Speeding up Simulations: Algorithms & Applications

David van der Spoel
Dept. of Cell & Molecular Biology

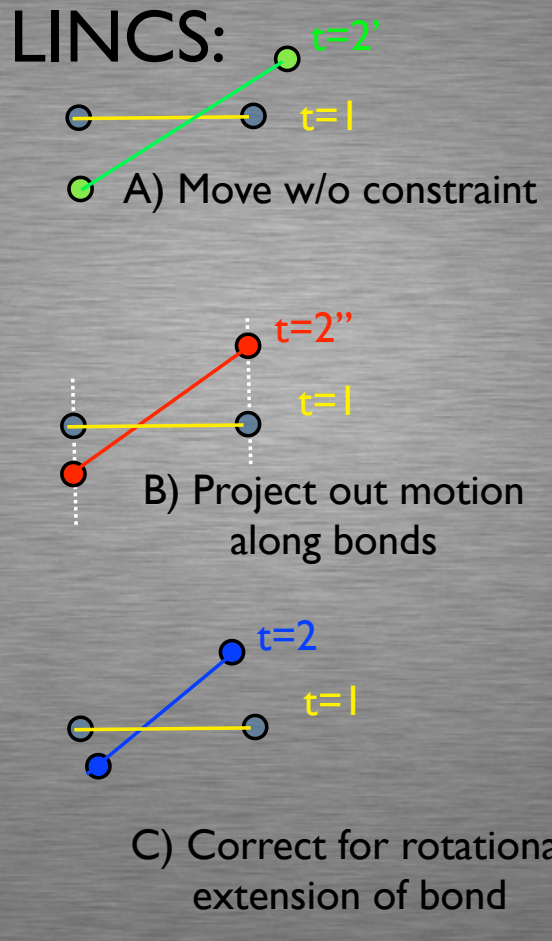
Uppsala Universitet, Sweden

The Need for Speed

- Software
- Hardware

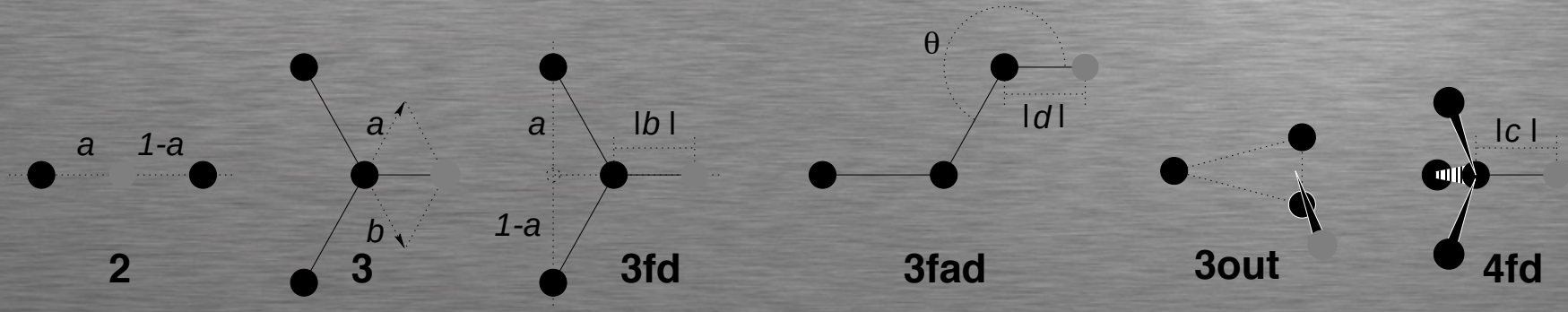
Constraint Simulations

- Δt limited by fast motions - 1fs
  - Remove bond vibrations
- SHAKE (iterative, slow) - 2fs
  - Problematic in parallel
  - Compromise: constrain h-bonds only - 1.4fs
- LINCS in GROMACS:
  - LINear Constraint Solver
  - Approximate matrix inversion expansion
  - Fast & stable -
  - Non-iterative
  - Enables 2-3fs timesteps
  - works in parallel



Going Further: Virtual Sites

- Next fastest motions is hydrogen angle vibrations and rotations of CH_3/NH_2 groups
- Try to remove them:
  - Construct (ideal) H position from heavy atoms. CH_3/NH_2 groups are made rigid
  - Calculate forces, and then project then back on the heavy atoms
  - Integrate only heavy atom positions, reconstruct hydrogens next step
- Our normal simulation setup is to use 4-5fs timesteps with NS every 5 steps, vsite hydrogens and LINCS

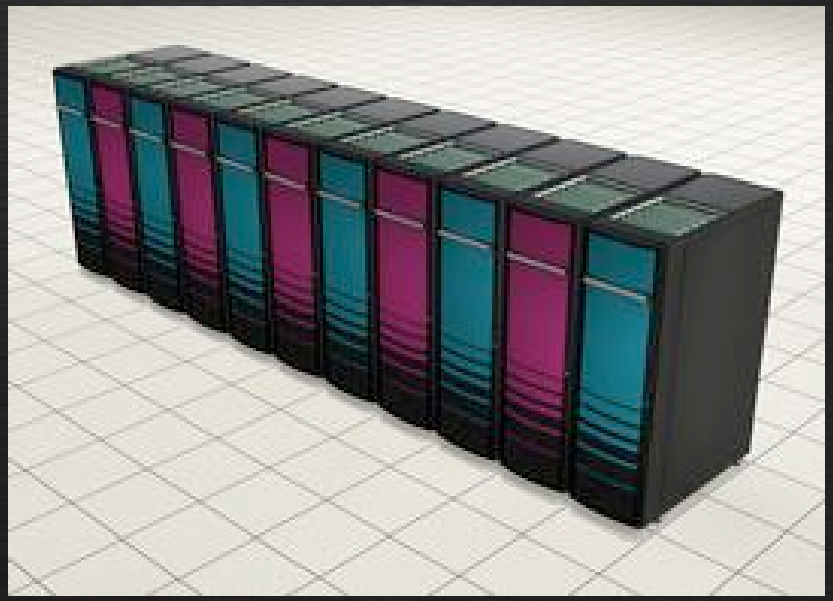


Gromacs virtual site types

Coarse-Graining & Vsites

- Coarse-Grained force fields are getting more popular
- Problem how to interface it with detailed regions
- Virtual interaction sites could be quite effective:
  - Alt 1: Construct all other atoms from CG sites
  - Alt 2: Construct CG sites from other atoms
- Framework for future multiscale modeling simulations!

The Need for Speed

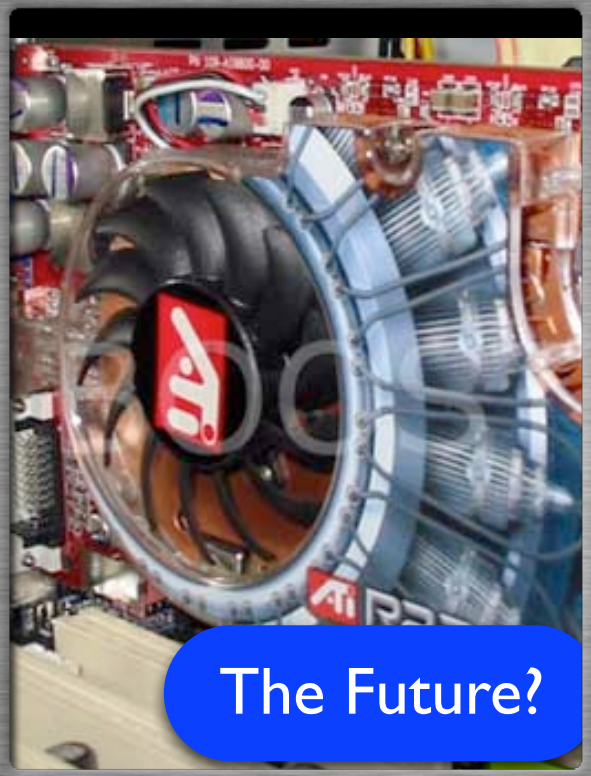


Cray XT-4

IBM Blue Gene

Why is GROMACS fast?

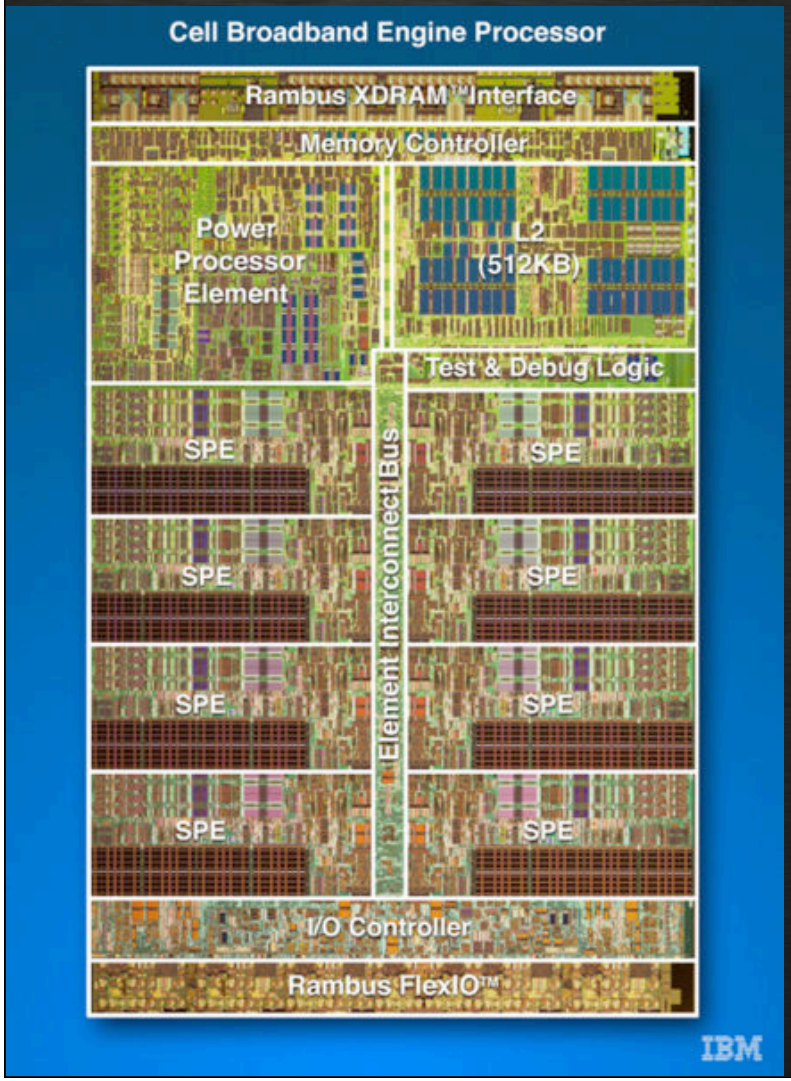
- Algorithmic optimization:
  - No virial in nonbonded kernels
  - Single precision by default
  - Tuning to avoid expensive statements such as PBC checks
  - Triclinic cells everywhere: saves 15-20% for a given system radius
- Optimized 1/sqrt(x)
  - Used ~75,000,000 times/sec
  - Assembly innerloops for x86, x86-64, ia64, Altivec, VMX,



The Future?


PLAYSTATION 3
PLAYSTATION 3
PLAYSTATION 3
PLAYSTATION 3
PLAYSTATION 3
PLAYSTATION 3
PLAYSTATION 3
PLAYSTATION 3

Cell Processor



- 64 bit Power CPU
- 512 kb Cache
- 8 Synergistic Processing Units

Protein Folding Properties from MD Simulations

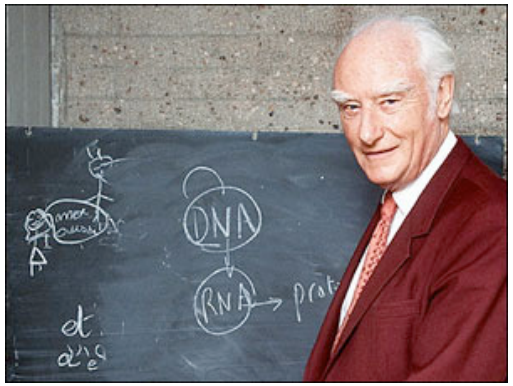
The Protein Folding Problem

"... everything living things do can be understood in term of the wiggling and jiggling of their atoms."

Richard Feynman

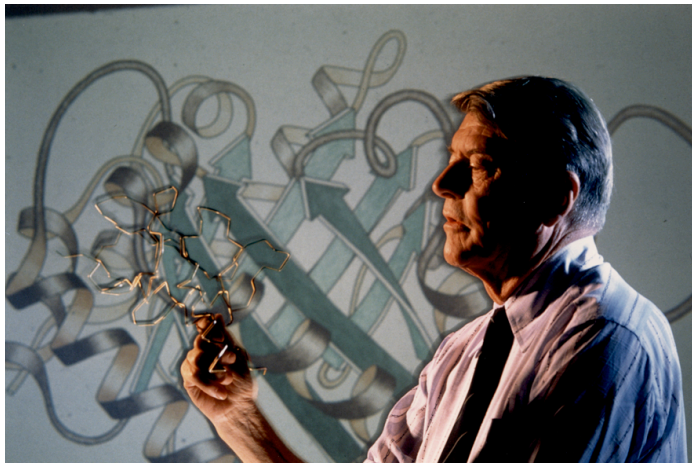
"folding is simply a function of the the order of the amino acids."

Francis Crick



Christian Anfinsen

"...the conformation that a protein assumes [...] is the one that is thermodynamically the most stable."



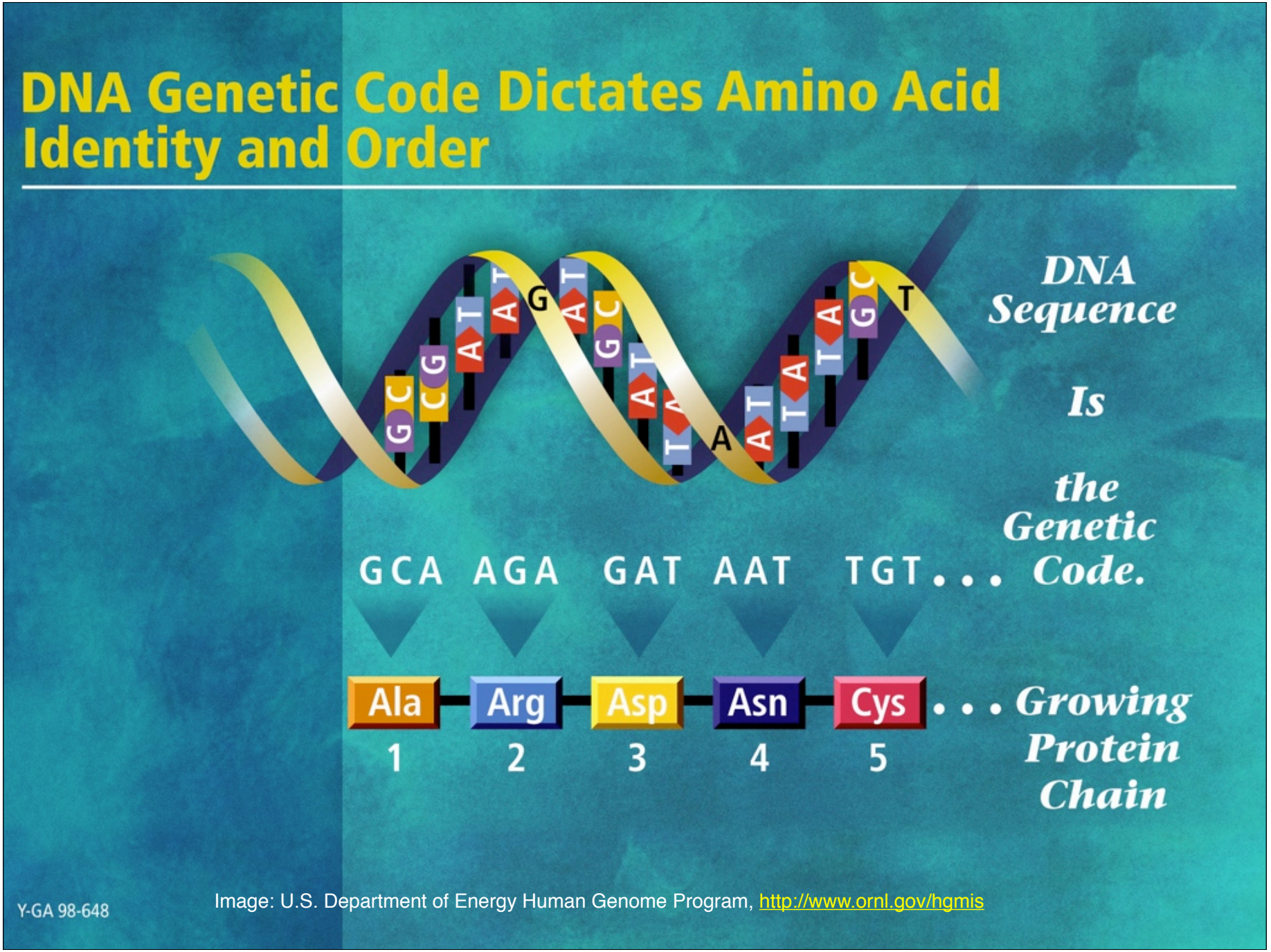


Image: U.S. Department of Energy Human Genome Program, http://www.ornl.gov/hgmis

Molecular dynamics simulations

- Calculate energies and forces using a classical force field
- Solve Newton's equations of motion (F = m a)
- Can in principle describe processes such as protein folding

- Efficient software: GROMACS (http://www.gromacs.org)

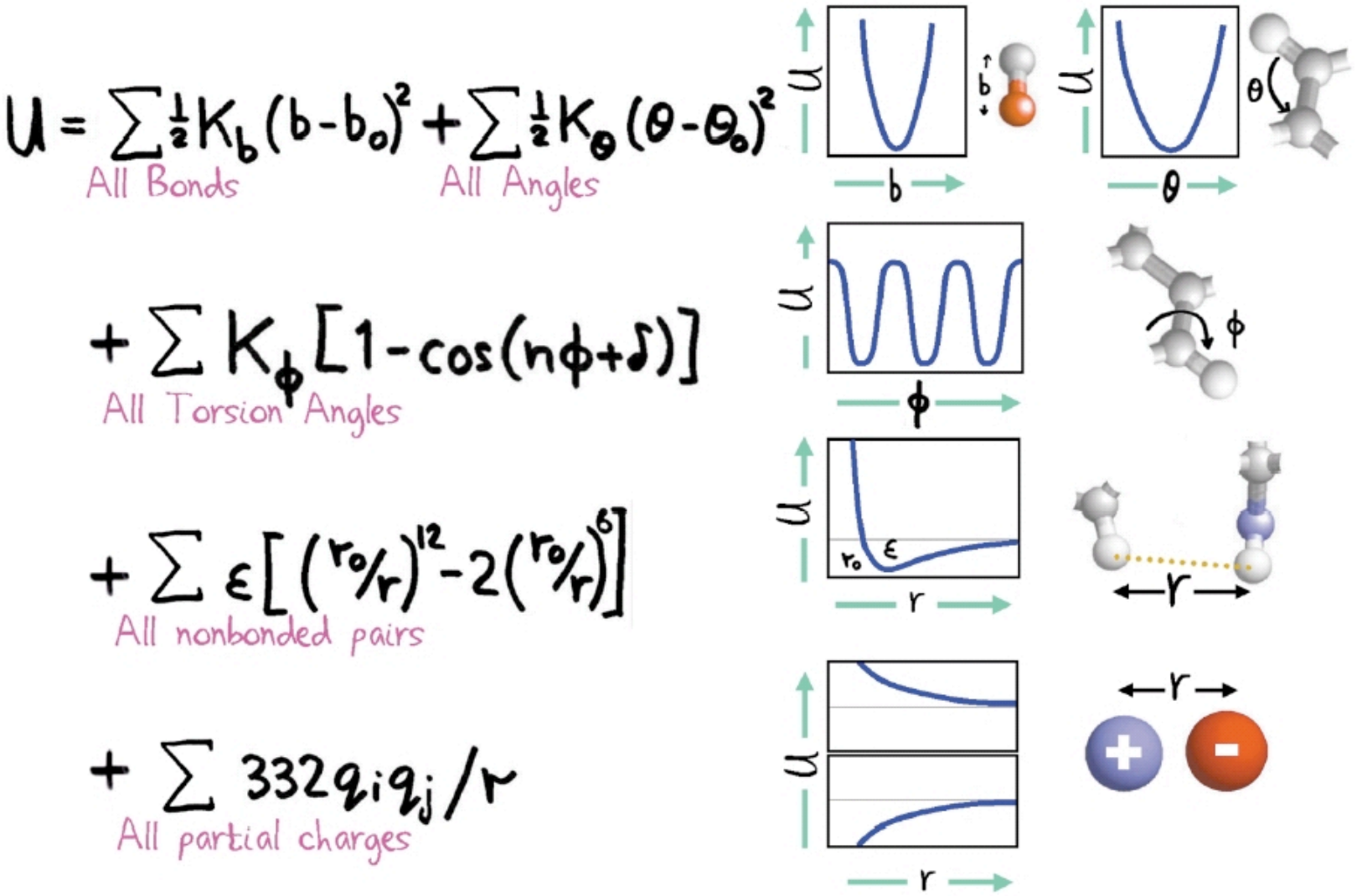
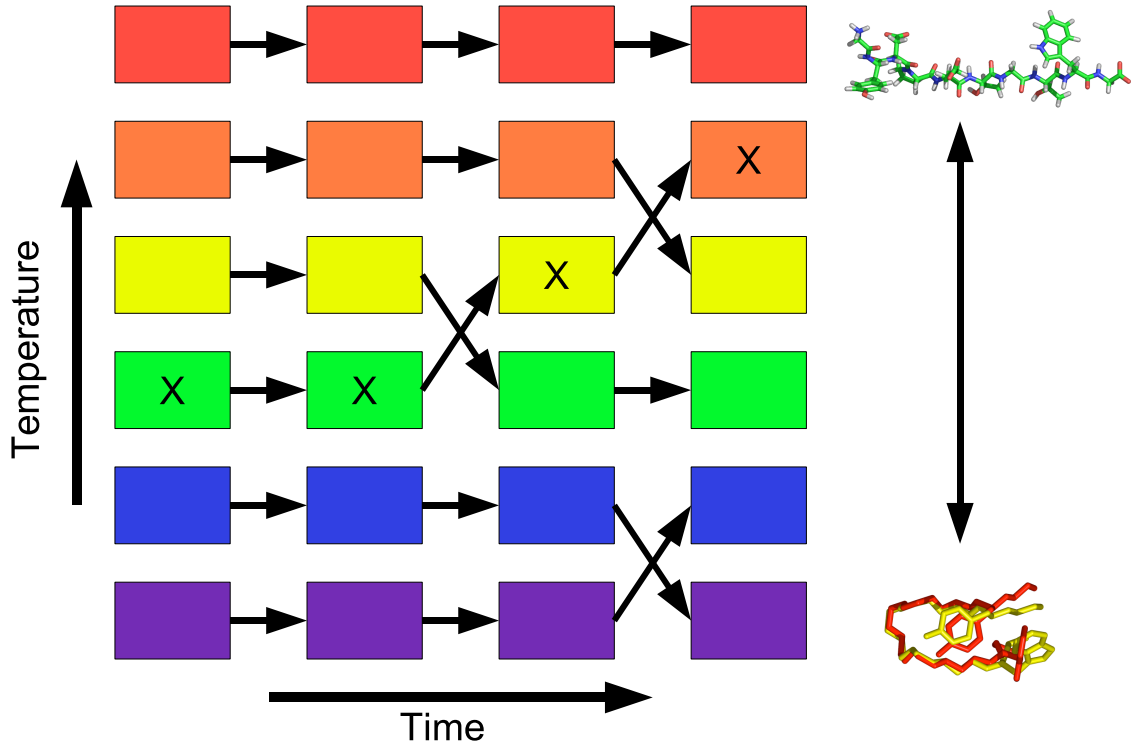


Fig. 1 The total potential energy of any molecule is the sum of terms allowing for bond stretching, bond angle bending, bond twisting, van der Waals interactions and electrostatics. Many properties of a biomolecules can be simulated with such an empirical energy function.

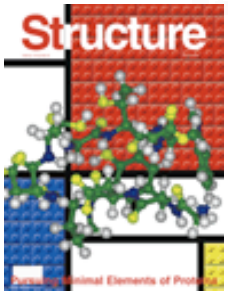
M.Levitt, Nat. Struct. Biol. 8, 392-393 (2001)

Replica exchange MD simulations

- Run multiple copies of a simulation at different temperatures, e.g. 280 K, 285 K, 290 K, etc.
- Exchange coordinates between adjacent temperatures every N time steps, based on a Metropolis criterion.
- Enhanced sampling at many temperatures



Chignolin

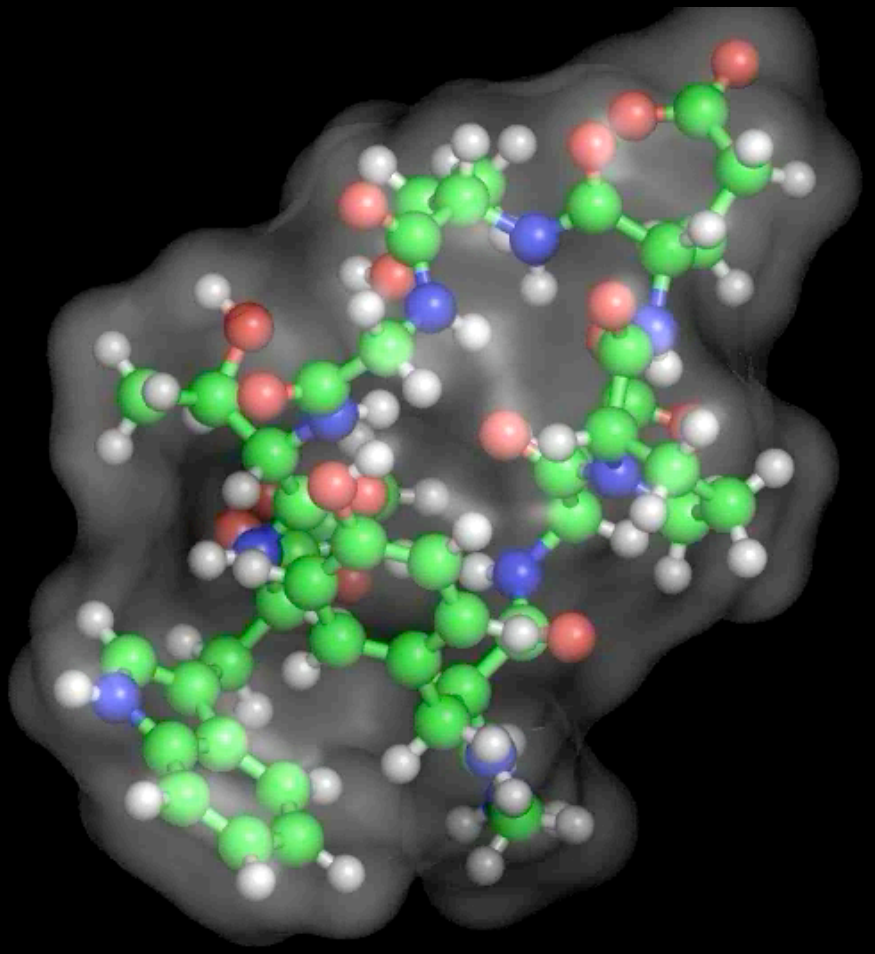


Honda et al,
Structure 2004

- The world's smallest Protein?
- 10 amino acids long
- Forms a stable β-hairpin in solution

NMR solution structure derived from 174 NOE interatomic distance restraints

18 NMR Structures of Chignolin



MD Simulations of Chignolin

Long classical MD trajectories

- 1.8 and 2.0 μs @ 300 K, 0.5 ns @ 277 K, 367 K.
- Explicit solvent (TIP4P)
- Particle mesh-Ewald treatment of Coulomb interactions
- OPLS force field

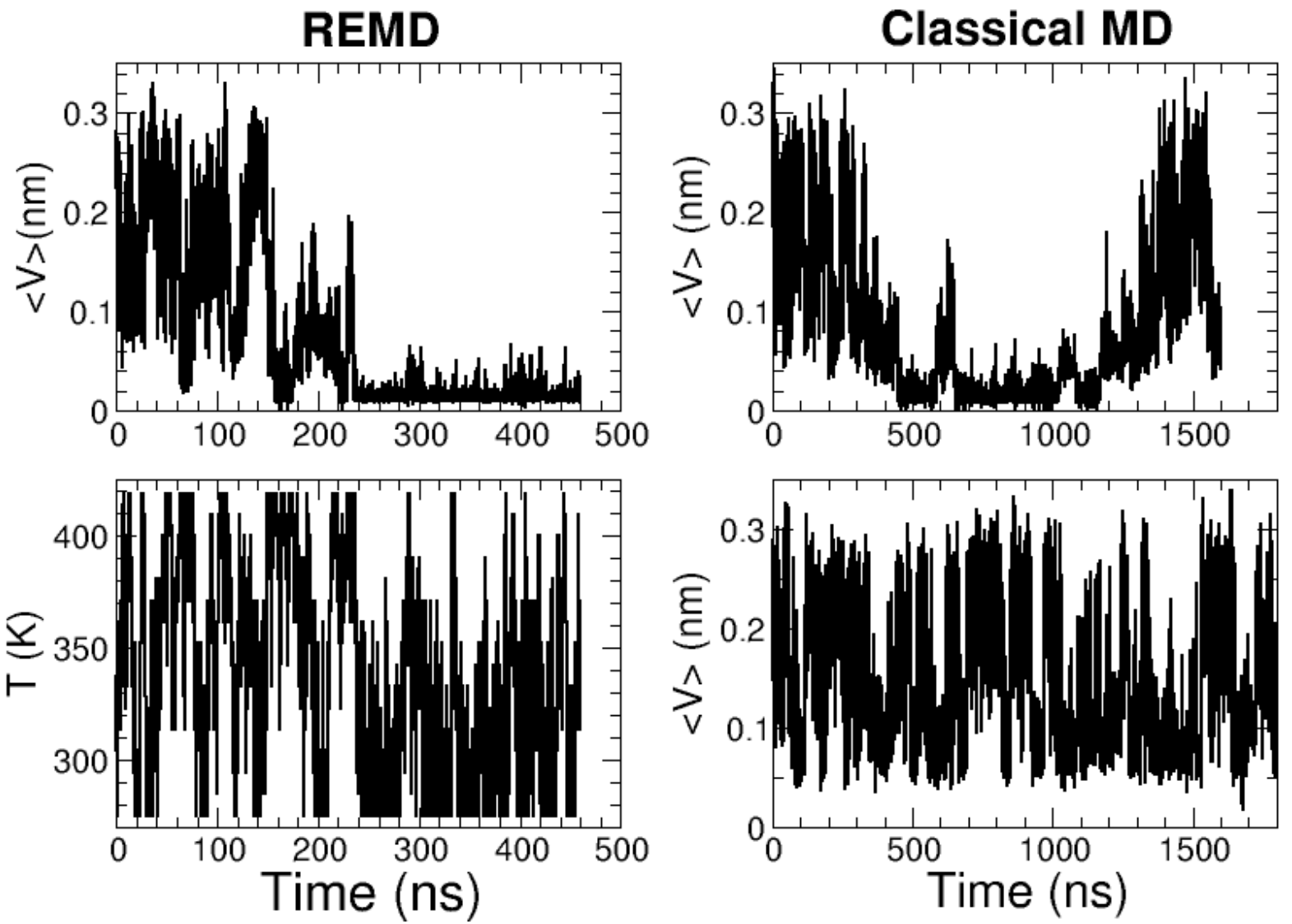
REMD trajectories

- 16 different T from 275 to 420 K
- 510 ns

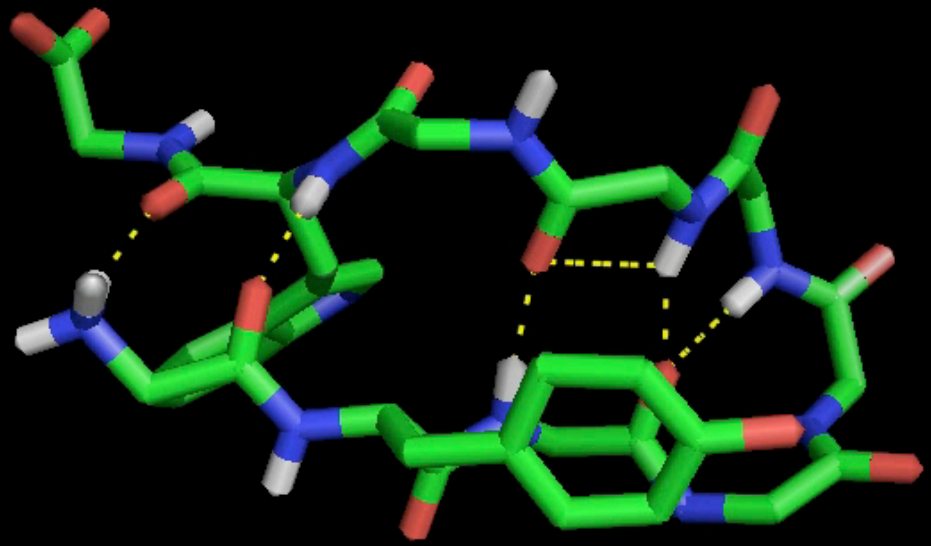
Analysis of simulations

- Distance violations <V> from experimental data
- Gibbs energy landscape

Convergence of MD simulations at T = 300 K



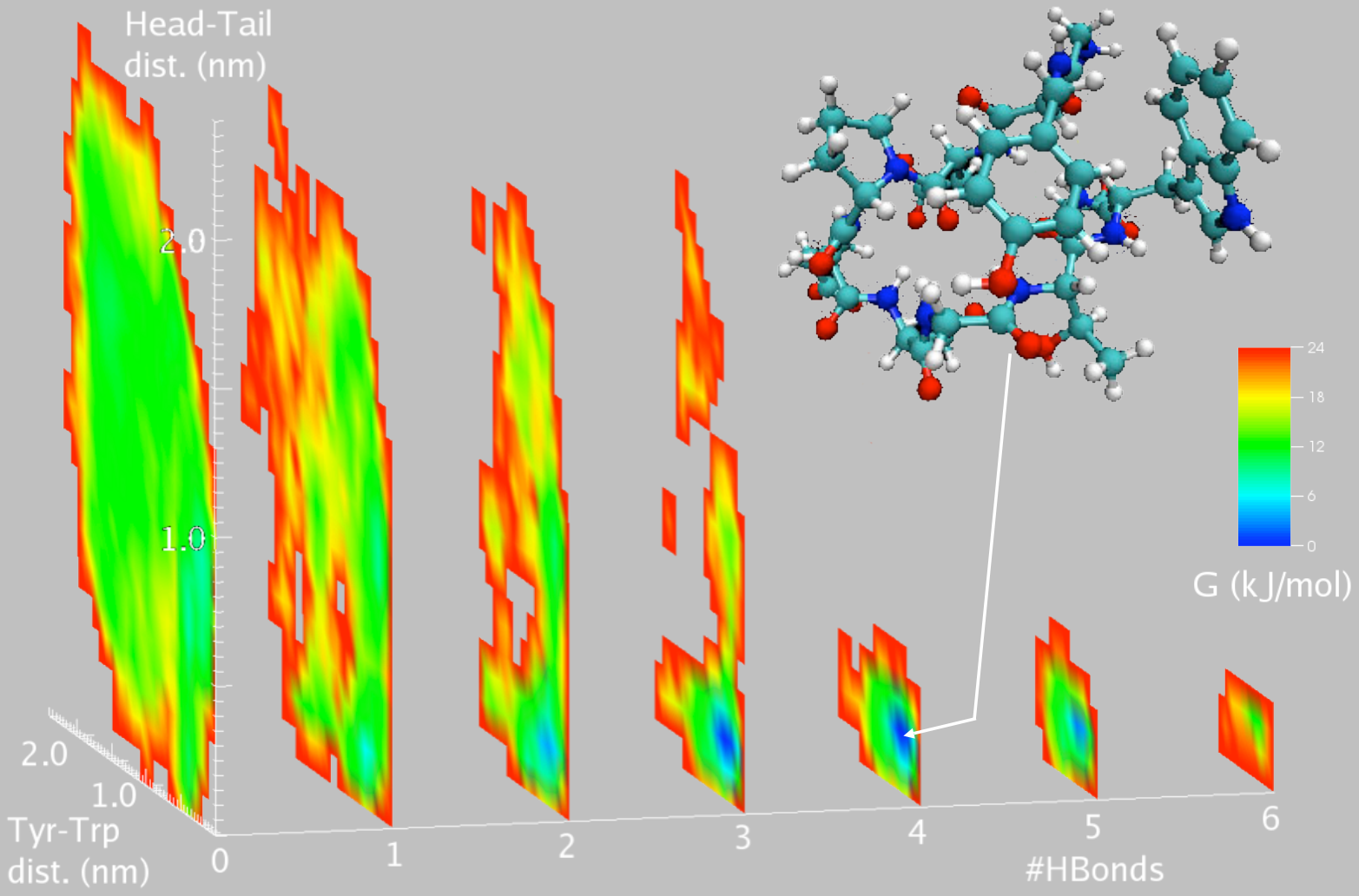
A Folding Event from Classical MD



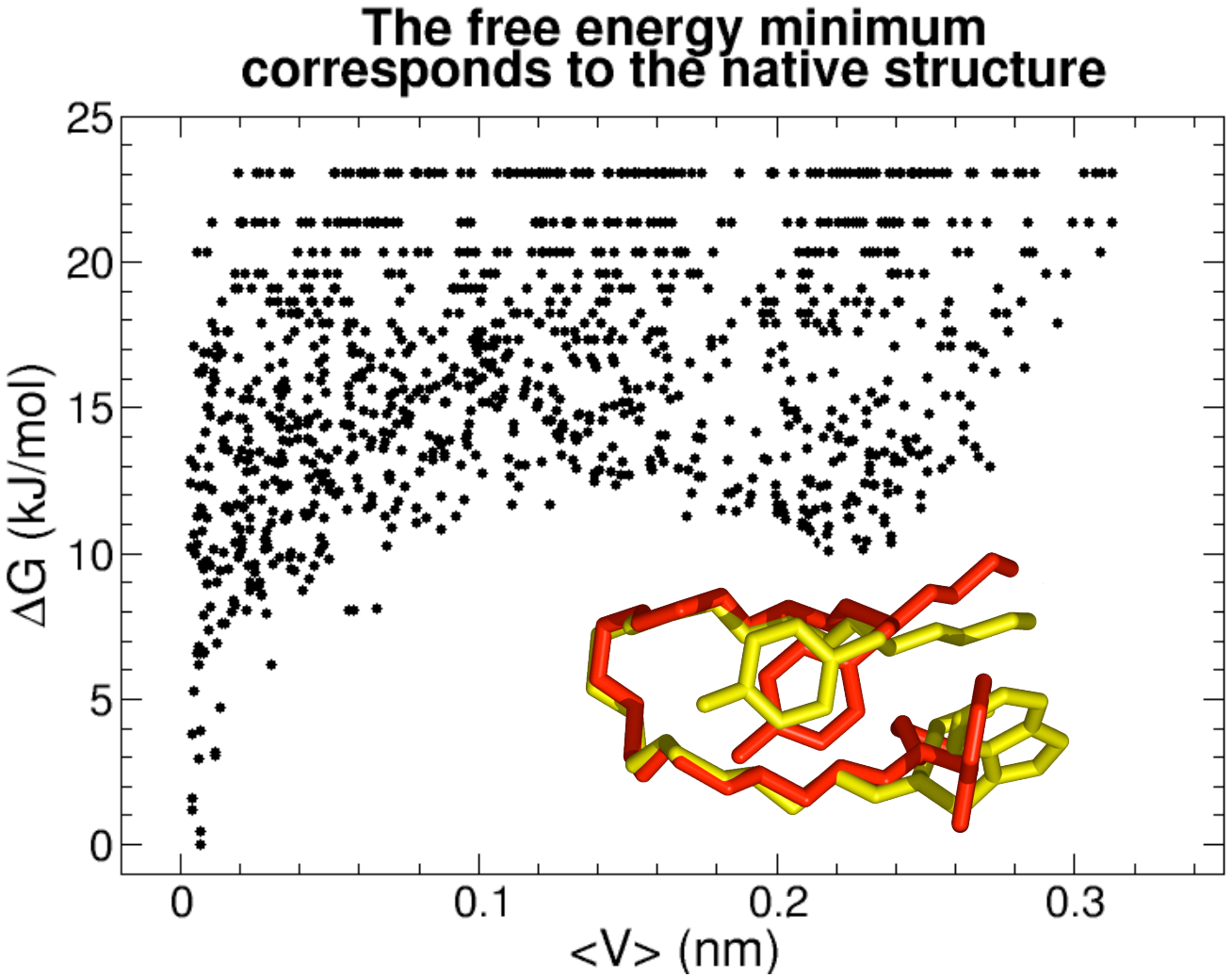
Energy landscape analysis @ 300 K

- Determine a suitable space (in this case in 3 dimensions)
- Make a histogram of the space and compute relative probabilities P(x,y,z) of finding a protein conformation in the bin
- Find most probable bin ==> P_{min}
- $\Delta G = -k_B T \ln P(x,y,z)/P_{min}$
- $\Delta G_{min} = 0$
- g_sham program

Energy Landscape

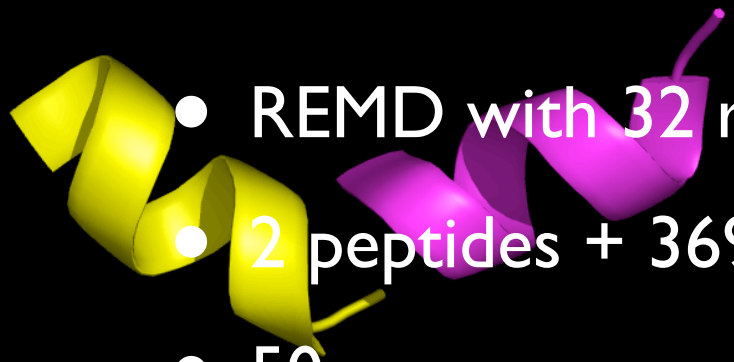


Thermodynamic hypothesis (Anfinsen)

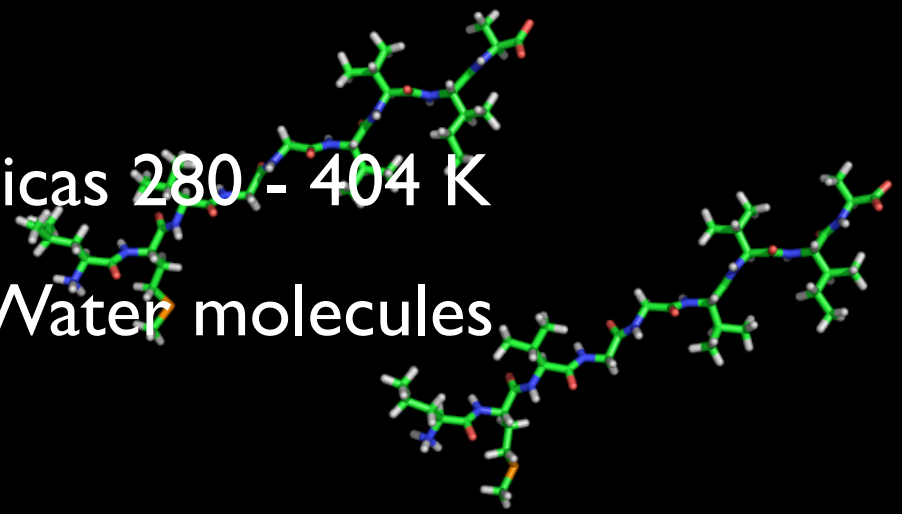


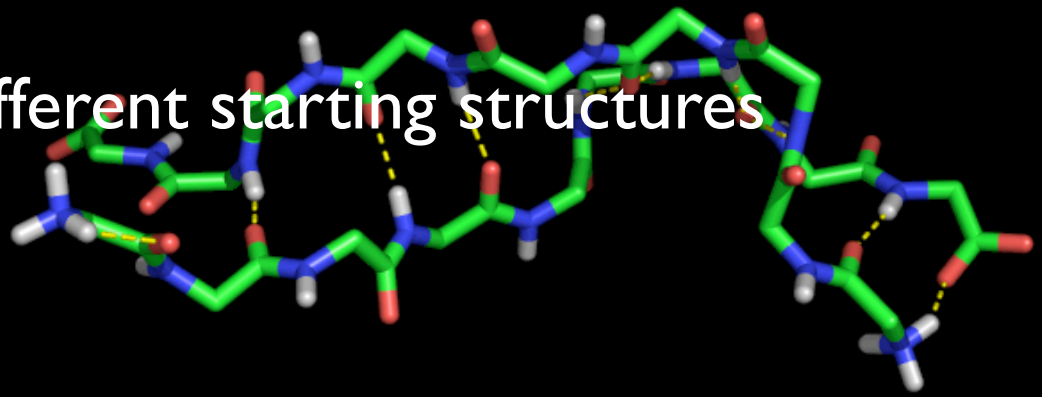
M. Seibert *et al.* J. Mol. Biol. **354** (2005) 173-183

Amyloid A > B Conversion

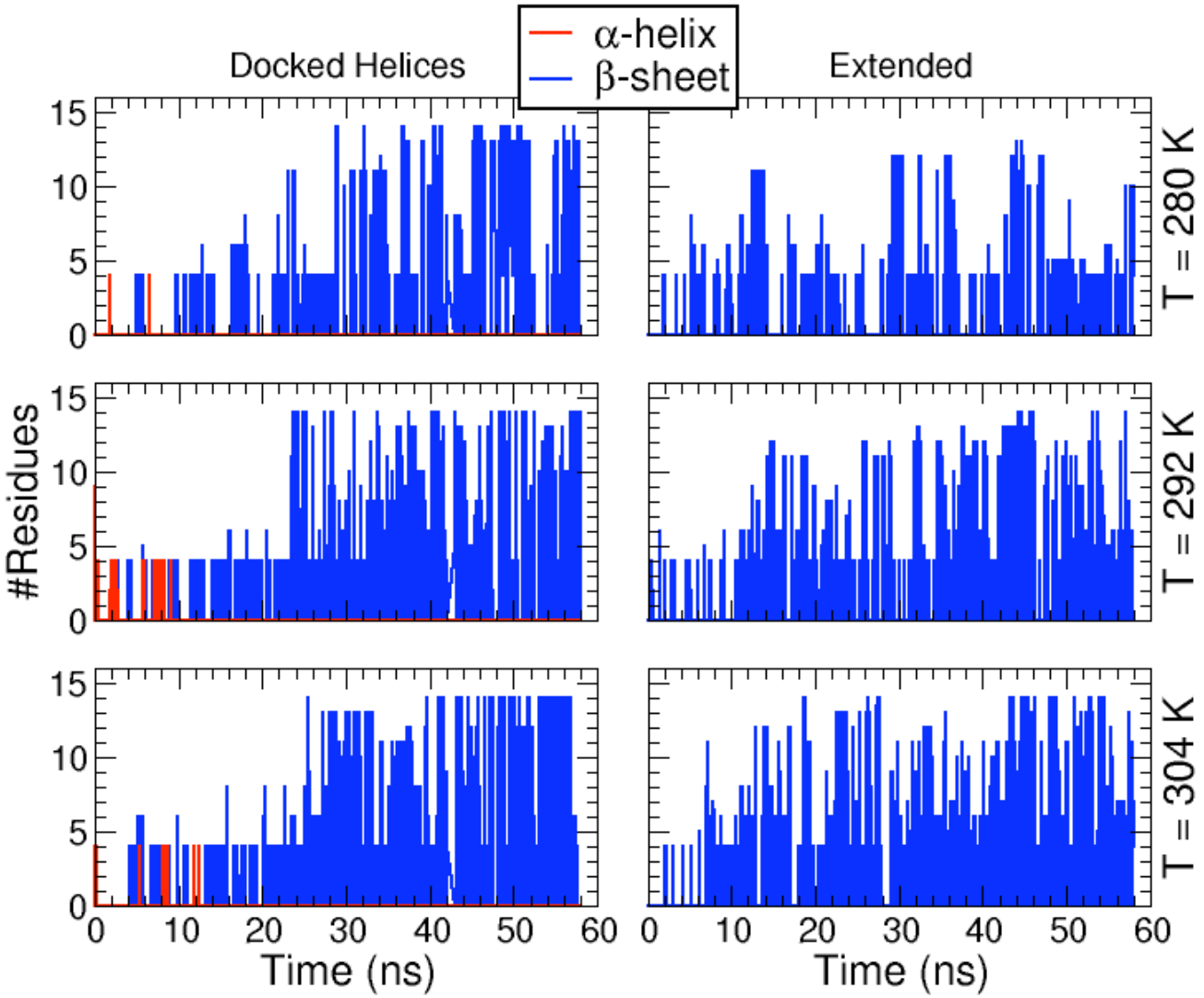


- REMD with 32 replicas 280 - 404 K
- 2 peptides + 3690 Water molecules
- 50 ns
- OPLS/AA + TIP4P
- Different starting structures





Secondary Structure



Folding kinetics from MD

- Using (RE)MD we obtain many trajectories in which proteins fold and unfold repeatedly
- Decide whether a protein is folded based on e.g. RMSD to the native state
- Compute the change in fraction folded F(t) for simulation m:

$$\frac{dF_m(t)}{dt} = k_f U_m(t) - k_u F_m(t)$$

Folding kinetics from MD

The rate constants are defined by:

$$k_u = A_u \mathrm{e}^{-\beta E_A^u}, \quad k_f = A_f \mathrm{e}^{-\beta E_A^f}$$

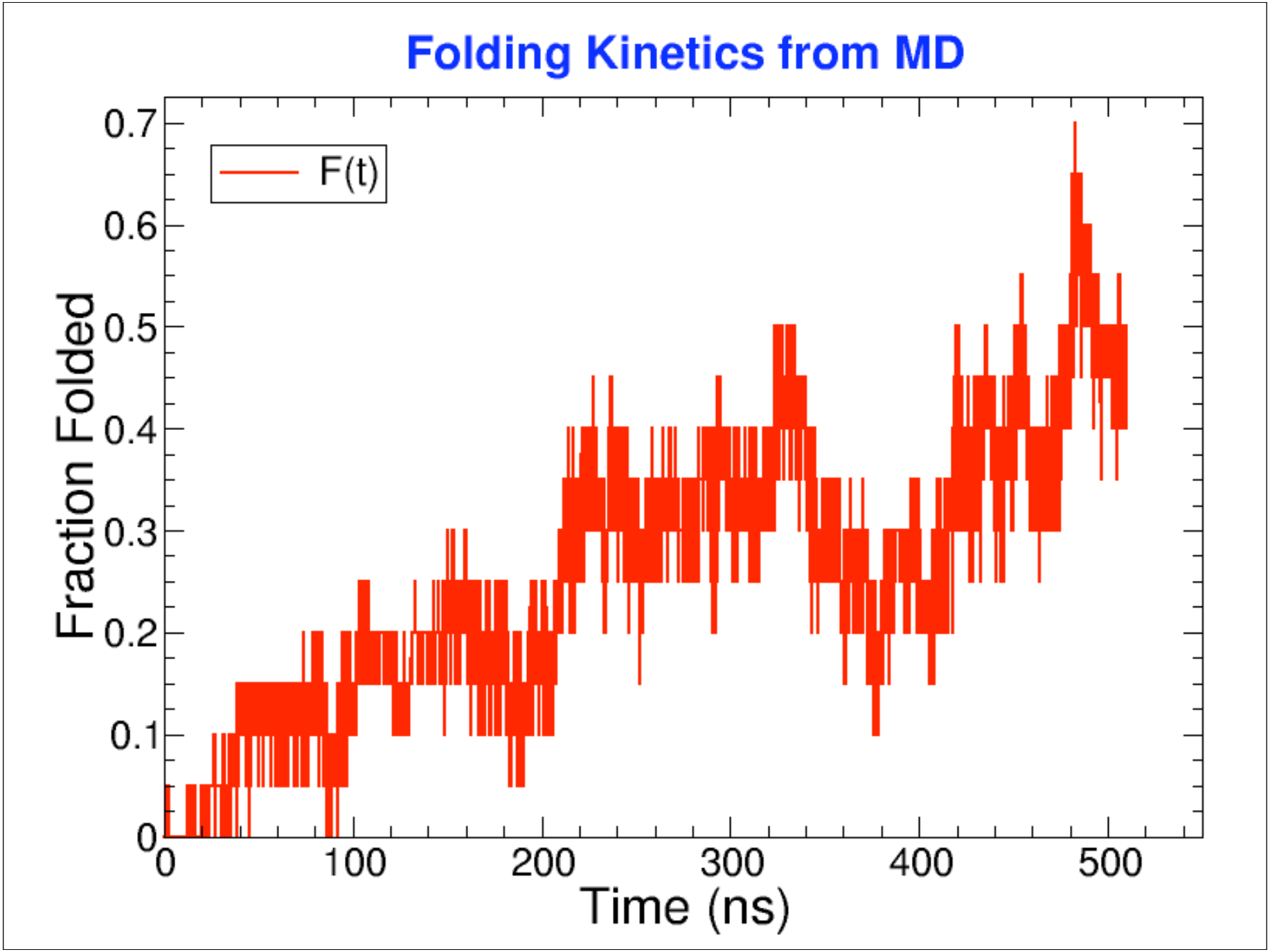
Folding kinetics from MD

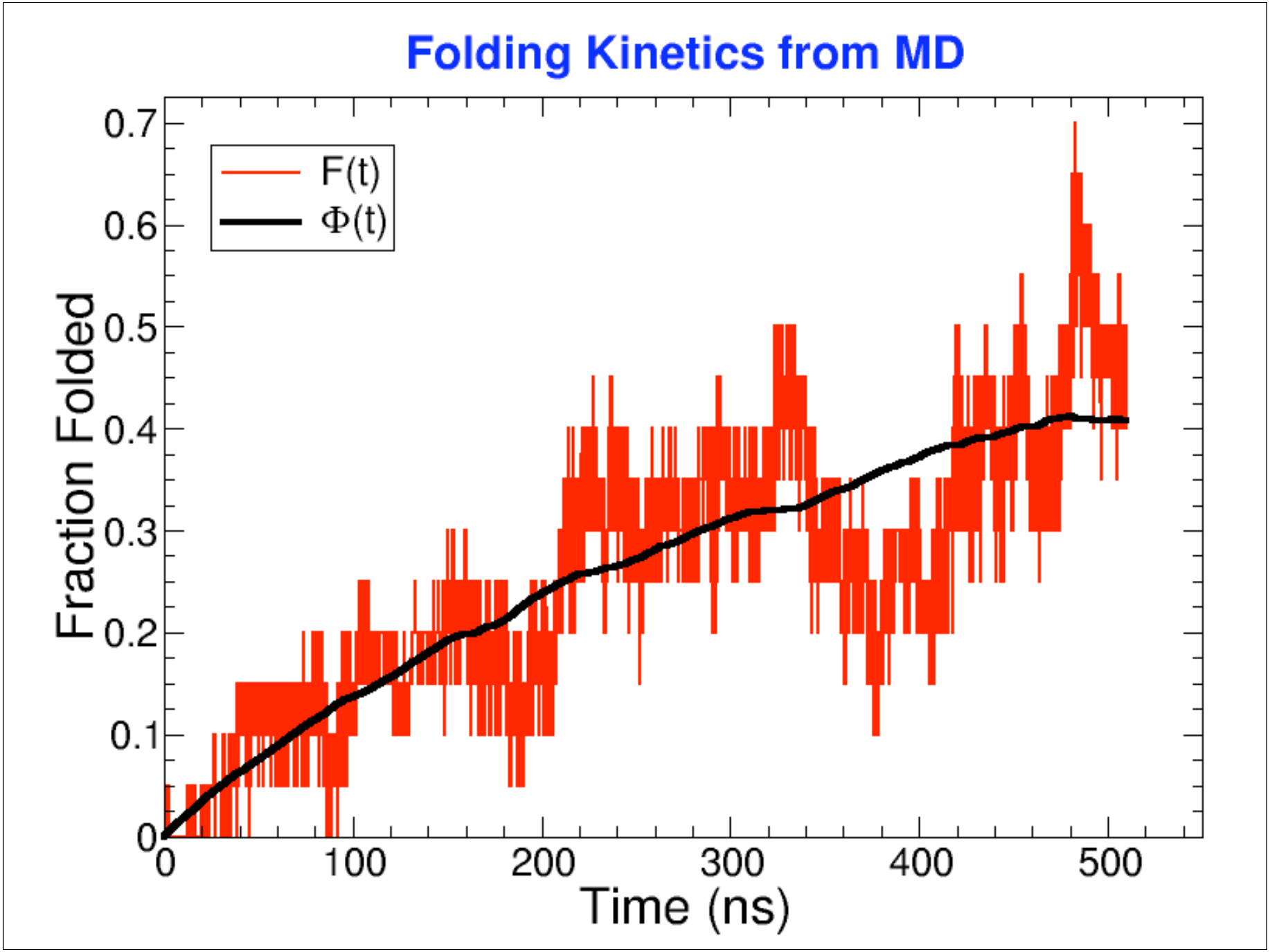
Now make $\beta = 1/k_B T$ time dependent:

$$\frac{dF_m(t)}{dt} = A_f \mathrm{e}^{-\beta_m(t) E_A^f} U_m(t) - A_u \mathrm{e}^{-\beta_m(t) E_A^u} F_m(t)$$

Folding kinetics from MD

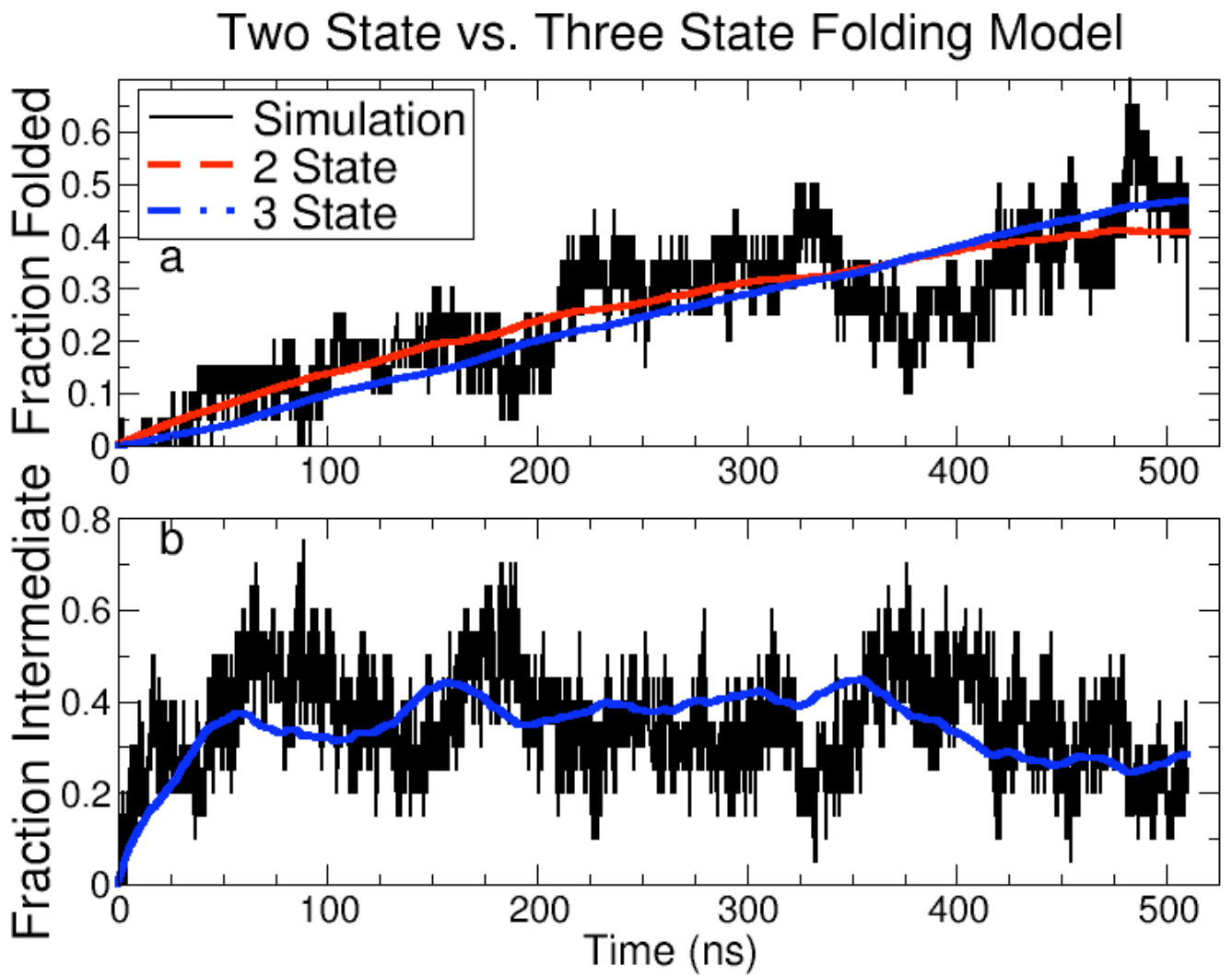
- Given a population of proteins in different states we predict the change in fraction folded by numerically integrating dF/dt averaged over many trajectories.
- From a numerical fit $\Phi(t)$ of the computed integral to $F(t)$ we can compute the four constants governing the kinetics: E_f, E_u, A_f, A_u
- The four constants do not depend on temperature


Folding Kinetics from MD
F(t)
Fraction Folded
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
100
200
300
400
500
Time (ns)


Folding Kinetics from MD
F(t)
Φ(t)
Fraction Folded
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
100
200
300
400
500
Time (ns)

Folding kinetics from MD @ 300 K

	Fold	Unfold	
A	9.3e-5(0.1)	0.094(0.01)	(1/ps)
E_A	11.2(1)	30.7(1)	(kJ/mole)
	Simulation	Exper.	
τ_f	1.0(0.3)		(μs)
τ_u	2.6(0.4)		(μs)
ΔG	2.4(0.7)	1.1(0.7)	(kJ/mole)
ΔH	19.6(2)	27.1(1)	(kJ/mole)
TΔS	17.1(1)	26.0(1)	(kJ/mole)
T_m	340(9)	312(2)	(K)

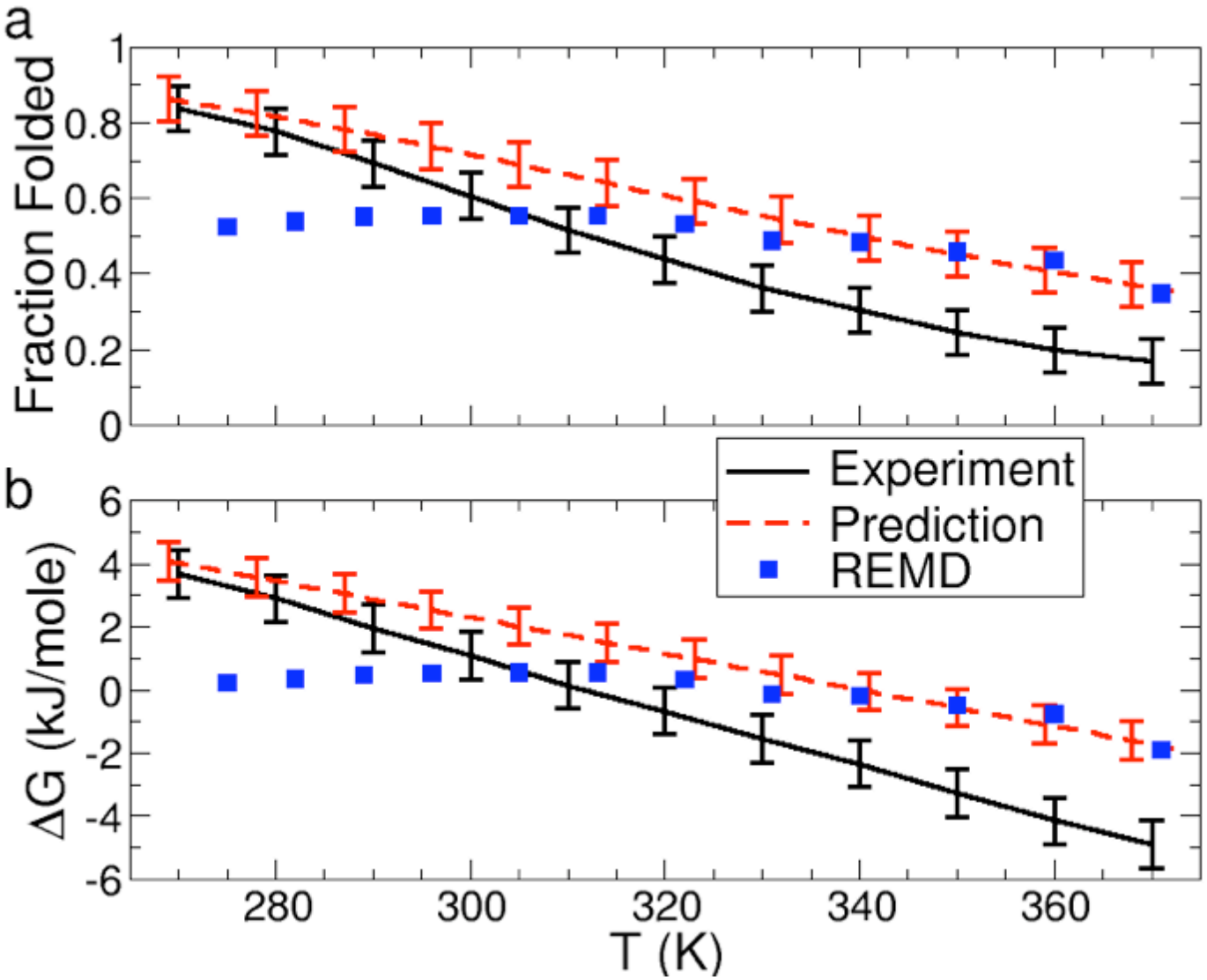

Two State vs. Three State Folding Model
Simulation
2 State
3 State
a
b
Fraction Folded
Fraction Intermediate
Time (ns)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
100
200
300
400
500

Folding kinetics from MD

- Based on the parameters E_f, E_u, A_f, A_u we can predict the folding equilibrium as a function of temperature

$$F(t=\infty) = \frac{k_u}{k_u + k_f}$$

Folding kinetics from MD



D. van der Spoel & M. Seibert, Phys. Rev. Lett. **96**, 238102, 2006

Chignolin Summary

- Chignolin's native state can be predicted to an accuracy of < 1.9 Å (all-atom) RMSD by *ab initio* molecular dynamics simulations.

- The native state can be identified on a Gibbs free energy landscape without direct reference to experimental data.

- Kinetics of folding can be predicted based on a heterogeneous ensemble of (RE)MD trajectories, giving information on longer time scales than what was actually simulated

- The temperature dependence of the folding/ unfolding equilibrium is reproduced quite well

Reduce Simulation Coupling to Improve Performance

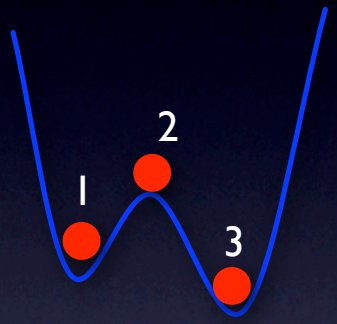
- Collect thermodynamic data instead of waiting for rare events (scales 100%)
- Weakly coupled simulations (10Mbit ethernet)
  - Replica-Exchange - Works great in GROMACS
- Non-Coupled simulations:
  - Distributed computing (dial-up/ADSL)

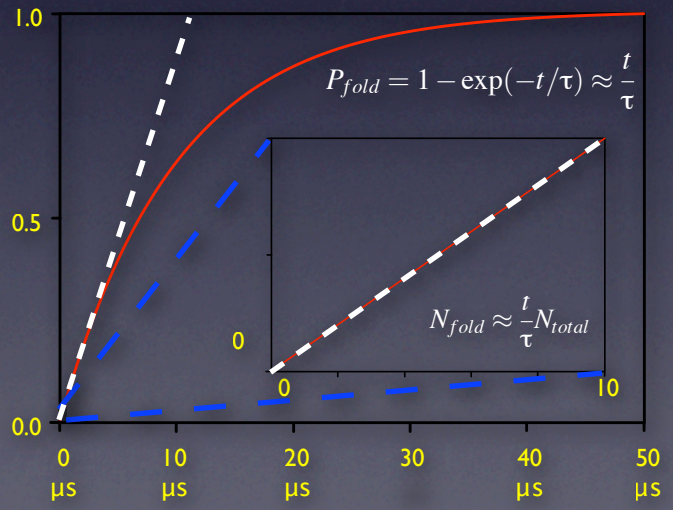
Distributed Computing

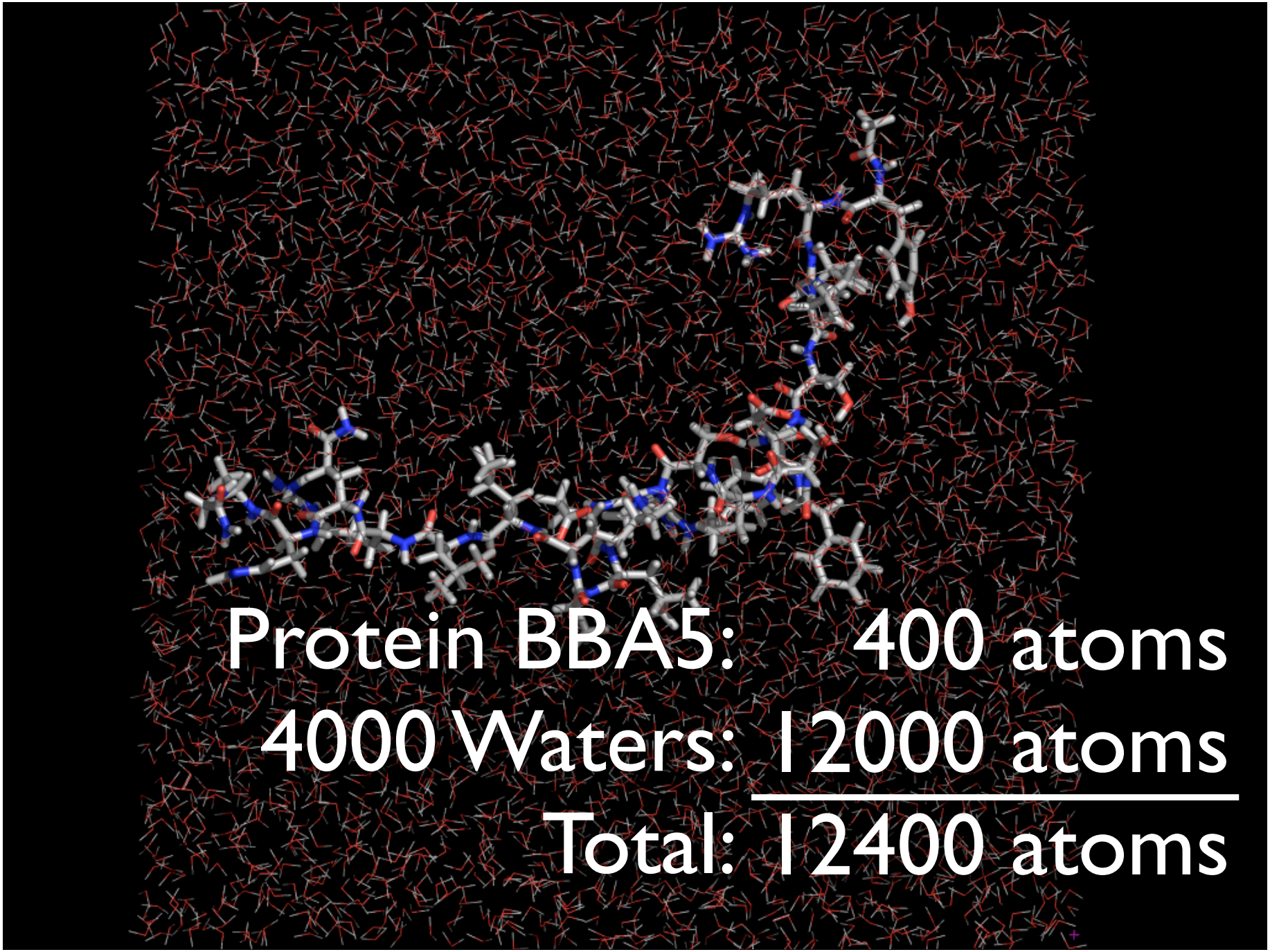
- Folding@Home

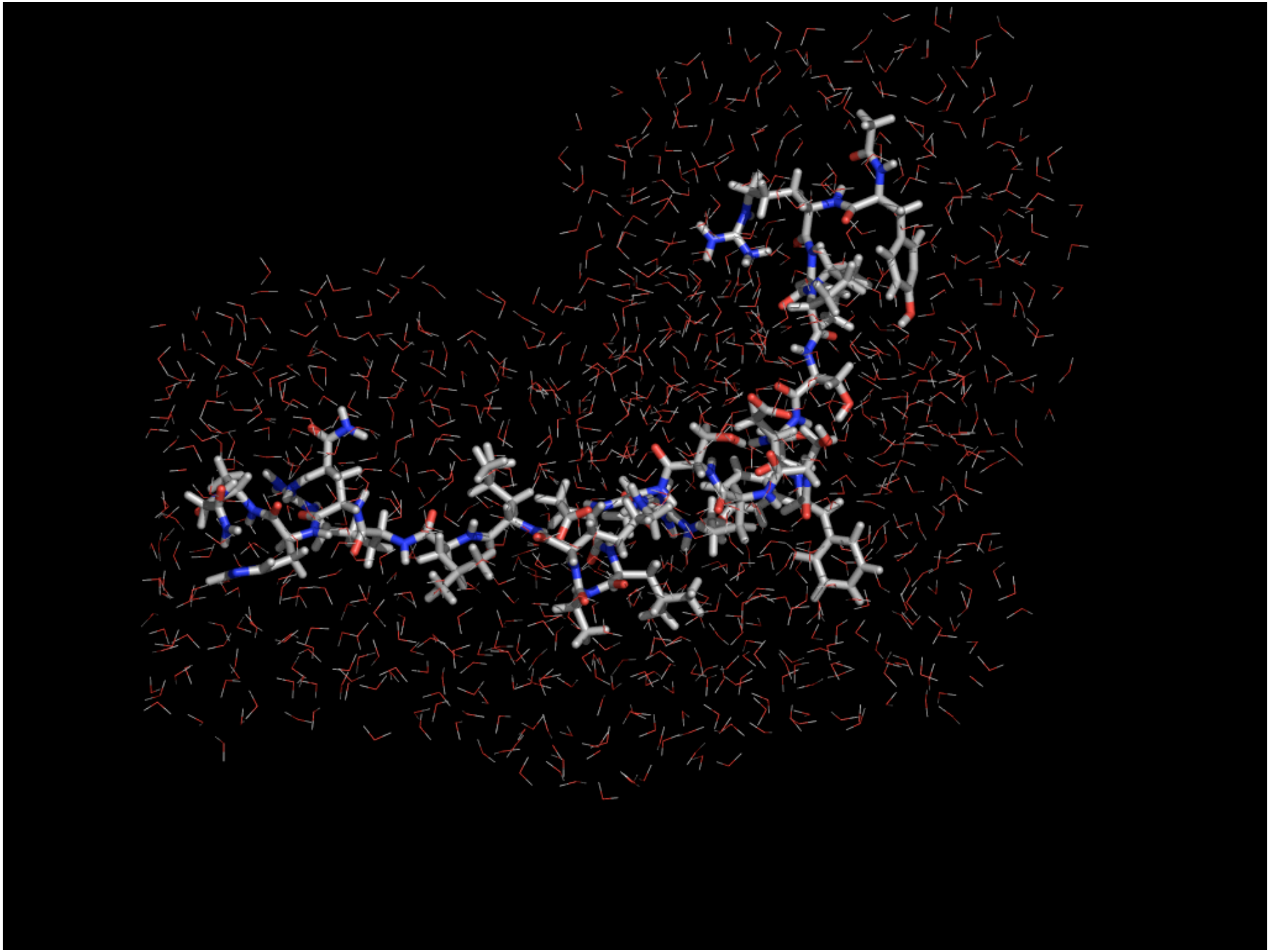
Distributed Computing Protein Folding

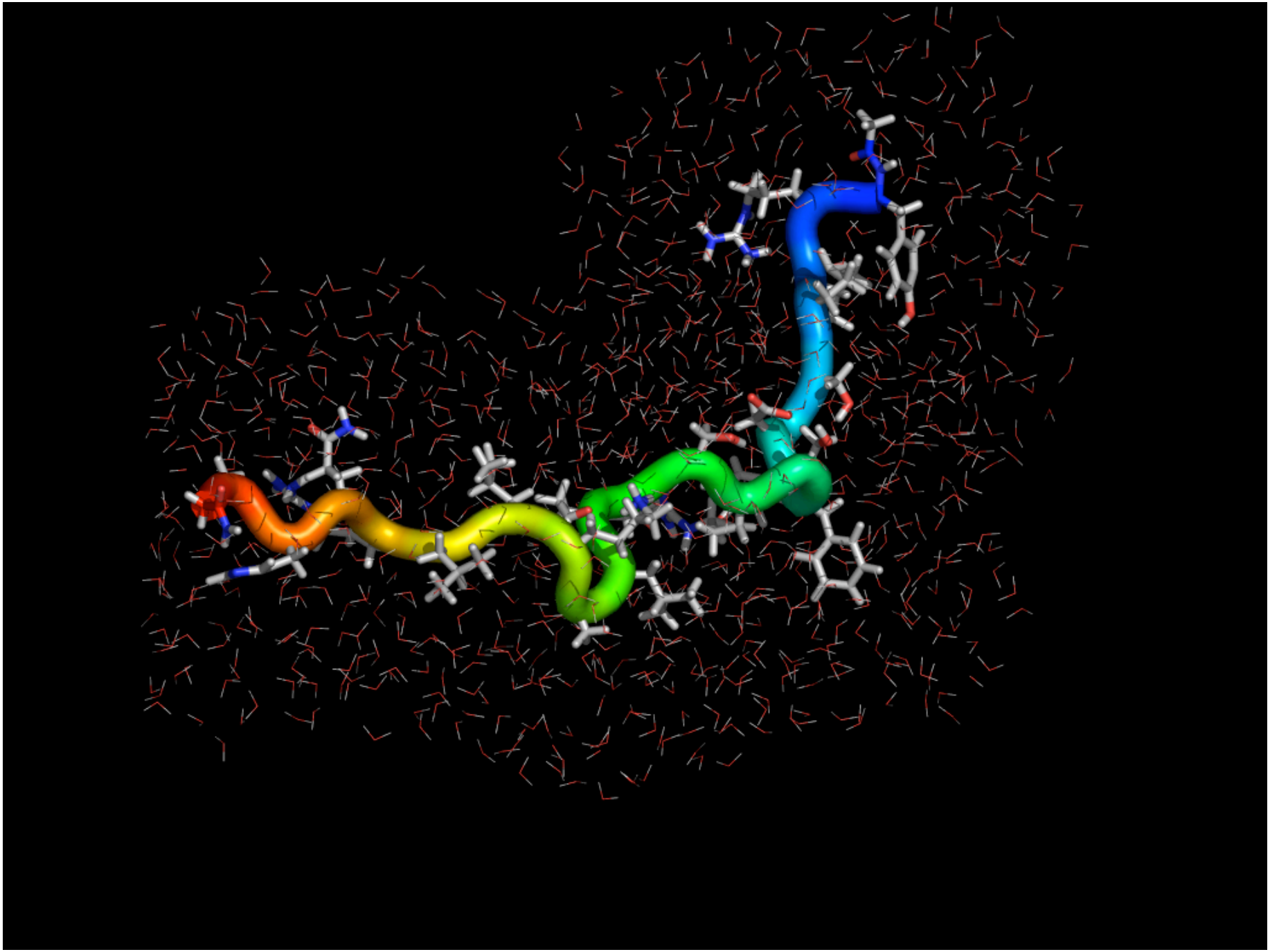
- Folding is approximately a 1st order transition
- BBA5: Folding time is ~10 μs
- Probability of folding in short simulation is small, but >0
- Perform 10,000 independent 10ns simulations instead of a single 100 μs one
- Run GROMACS as screensaver in Folding@Home

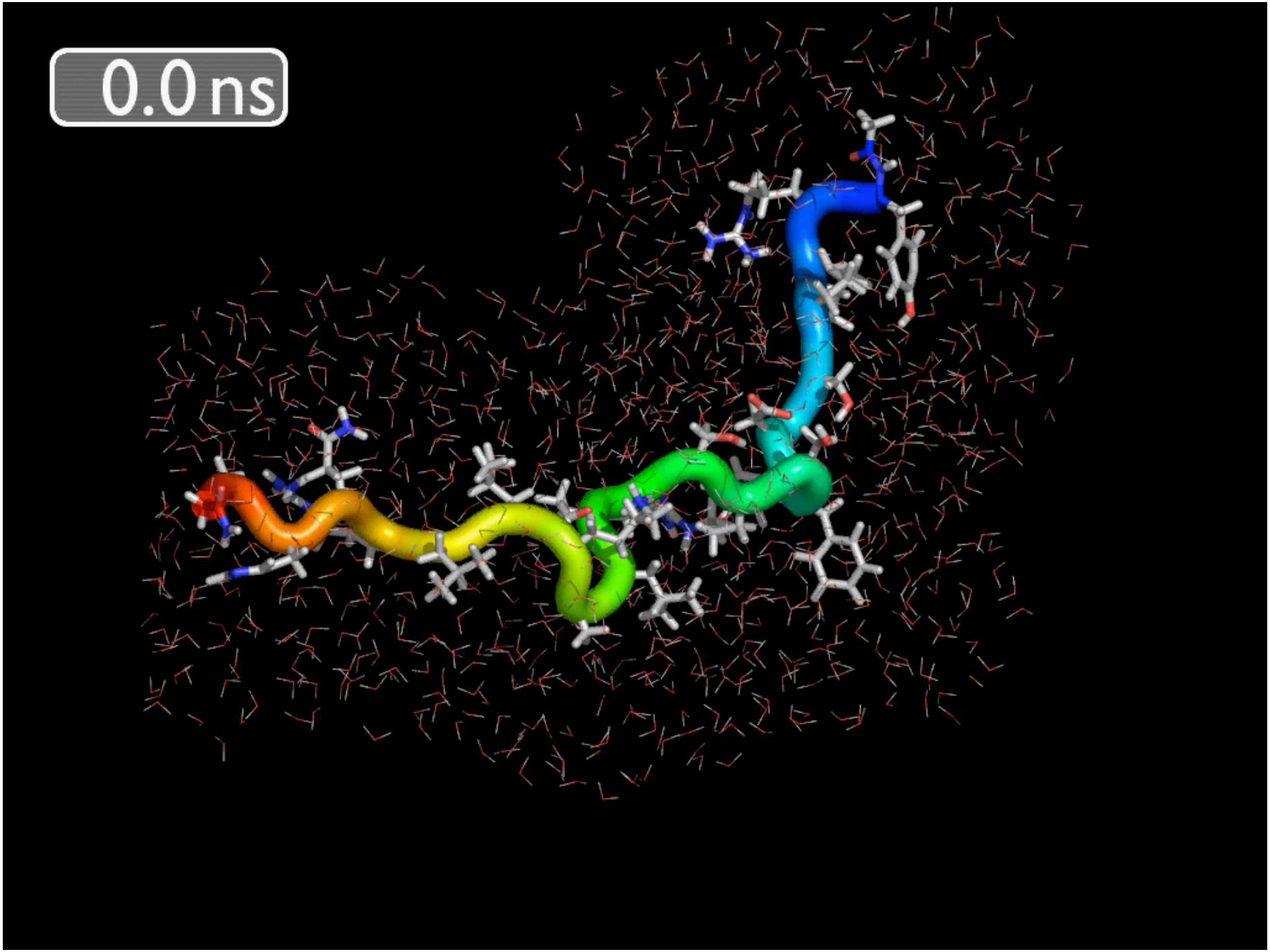




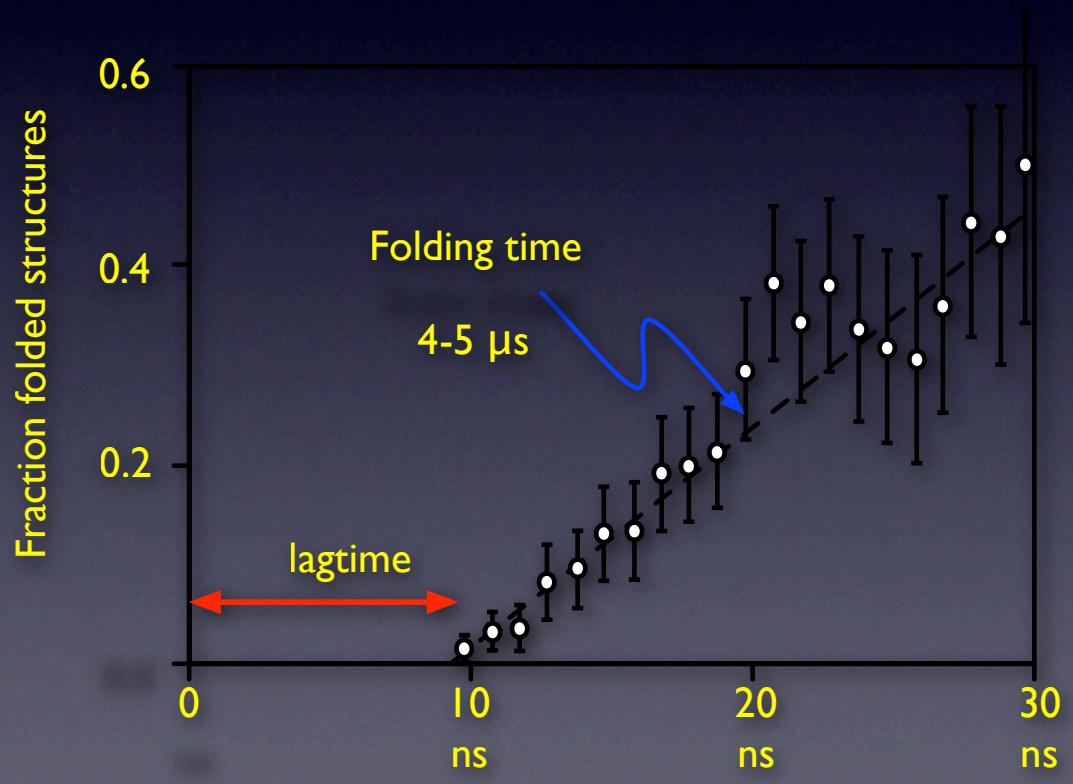

Protein BBA5: 400 atoms
4000 Waters: 12000 atoms
Total: 12400 atoms






0.0 ns

Fold Fraction over time



Summary - Speed

- Many time-saving techniques implemented in GROMACS 3.3
- Very good parallel scaling will be available in GROMACS 4.0
- REMD or other high level algorithms may be used to speed up convergence
- Algorithms exist that allow to make predictions on time scales (way) beyond the simulated ones

Outlook

- REMD calculations can be speeded up easily by running each replica on multiple nodes allowing longer trajectories and/or larger proteins. Implemented in GROMACS 4.0b
- Better force fields necessary to get the higher temperatures correct.
- Investigation of more complex folding events (with intermediates, pathway dependencies)

Acknowledgements

Protein Folding Properties from MD Simulations

- **UPPSALA:** Alexandra Patriksson, Marvin Seibert, Michiel van Lun, Filipe Maia, Nicusor Timneanu
- **TARTU:** Csaba Hetényi
- **GRONINGEN**: Paul van Maaren

THE GROMACS TEAM

- **STOCKHOLM:** Erik Lindahl
- **MAINZ:** Berk Hess