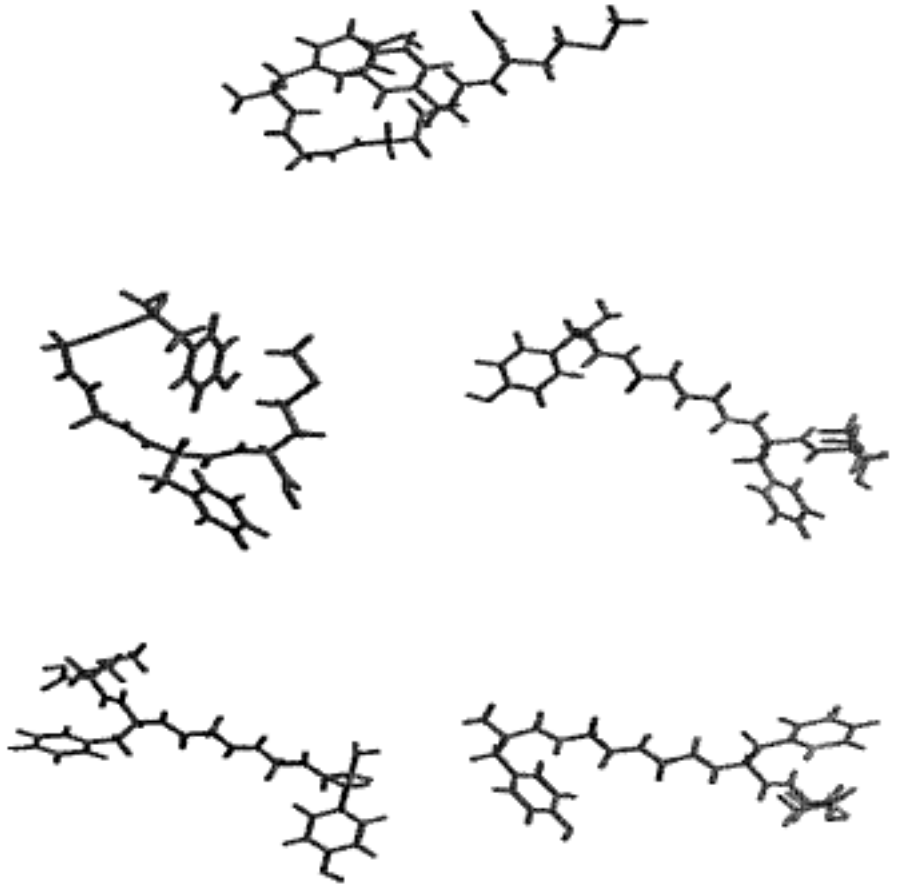


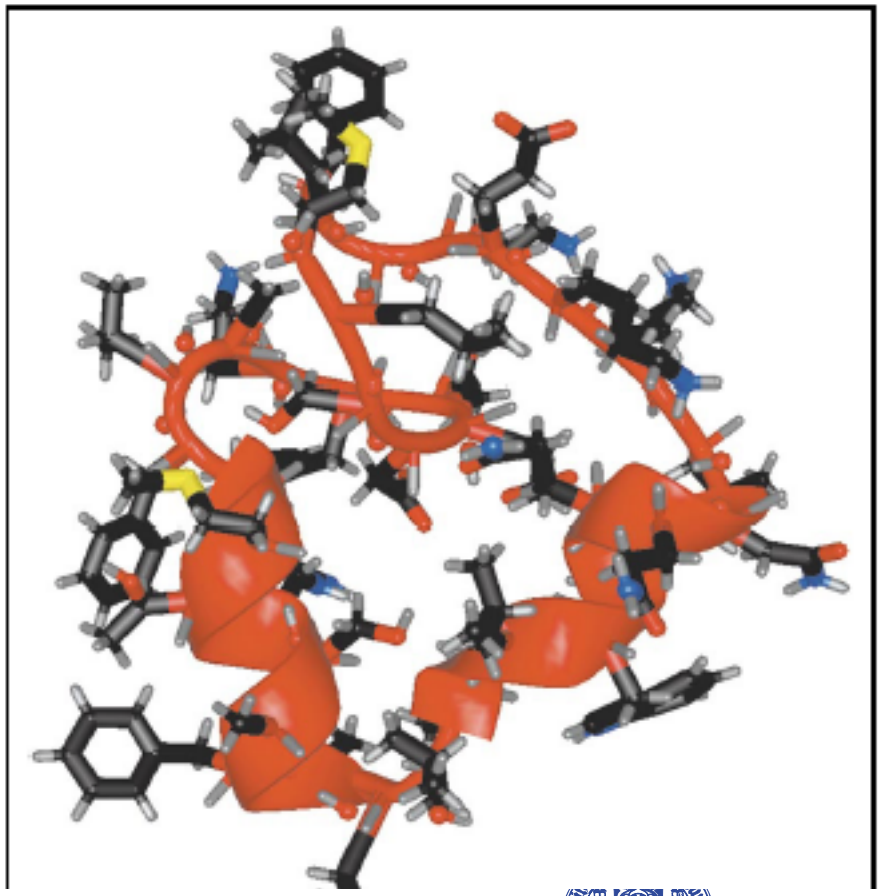
FIG. 11. FEGM structures for solvated met-enkephalin. The top figure is the PEGM and the FEGM for 100 K. The structures at other temperatures (200, 300, 400, 500) are shown left to right, top to bottom.

Protein Folding

Brute Force MD

- Duan&Kollman, 1998, IBM Systems Journal, v. 40, N 2, 2001, p. 297.
- 1 μs simulation of 36 residue peptide from unfolded state (time of folding 10-100 μs).
- protein + 3000 water molecules with time step of 2 fs.
- 4 months on 256 processor parallel supercomputer.
- *Blue Gene: A vision for protein science using a petaflop supercomputer*, www.research.ibm.com/journal/sj/402/allen.html

Figure 2 The structure of an intermediate state (at 350 nanoseconds) in one of the 500-nanosecond simulations



- T. Schlick, Computing in Science & Engineering, 2000, v. 6, N 2, p. 38.

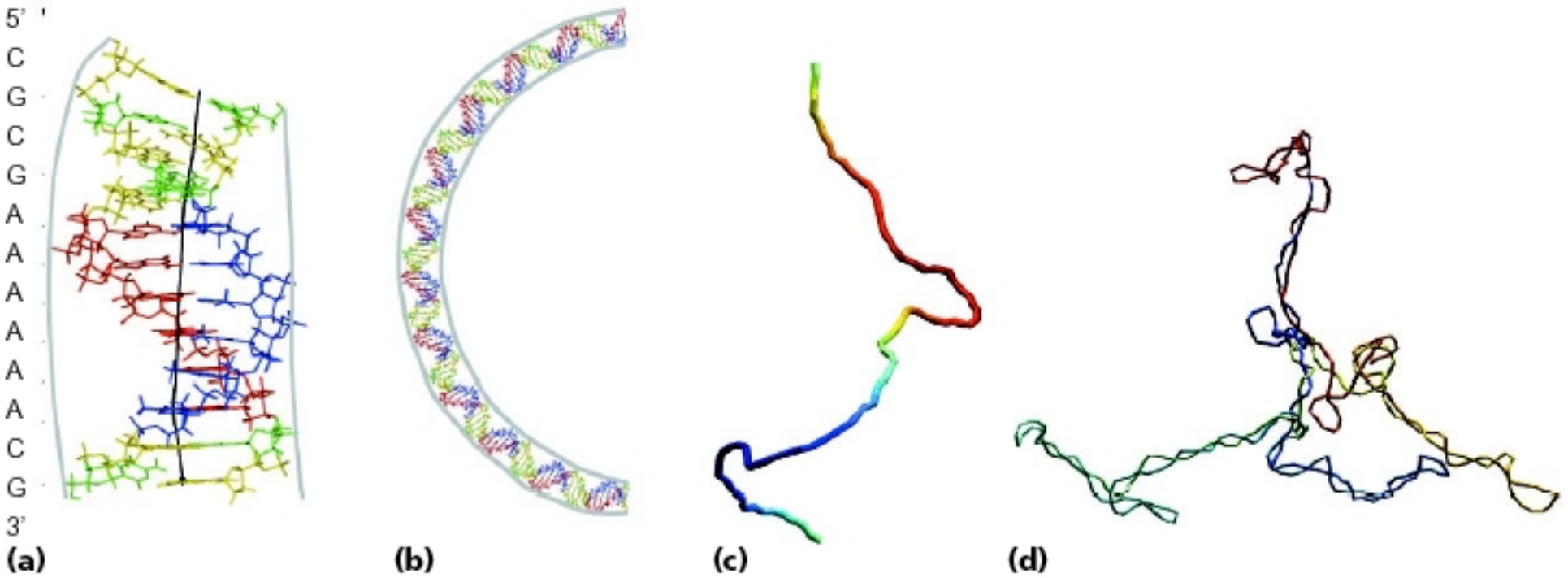


Figure 2. Models of DNA at four different length scales: (a) an A-tract dodecamer with an overall curvature of 11°, (b) a model of 120 base pairs of a phased A-tract sequence, (c) linear DNA of 1.2 kbp, and (d) supercoiled DNA of 12 kbp. Our computed dodecamer by all-atom molecular dynamics served as the model for constructing the 120-base-pair system; the larger linear and supercoiled structures are representative of the thermal equilibrium ensemble, as generated by Brownian dynamics simulations. The curve for the long DNA represents the double helix.

Protein Folding

Sequencing and Bioinformatics

- Knowledge based methods (protein databank - ca. 15000 structures).
- Proteins with similar primary structures (sequences) tend to have similar three dimensional structure.
- Sequence alignment - based on some scoring.
- Happens to be a hard computational problem.
- Point accepted mutation matrix (a probability to change one amino acid to another).
- Dynamic programming (another global minimum search method).
- Heuristic search: BLAST (basic local alignment search tool).
- Competition CASP: predictioncenter.llnl.gov/casp4/Casp4.html

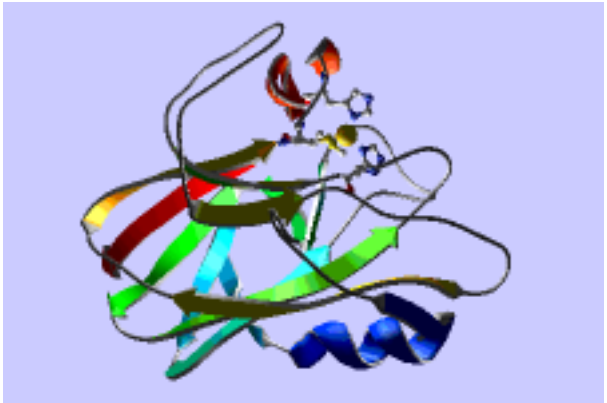
Protein Folding

Restrained Molecular Dynamics

- Experiments to determine molecular structure:
  - X-ray crystallography,
  - NMR - Nuclear Magnetic Resonance.
- $SS = \Sigma_i (y_i^{ex} - y_i^{calc})^2$, by itself quite a hard problem.
- It is impossible to obtain all atomic coordinates.
- RMD - a molecular dynamics combined with simulated annealing when the potential energy is modified to include experiments

$$E_{tot} = V(r) + a \cdot SS$$

- a - scaling factor.



http://www.usm.maine.edu/~rhodes/

Docking

- To predict the structure of the intermolecular complex formed between two or more molecules.
- Depends on the free energy and the solvent.

Summary

- Conformational analysis
- Global optimization
- Structure of proteins
- Protein folding
- Docking