THE PROTEIN FOLDING PROBLEM

Alexey Onufriev,
Virginia Tech

www.cs.vt.edu/~onufriev

OUTLINE:

- Protein biochemistry 101

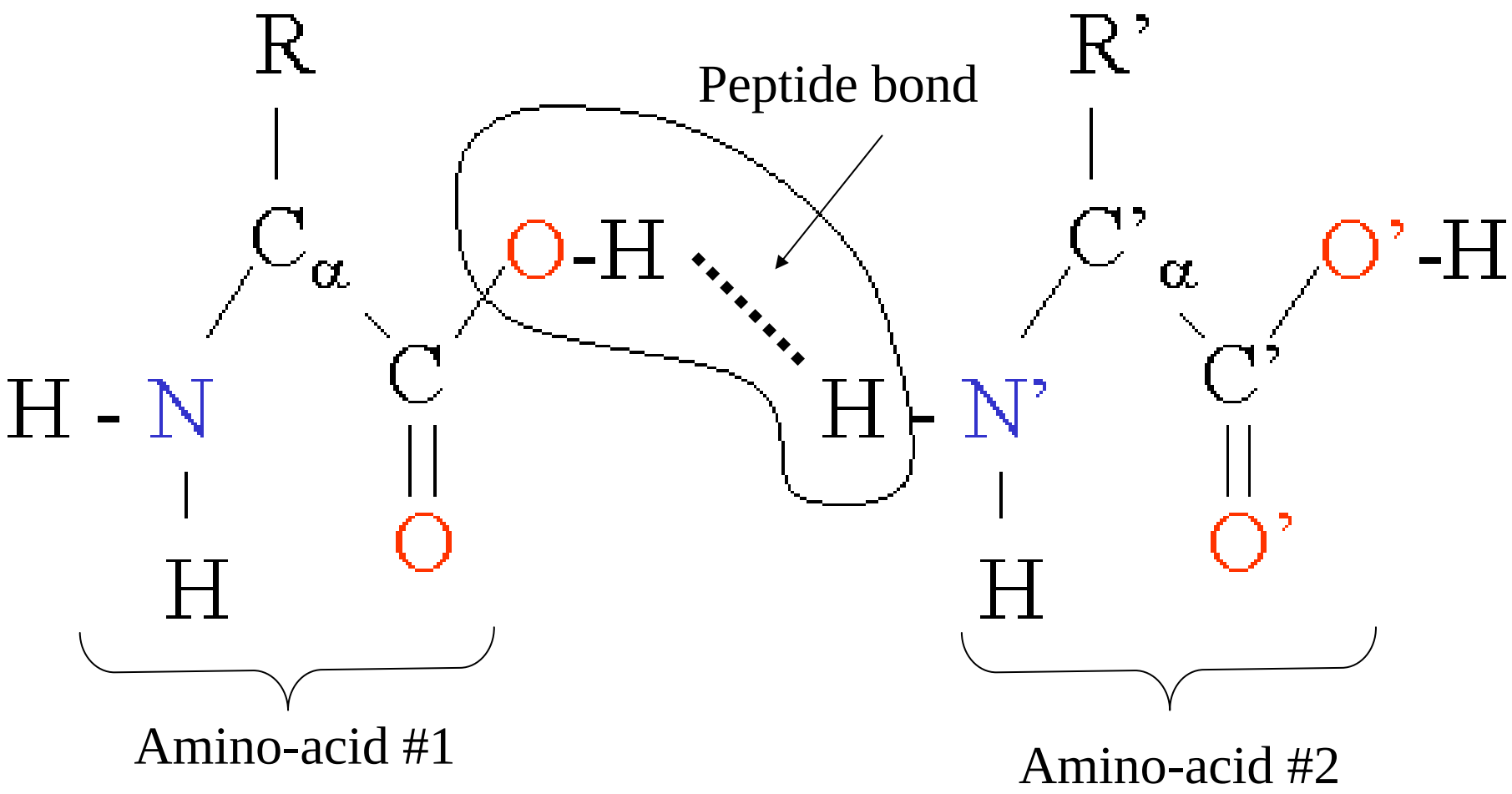
- Overview of available methods.
  A) Experiment
  B) Theory.
  Their pros and cons.

- Folding proteins in ``virtual water”. Energy landscapes.

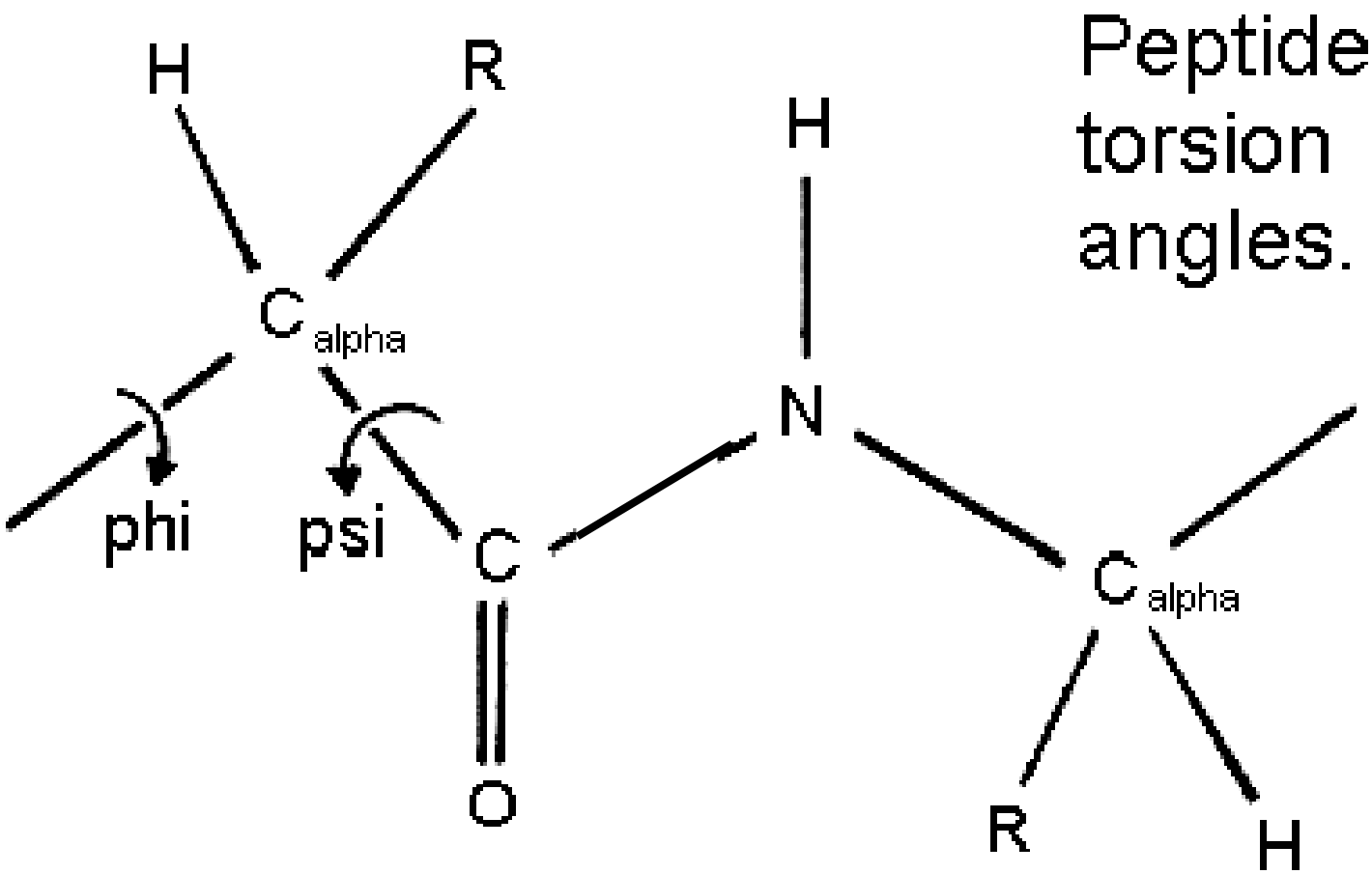
Claims: 1. The use of detailed, atomic resolution models is often critical.
2. Simple physics works on very complex biological systems.

Protein Structure in 3 steps.

Step 1. Two amino-acids together (di-peptide)



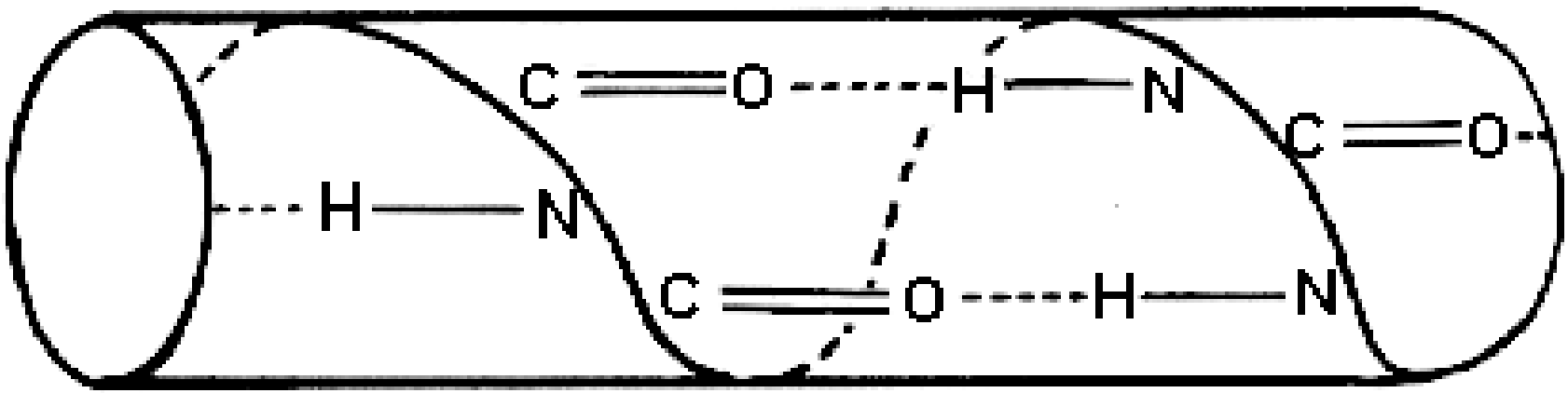
Step 2: Most flexible degrees of freedom:



Protein Structure in 3 steps.

Sometimes, polypeptide chain forms helical structure:

Toilet roll representation of the main chain hydrogen bonding in an alpha-helix.

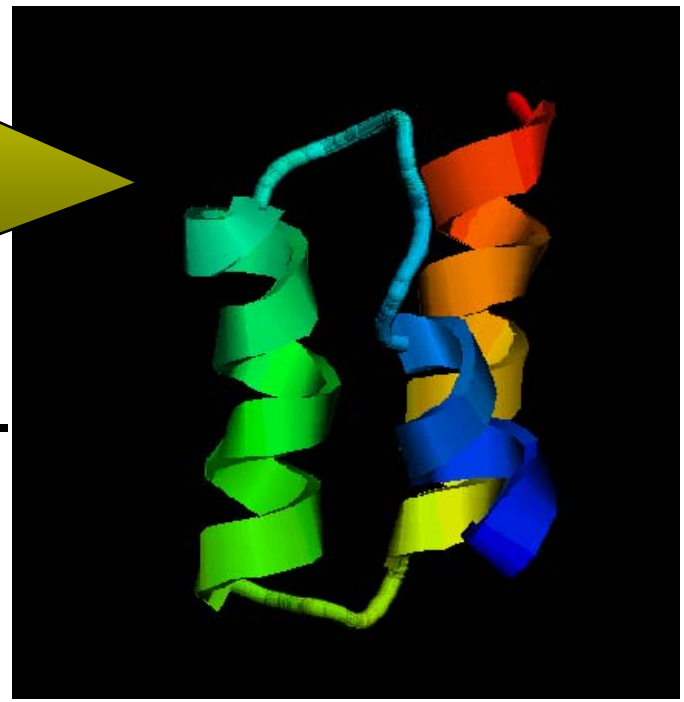


Protein folding problem #1

Amino-acid sequence – translated genetic code.

MET—ALA—ALA—ASP—GLU—GLU--….

How?



Experiment: amino acid sequence uniquely determines protein's 3D shape (ground state).

Nature does it all the time. Can we?

Experimental methods:

- *X-ray* // "Gold standard". Atomic resolution.

// Crystal packing artefacts, not all proteins can be crystallized well, problems with large ones, membrane proteins, missing hydrogens, and, sometimes, big chunks of structure

- *NMR* // "true" structure in solution. Can get hydrogens. Can trace some dynamics (e.g. in folding).

// expensive, slow. Large errors -> low reolution in many cases. Can't get all atoms. No large structures.

- *Neutron Scattering* // perfect for hydrogens. Dynamics. // proteins in powder state, very expensive. Only very few structures.
- *Crio-EM* // very large structures (viruses). // low (10A) resolution.

Theoretical Approaches.

Heuristic

•*(homology modeling).*

Steps:

- template recognition
- backbone generation (threading)
- Loop modeling
- side-chain modeling
- Optimization + Validation

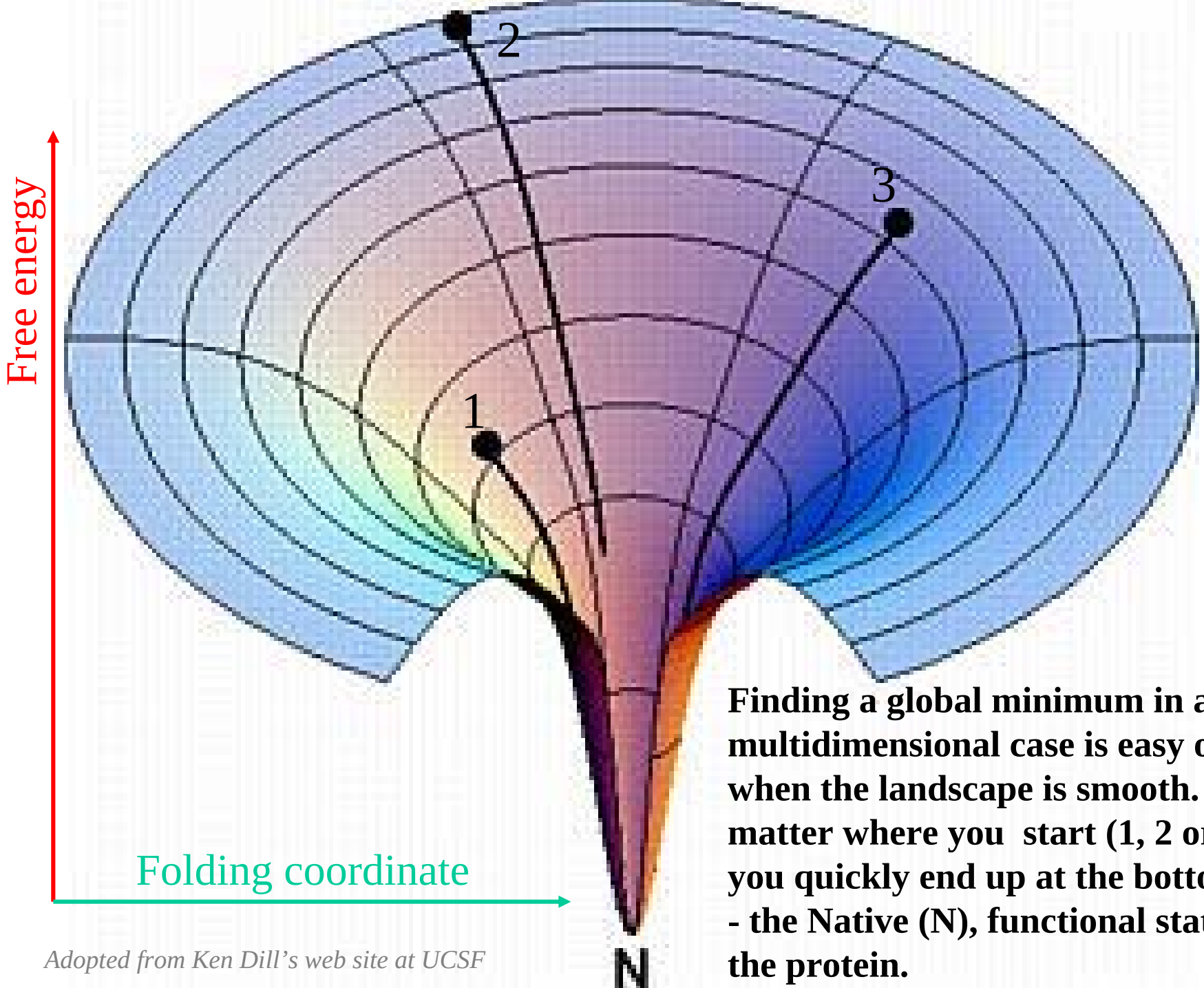
Ab-initio
(just use the right Physics and it will fold… Really?)

Homology Modeling

- Fast enough for approximate prediction of folds of fractions of whole genomes.
- For small proteins (< 90 residues), predicts structure to within 2-6 Angtroms error (compared to experiment)
- Drawbacks: 1) no template: no go.
  2) no atomic resolution
  3) hard to use to learn about the folding process.

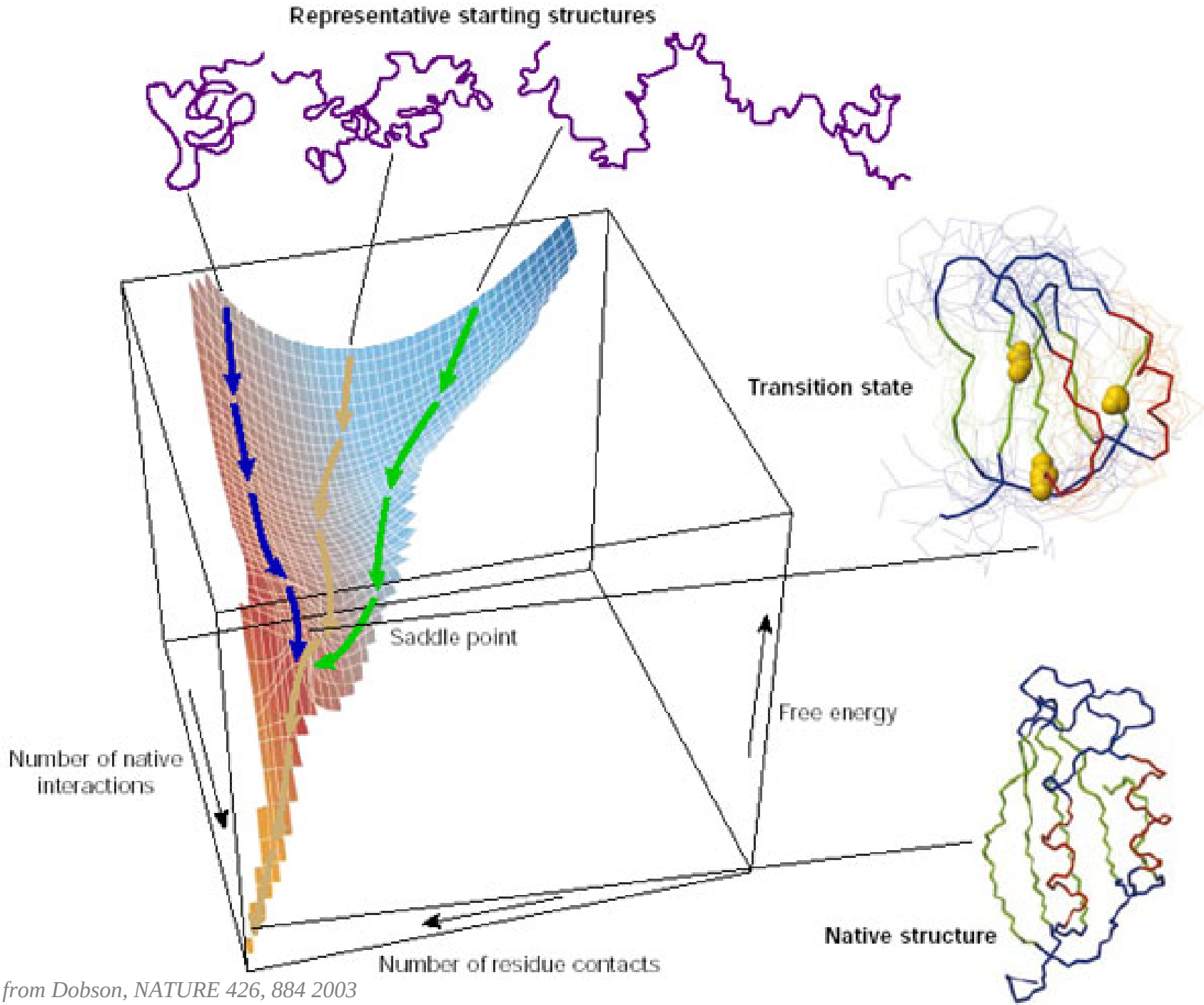
Protein folding problem = Minimization problem.

Objective function = System's energy.

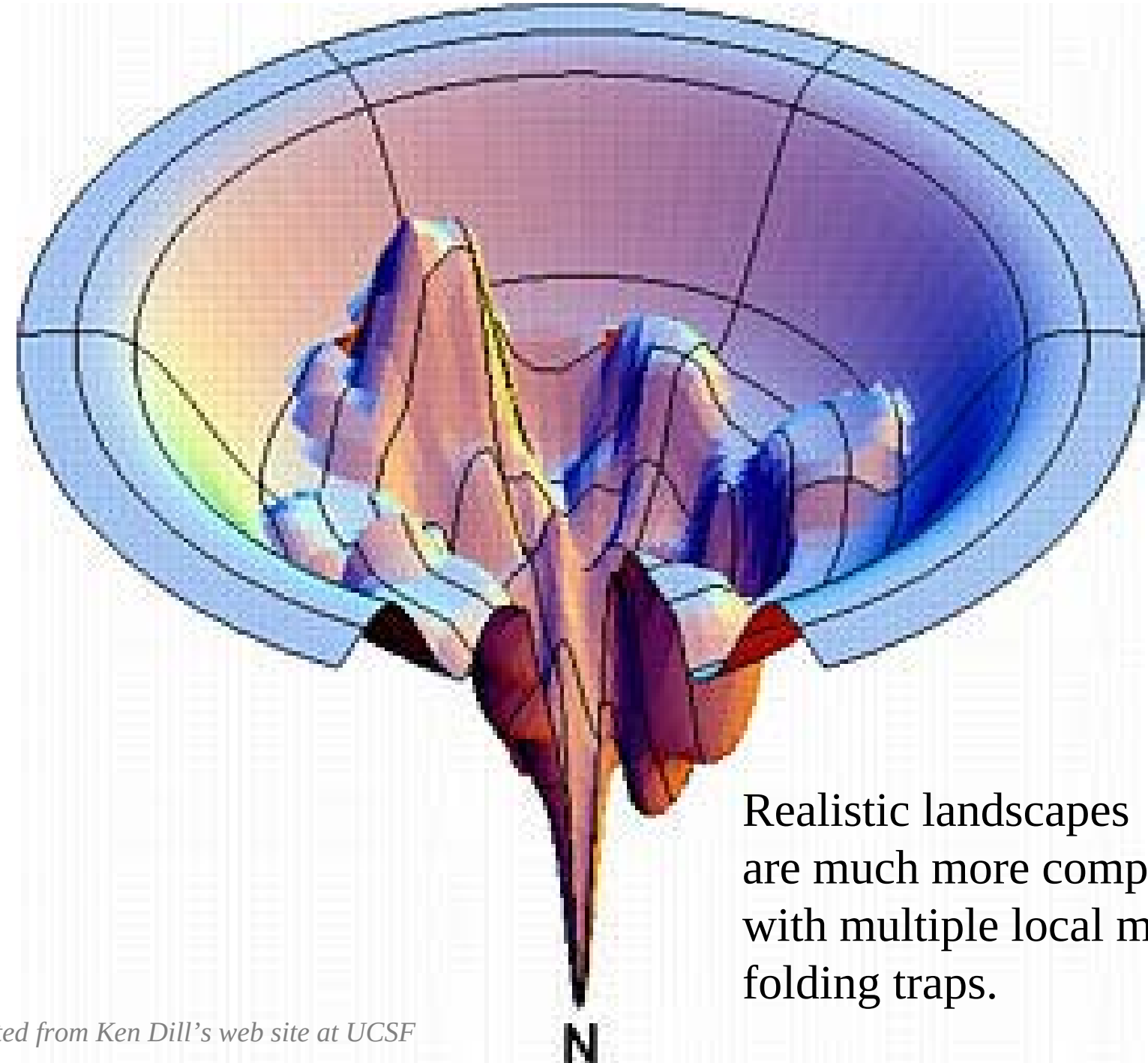


Finding a global minimum in a multidimensional case is easy only when the landscape is smooth. No matter where you start (1, 2 or 3), you quickly end up at the bottom -- the Native (N), functional state of the protein.

Adopted from Ken Dill's web site at UCSF

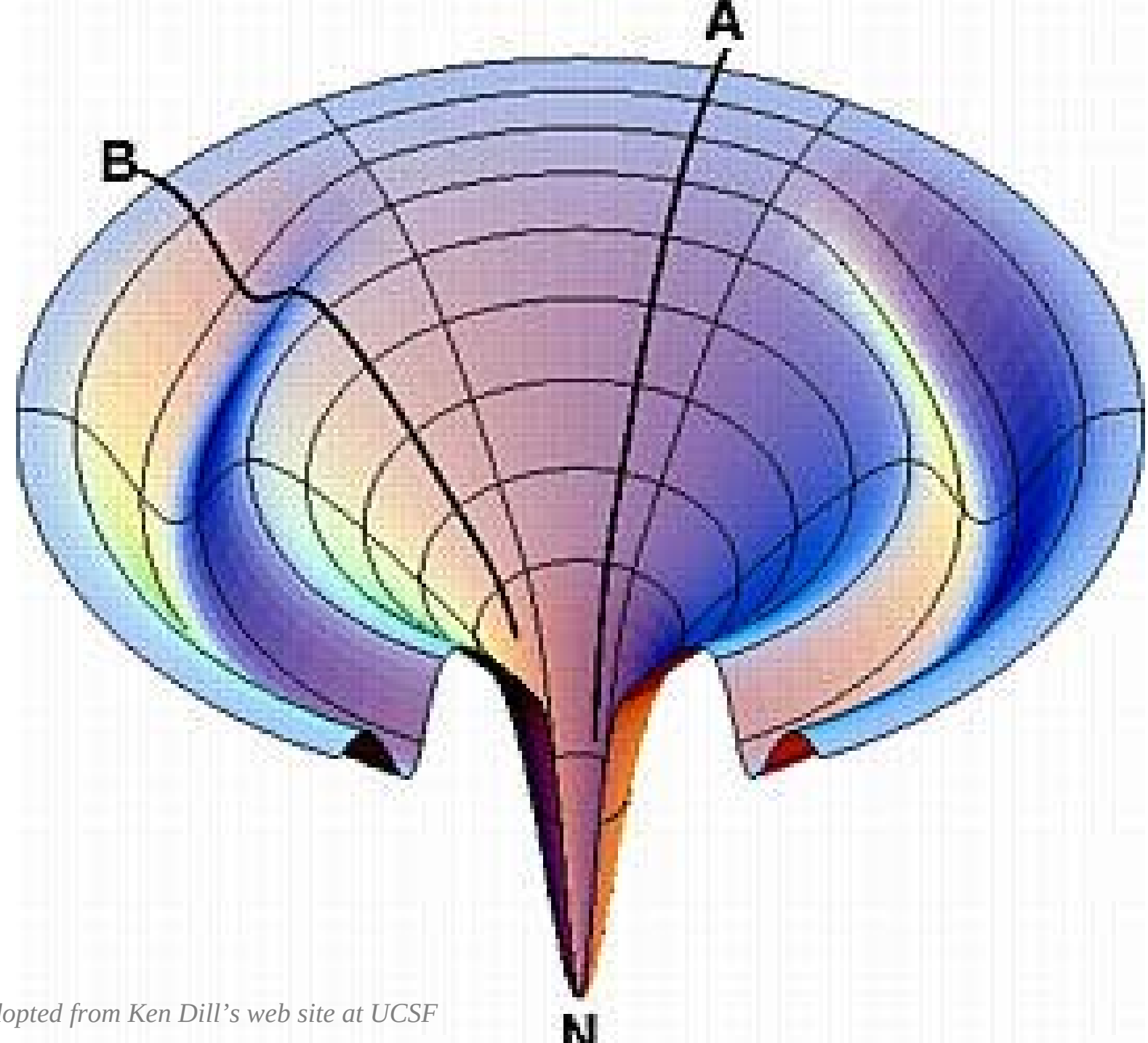


Adopted from Dobson, NATURE 426, 884 2003



Realistic landscapes
are much more complex,
with multiple local minima –
folding traps.

Adopted from Ken Dill's web site at UCSF



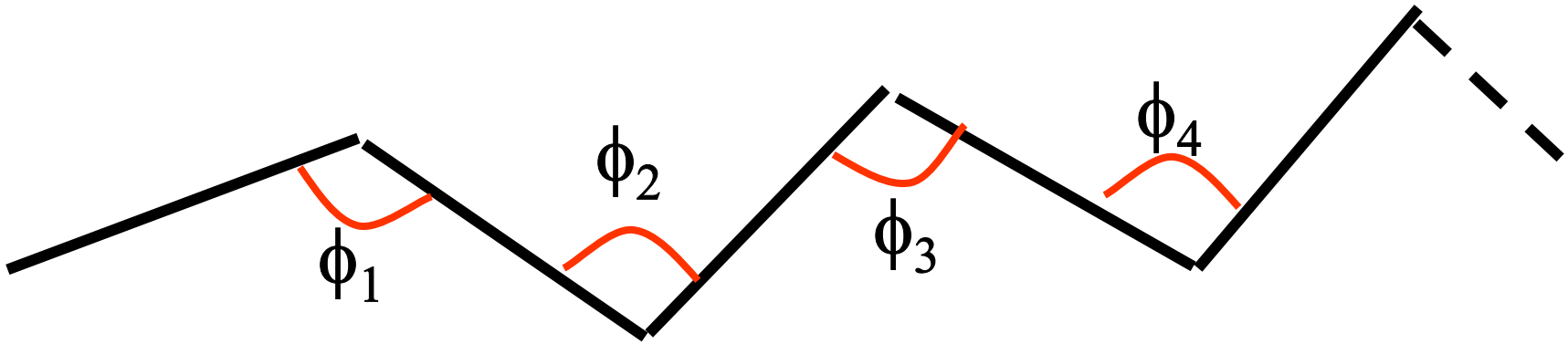
Adopted from Ken Dill's web site at UCSF

Protein Folding Problem #2

- How do folding proteins avoid kinetic traps and mis-folding (that can lead to diseases)?

The magnitude of the protein folding challenge:

Enormous number of the possible conformations of the polypeptide chain



A small protein is a chain of ~ 50 mino acids (more for most).

Assume that each amino acid has only 10 conformations (vast underestimation)

Total number of possible conformations: 10^{50}

Say, you make one MC step per femtosecond.

Exhaustive search for the ground state will take 10^{27} years.

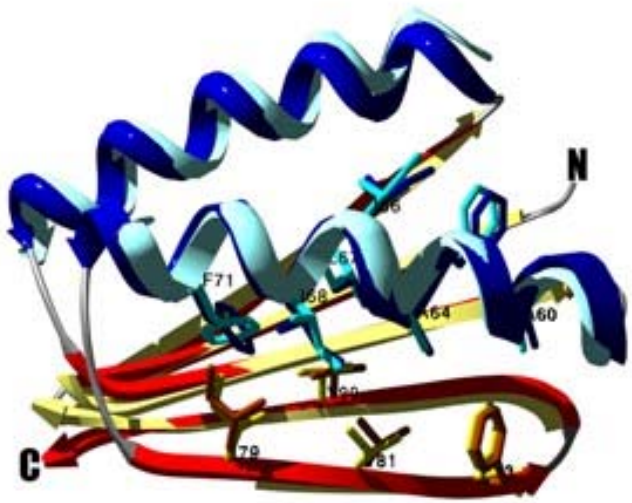
Why bother: protein's shape determines its biological function.

Levinthal's paradox:

- How can proteins EVER fold, given their mind boggling complexity (that is the number of degrees of freedom that the folding protein needs to search through to find the minimum energy state).

First-principles approach to the protein folding problem.

- Physics-based models.
  Recent succssess: top7 protein. Predicted completely from scratch (1A accuracy).



See also: http://folding.stanford.edu/

•One possible solution: model time-evolution of atoms in the protein; follow it from an unfolded state.

PRICIPLE:
Given positions of each atom x(t) at time t, position at t + Δt:

$$x(t + \Delta t) \sim x(t) + v(t)\,\Delta t + \tfrac{1}{2} * F/m * (\Delta t)^2$$

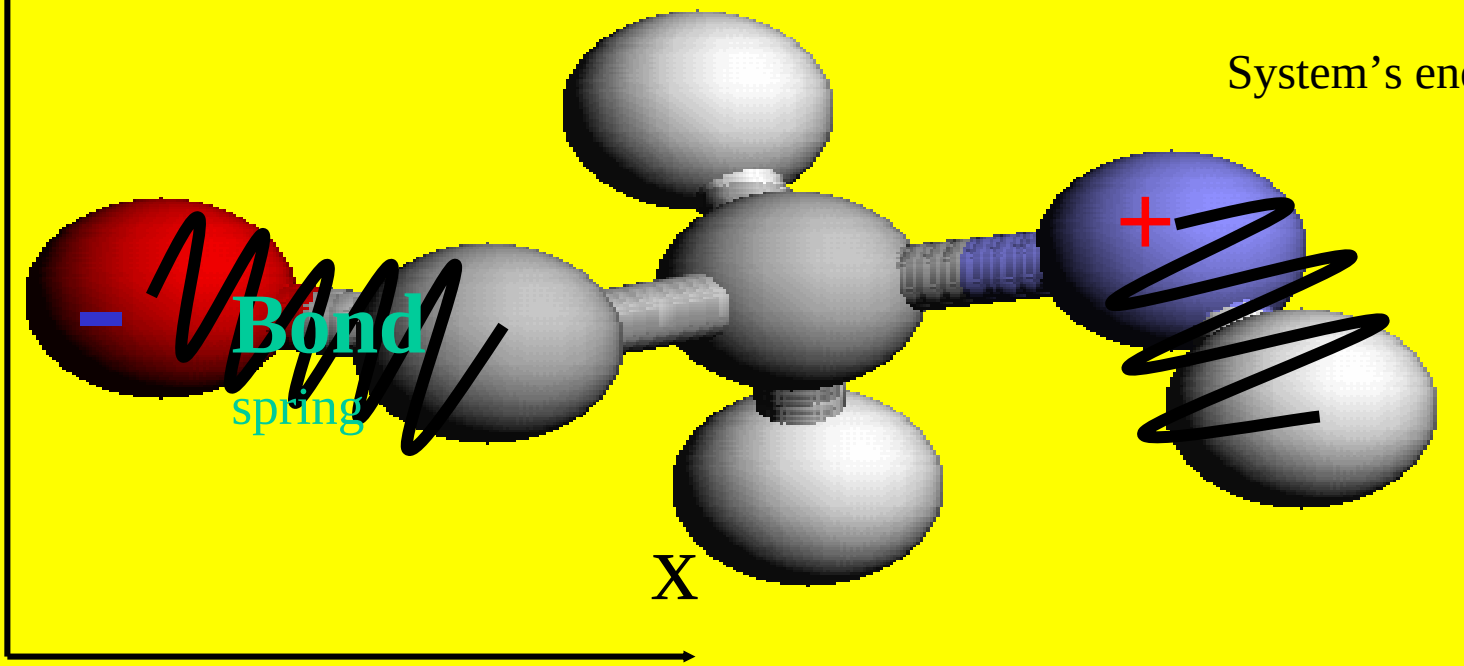
Key parameter: integration time step Δt. Controls accuracy and speed.

Principles of Molecular Dynamics (MD):

Each atom moves by Newton's 2nd Law: $F = ma$

$$F = dE/dr$$

System's energy



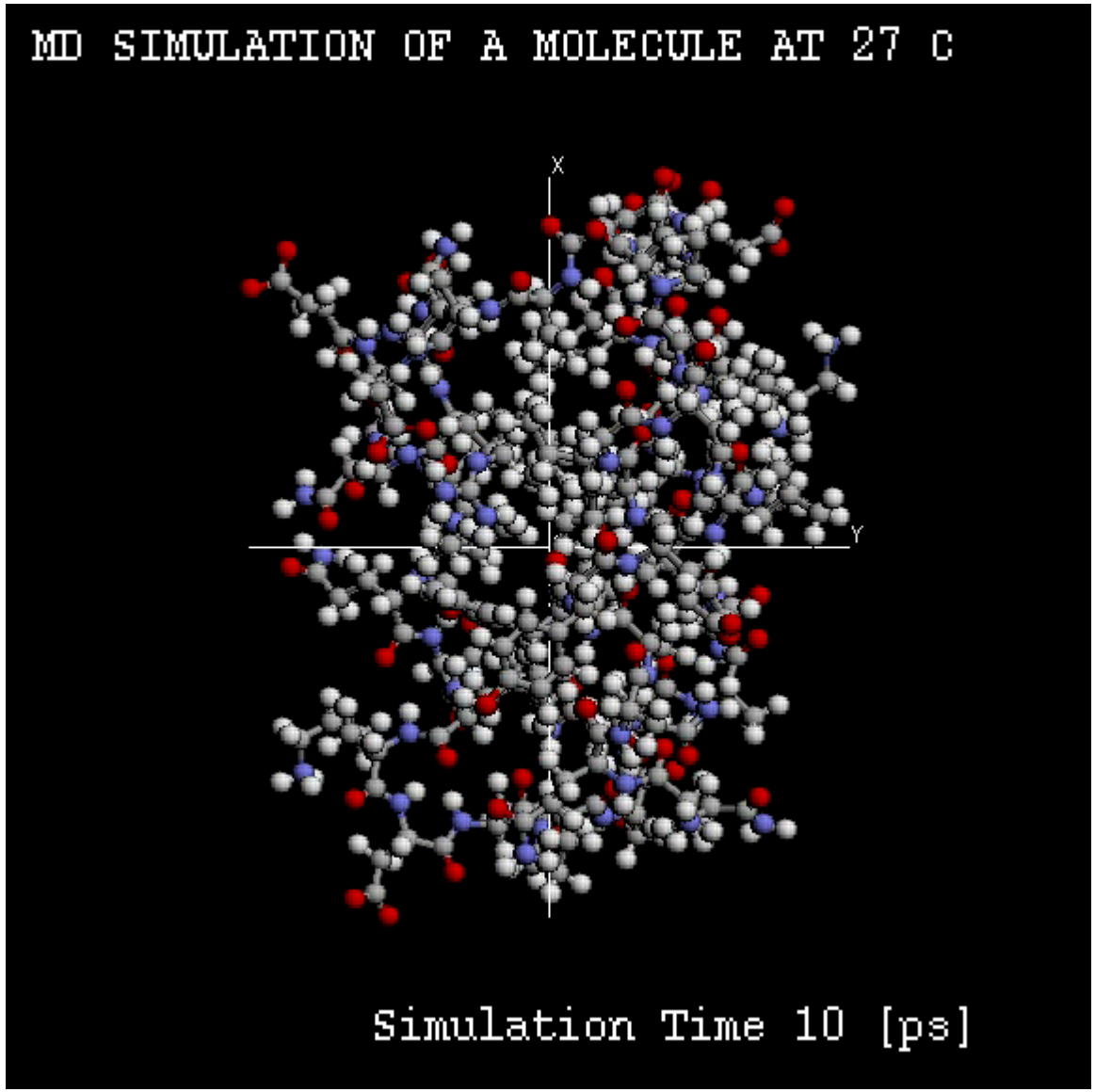
$$E = Kr^2 + A/r^{12} - B/r^6 + Q_1Q_2/r + \ldots$$

Bond stretching

VDW interaction

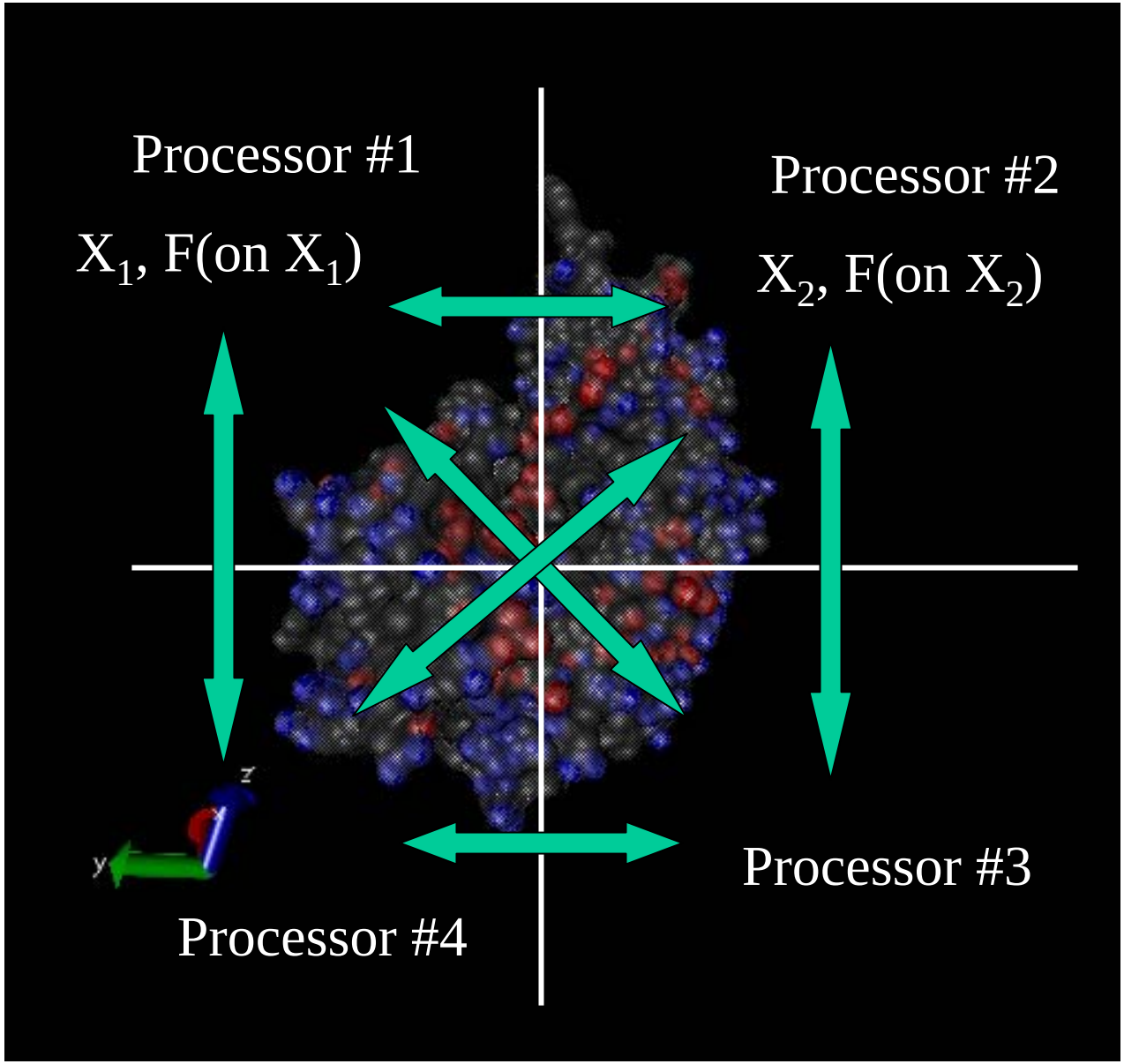
Electrostatic forces

Computer simulations give us positions of all atoms as a function of time.



At normal temperature real molecules are very much alive. Not just solid blocks.

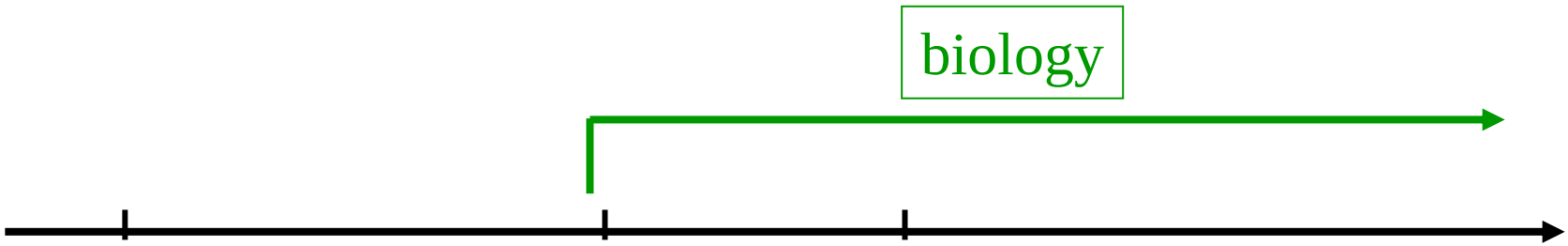
Parallel computation is key, good inter-processor communication is vital



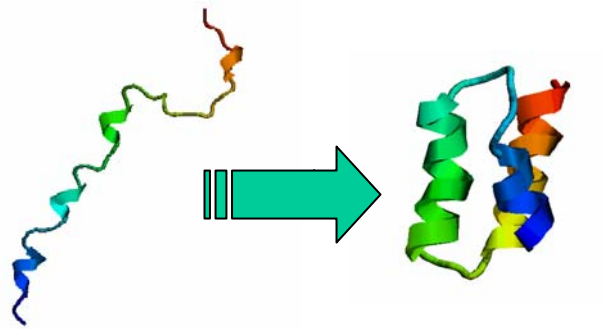
Force acting on each atom depends upon positions of every other atom in the system.

Computed coordinates have to be communicated between all processors at each step

Main Hurdle: biology is too slow (compared to Δt).



H-C bond
vibration



For stability, Δt must be at least an order of magnitude less than the fastest motion, *i.e* $\Delta t \sim$ **10^{-15} s.**

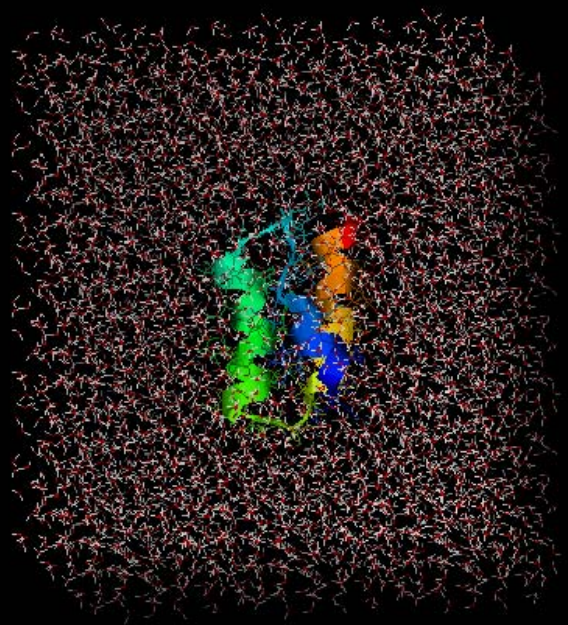
Example: to simulate folding of the fastest folding protein, at least $10^{-6}/10^{-15} = 10^{9}$ steps will be needed .

If you have a large PARALLEL machine, you can study things like:

1. Protein Folding
2. Dynamics of Protein – DNA interaction.
3. Target-ligand docking.
4. Etc.

Computational advantages of representing water implicitly, via a continuum solvent model

Explicit water



Large computational cost. Slow dynamics.

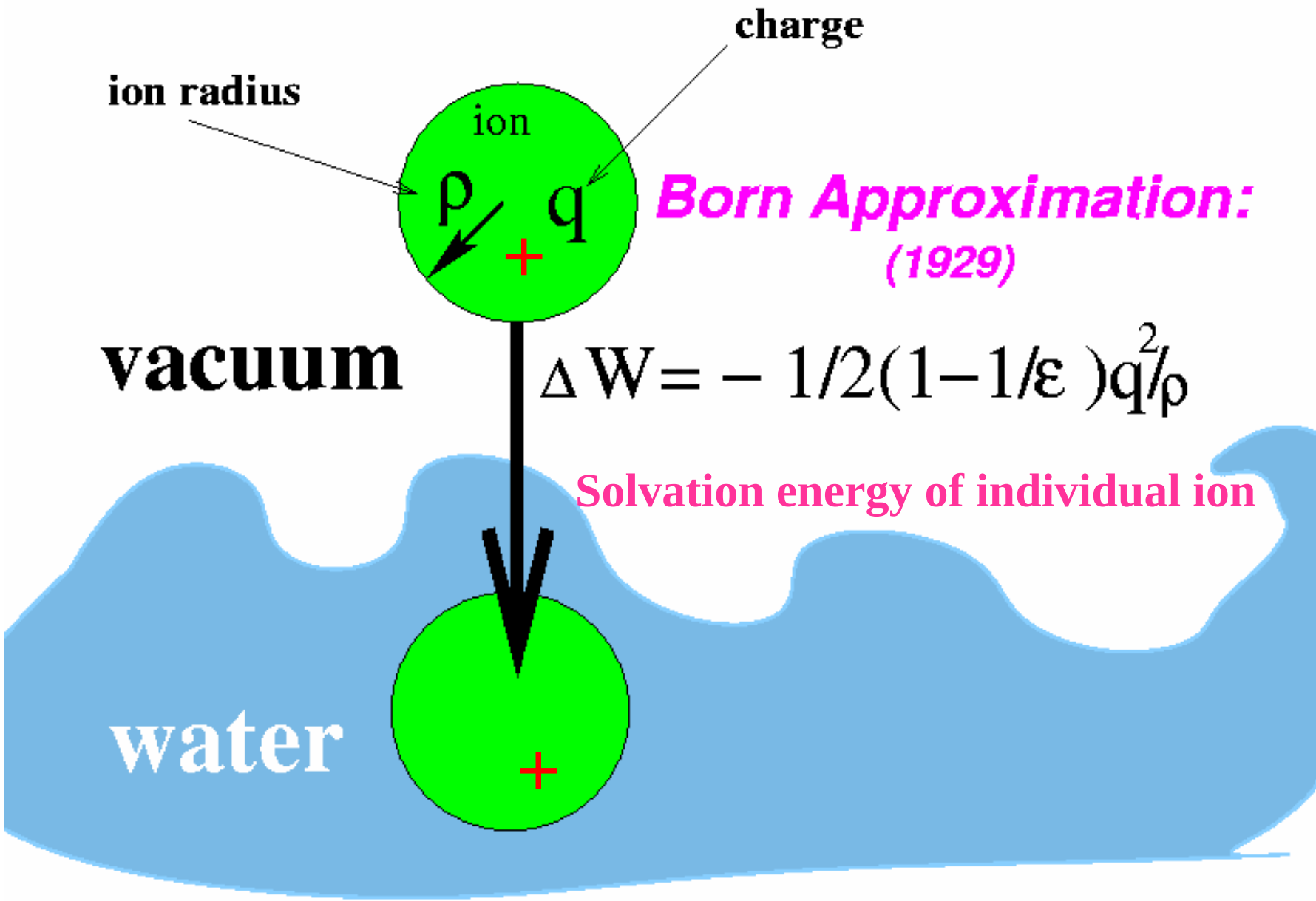
Implicit water as dielectric continuum

Low computational cost. Fast dynamics.

Other advantages:

1. **Instant dielectric response => no water equilibration necessary.**
2. **No viscosity => faster conformational transitions.**
3. **Solvation in an infinite volume => no boundary artifacts.**
4. **Solvent degrees of freedom taken into account implicitly => easy to estimate total energy of solvated system.**

Electrostatic Interactions are key!


charge
ion radius
ion
ρ
q
+
Born Approximation:
(1929)
vacuum
ΔW = − 1/2(1−1/ε)q²/ρ
Solvation energy of individual ion
water
+

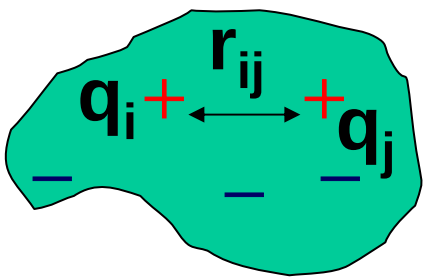
The generalized Born approximation (GB):

Total electrostatic energy

Solvent polarization, ΔW

$$E^{elec} \approx \underbrace{\frac{1}{2}\sum_{i\neq j}\frac{q_i q_j}{r_{ij}}}_{\text{Vacuum part}} - \overbrace{\frac{1}{2}\left(1-\frac{1}{\epsilon_w}\right)\sum_{ij}\frac{q_i q_j}{f_{ij}^{\mathrm{GB}}}}$$

Function to be determined.



molecule

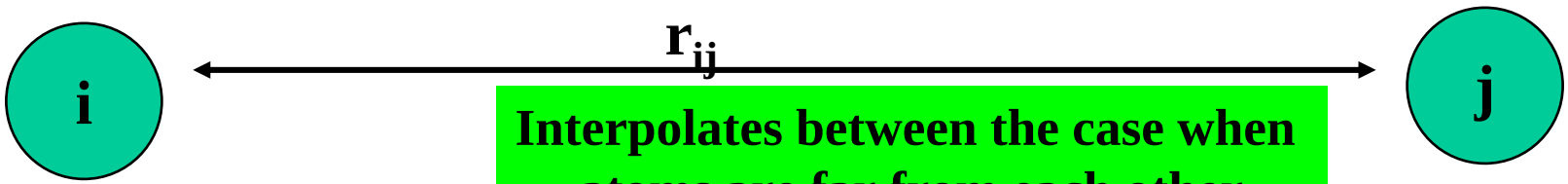
The “magic” formula:

$$f = \left[\underbrace{r_{ij}^2}_{\to 0} + \overbrace{R_i R_j \exp\left(-r_{ij}^2/4R_iR_j\right)}^{\to 0} \right]^{1/2}$$

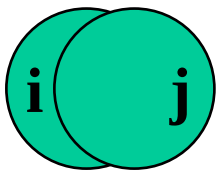
$\to 1$

$f \to r_{ij}$ $E \sim 1/r$

$f \to (R_iR_j)^{1/2}$ $E \sim 1/R$

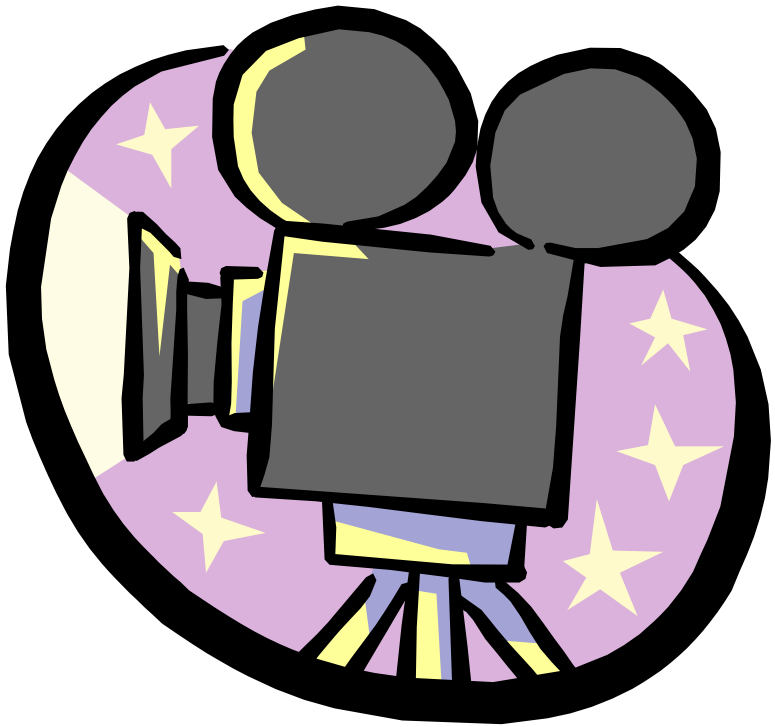


Interpolates between the case when atoms are far from each other and Coulomb’s law is recovered



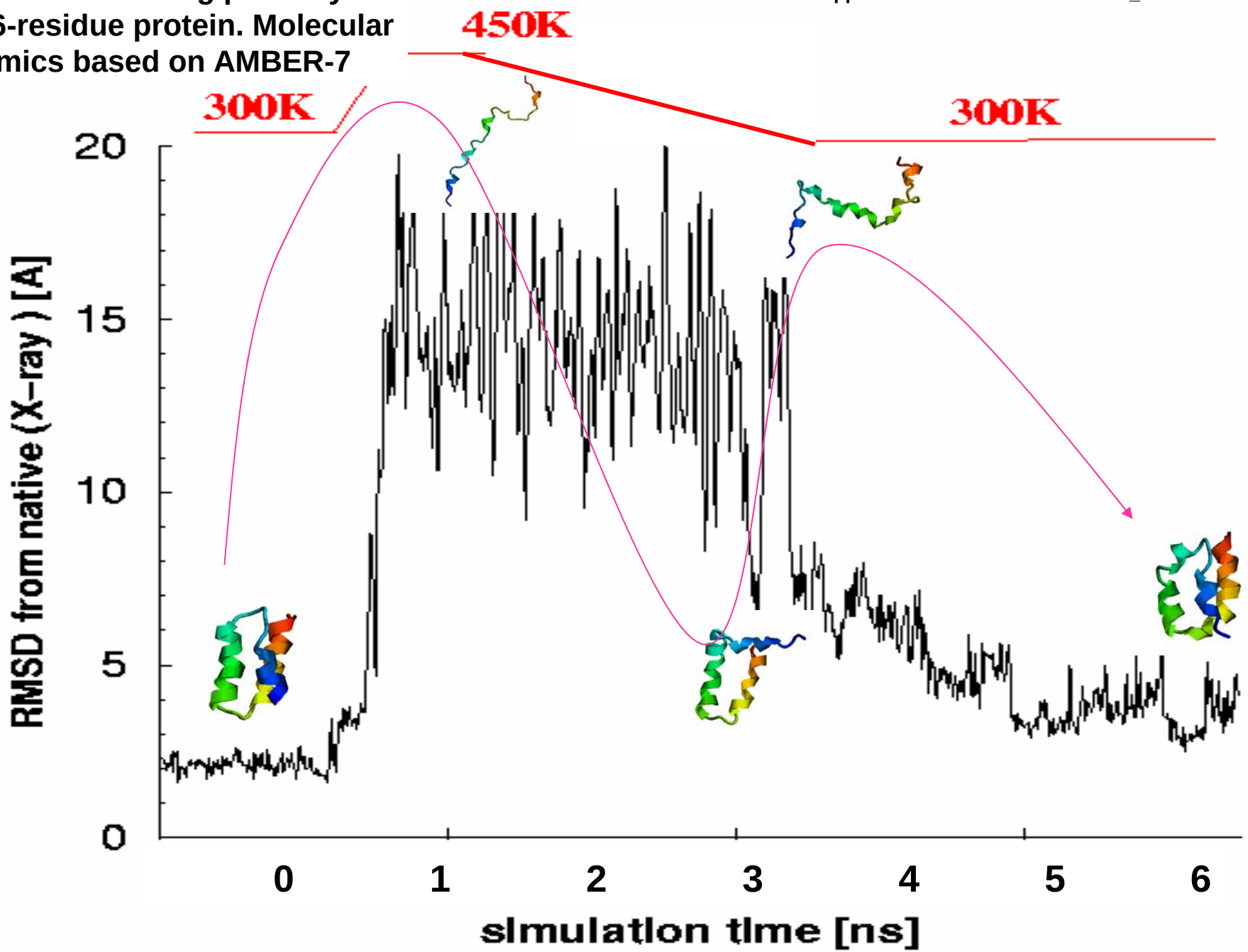
And when they fuse into one, and Born’s formula is recovered

details: Onufriev et al. J. Phys. Chem. 104, 3712 (2000)
Onufriev et al. J. Comp. Chem. 23, 1297 (2002)
Onufriev et al. Proteins, 55, 383 (2004)



Simulated Refolding pathway of the 46-residue protein. Molecular dynamics based on AMBER-7

Movie available at: www.scripps.edu/~onufriev/RESEARCH/in_virtuo.html



NB: due to the absence of viscosity, folding occurs on much shorter time-scale than in an experiment.

Folding a protein *in virtuo*
using Molecular Dynamics based on the Generalized Born (implicit solvation) model.

Simulation time: overnight on 16 processors.

Protein to fold: 46 -residue protein A (one of the guinea pigs in folding studies).

Protocol details: AMBER-7 package, parm-94 force-field.
New GB model.

Recent landmark attempt to fold a (36 residue) protein *in virtuo* using Molecular Dynamics:

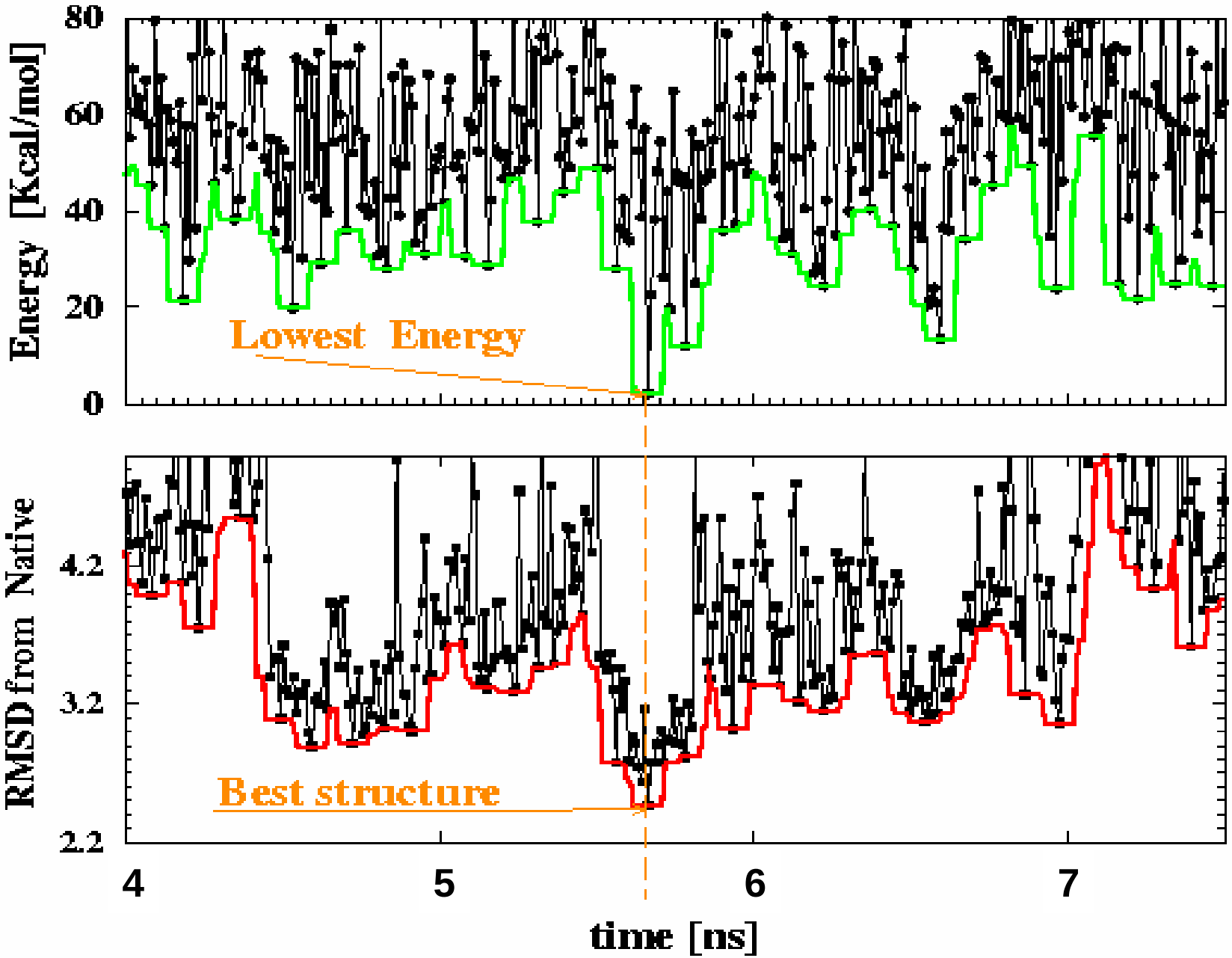
Duan Y, Kollman, P *Science, 282 740 (1998).*

Simulation time: 3 months on 256 processors
= 64 years on one processor.

Result: partially folded structure.

Problem: explicit water simulation are too expensive computationally – can't wait long enough.

The bottom of the folding funnel.

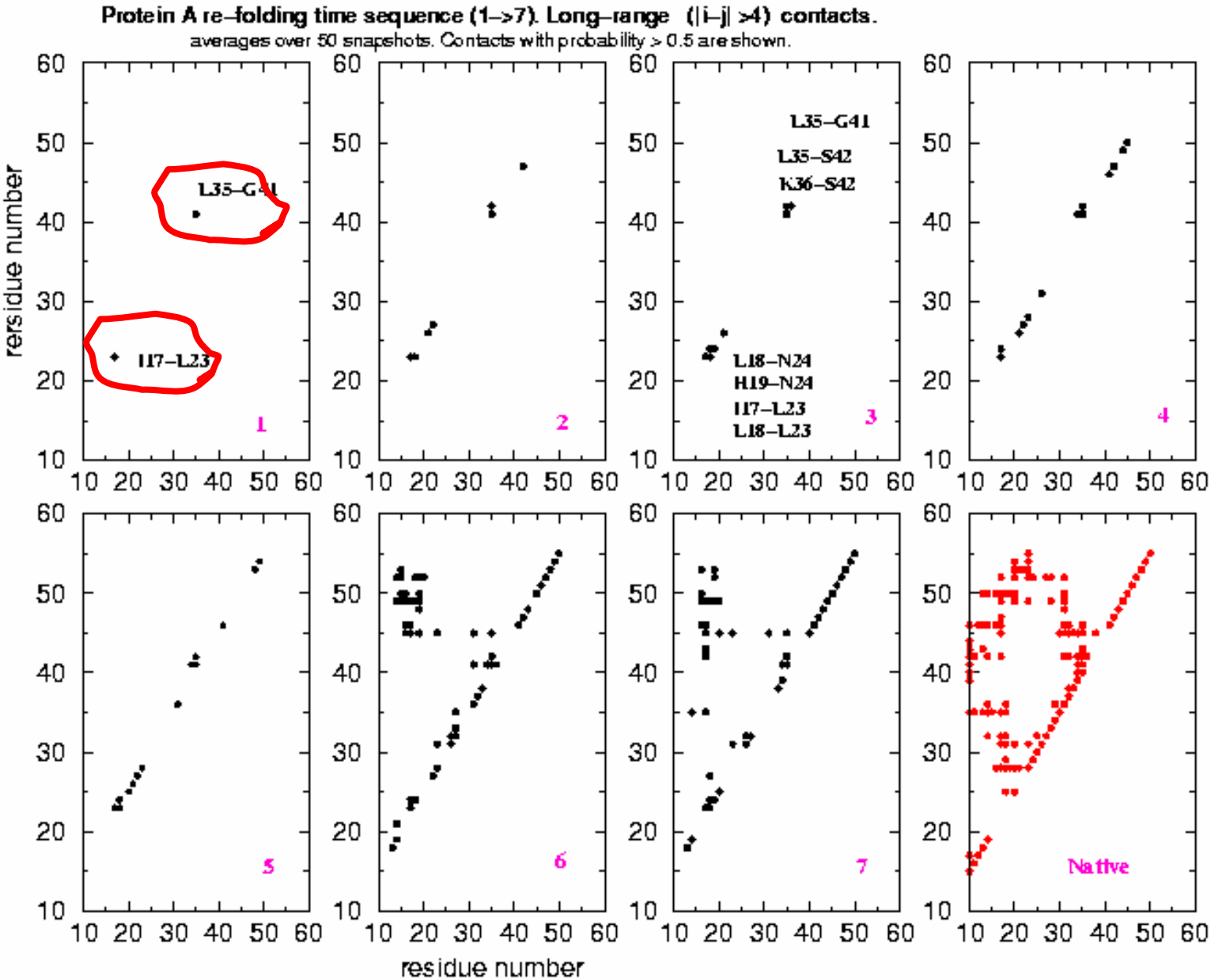


Conclusions:

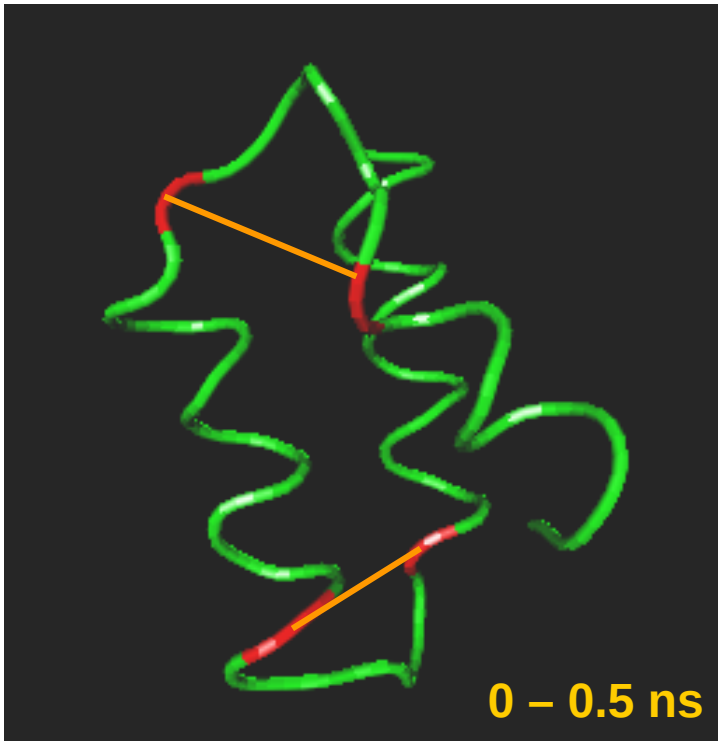
- **Molecular Dynamics based on the improved GB model can be used to fold a 46-residue protein (to backbone RMSD to X-ray 2.4 A, starting from an unfolded state at 450 K.)**

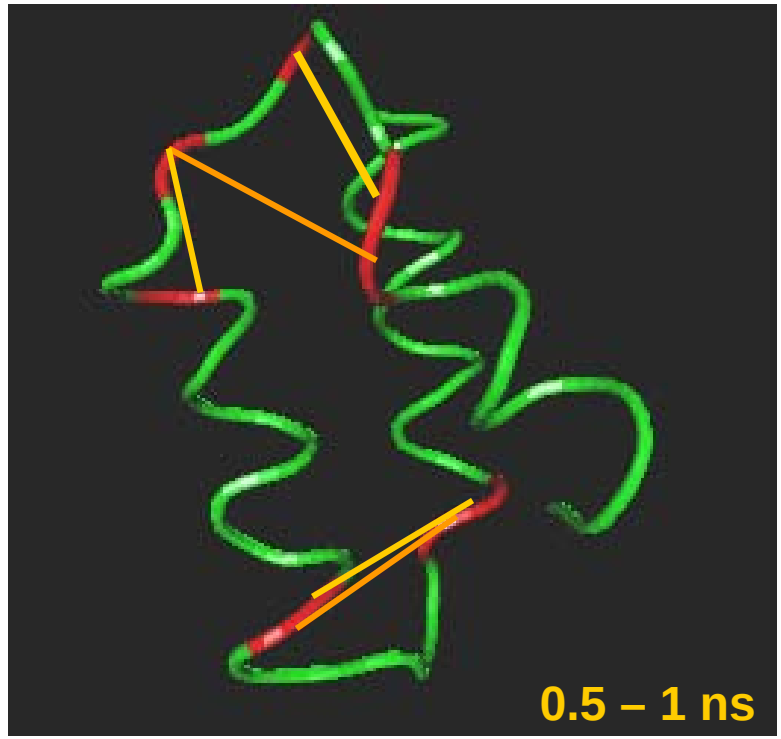
- **Contacts formed in the early stages of folding between residues in the loop regions may direct fast formation of the correct topology.**

Protein-A re-folding steps. Formation of residue-residue contacts



Initial stages of re-folding. **Contacts are formed between residues in the loops.**
(mostly hydrophobic)





Contacts superimposed on the native backbone . t=0 ns corresponds to the unfolded structure.

Hypothesis: restricted motion in the loops may direct fast folding.

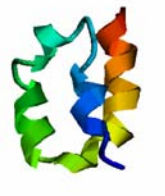
NMR evidence for restricted motions in the unfolded state of apomyoglobin:
Schwarzinger S., Wright, P., Dyson, J. Biochem. 41, 12681, (2002)

Restricted motion in the loops in the unfolded state may be important for fast folding. (directing formation of the correct topology)

1

T=450 K

T=300 K, 5ns



Correct topology

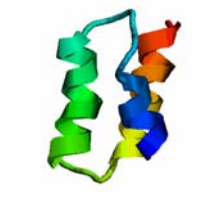
2

T=750 K

T=300 K; 6ns

Wrong topology High energy

3



T=750 K

T=300 K; 3ns

Correct topology

But loop fluctuations slightly restricted.

(ϕ,ψ) of 4 residues. Harmonic potential 2 kcal/mol/rad if deviate more than 20^0 from native values

Correct topology is achieved despite considerable fluctuations of dihedral angles in the high temperature unfolded state.

Conclusions to part I:

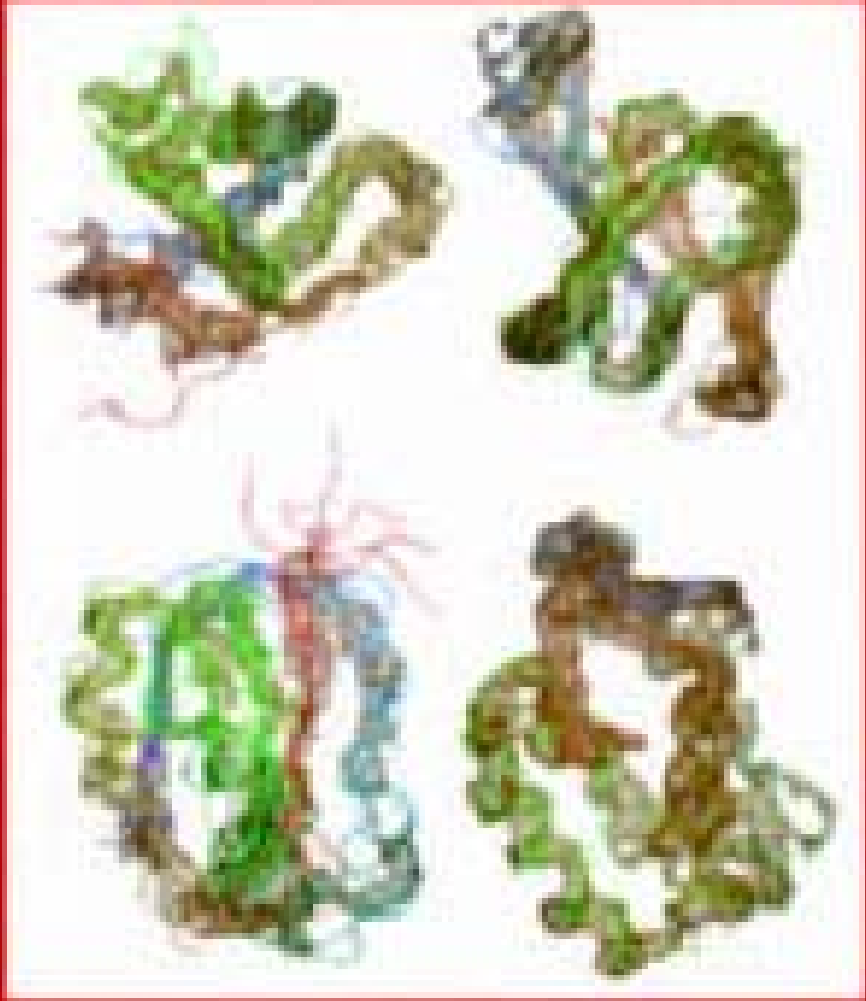
- **Molecular Dynamics based on the improved GB model can be used to fold a 46-residue protein (to backbone RMSD to X-ray 2.4 A, starting from an unfolded state at 450 K.)**

- **Contacts formed in the early stages of folding between residues in the loop regions may direct fast formation of the correct topology.**

nature

insight

protein misfolding



Restricted motion in the loops in the unfolded state may be important for fast folding.

If, instead of 450K, the protein in unfolded at 750K (the rest of the simulation protocol remains the same), it misfolds upon cooling. The misfolded structure **2** has the wrong topology (and higher energy) compared to the native fold **1** achieved in the previous simulation, and represents a kinetic trap. If, however, (ϕ,ψ) dihedral angles of loop residues (4 in each loop) are slightly restrained during the simulation, the protein finds the correct topology immediately upon cooling **3** . Experimentally, restricted motions in the loops are observed in 8M urea unfolded apomyoglobin (S. Schwarzinger, P. Wright, et al.)

Correct topology is achieved despite considerable fluctuations of dihedral angles in the high temperature unfolded state.

End result (300K)

Fluctuations in the unfolded state (0.1 – 1ns) red diamonds – values in the native state.