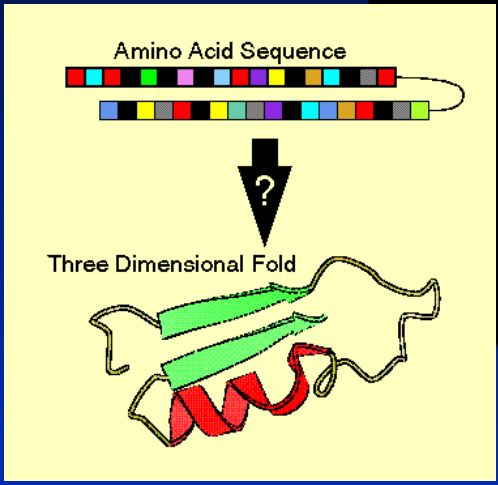
Center for Computational Biology and Bioinformatics

Protein folding problem:

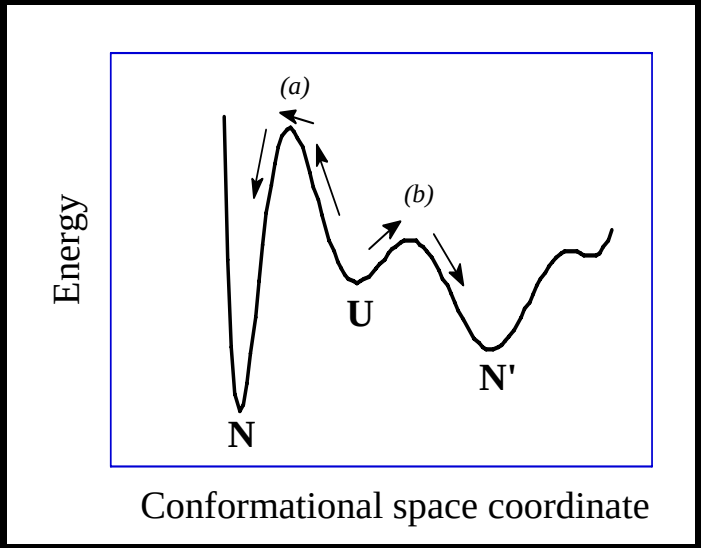
"Predicting 3-dimensional structure from amino acid sequence"



- A unique folded structure (native conformation, native fold) is assumed by a given sequence, although infinitely many conformations can be accessed.
- Which structure? (Protein folding problem)
- How? (Folding **kinetics**)

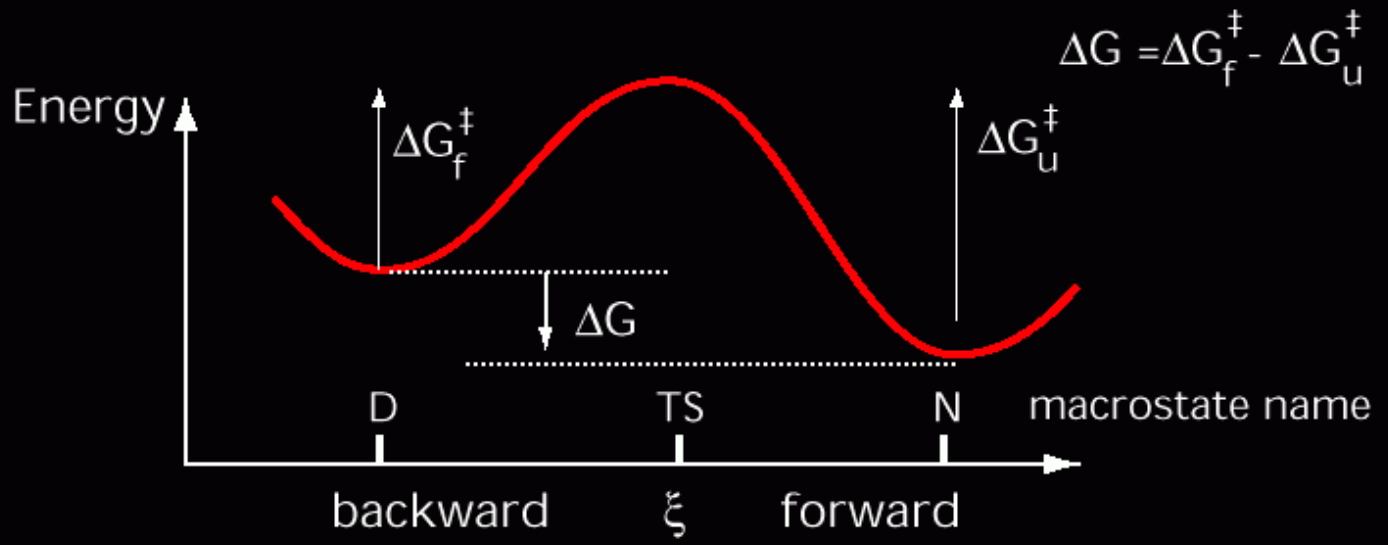
Basic postulate:

Thermodynamic equilibrium → Global energy minimum



Probability of conformation i ~ exp ($-E_i/RT$)

The Classical Transition State



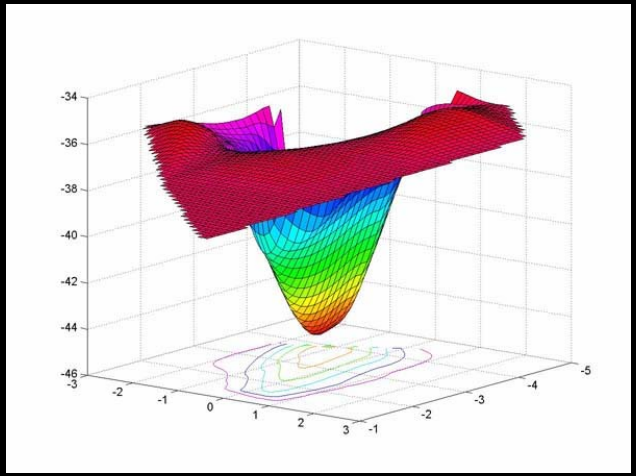
localized ensembles	$(C_1\ C_2\ C_3)$.....	$(C_i\ C_{i+1})$	(C_N)	microstates

- Macrostates are localized ensembles of microstates.
- States are in series and don't overlap.
- Single reaction coordinate. Forward & backward directions.

Folding is fast (μs-ms)

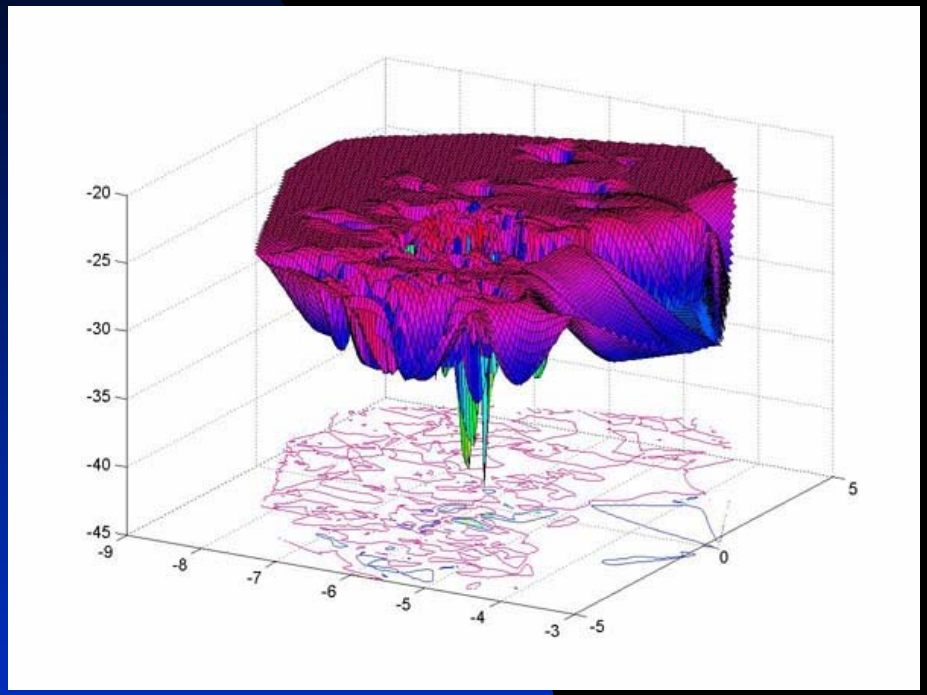
- Explanation: Multiple pathways

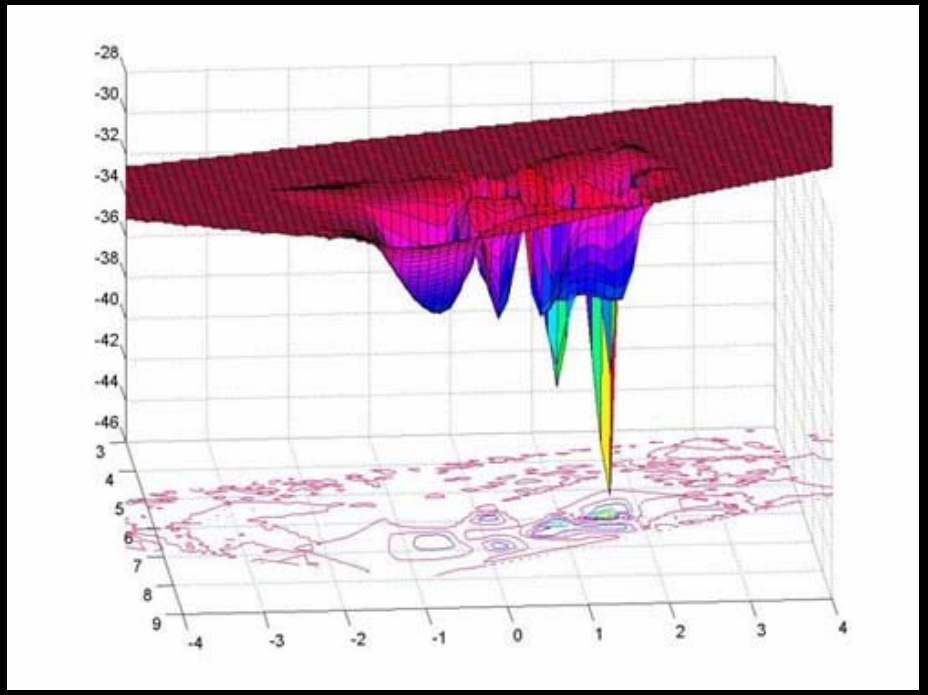
Representation of the energy surface as a funnel, rather than an energy curve as a function of reaction coordinate



Center for Computational Biology and Bioinformatics

Folding/unfolding energy landscapes





Reference

B. Ozkan, K.A. Dill & I. Bahar, Protein Sci. 11, 1958-1970, 2002.

Protein structure prediction

Three computational methods:

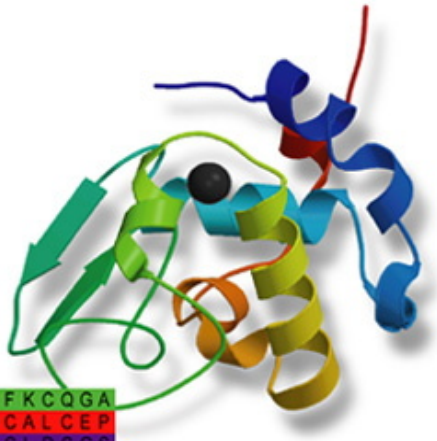
- Homology modeling
- Threading
- *Ab initio* simulations

CASP (Critical Assessment of Structure Prediction)

Homology/comparative modeling

Modeller

Program for Comparative Protein
Structure Modelling by Satisfaction
of Spatial Restraints



MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (2, 3), and can perform many additional tasks, including de novo modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc. MODELLER is written in Fortran 90 and runs on the Pentium PC's (Linux and Win XP), Apple Macintosh (OS X) and workstations from Silicon Graphics (IRIX), Sun (Solaris), IBM (AIX), and DEC Alpha (OSF/1).

http://guitar.rockefeller.edu/modeller/modeller.html (A. Sali)

SWISS-MODEL

An Automated Comparative Protein Modelling Server accessible via the ExPASy (Expert Protein Analysis System) web server (by Peitsch et al.)

STEPS:

1. Search for suitable templates (from ExNRL-3D , using BLAST)
2. Check sequence identity with target
(SIM will select all templates with sequence identities above 25% and N> 20)
3. Create ProModII jobs
4. Generate models (ProModII) using known 3-d templates
5. Energy minimization with Gromos96

http://swissmodel.expasy.org/SWISS-MODEL.html

Three levels of sequence similarity

- Above 30 % sequence identity

- The region 20-30 % ***Twilight Zone***

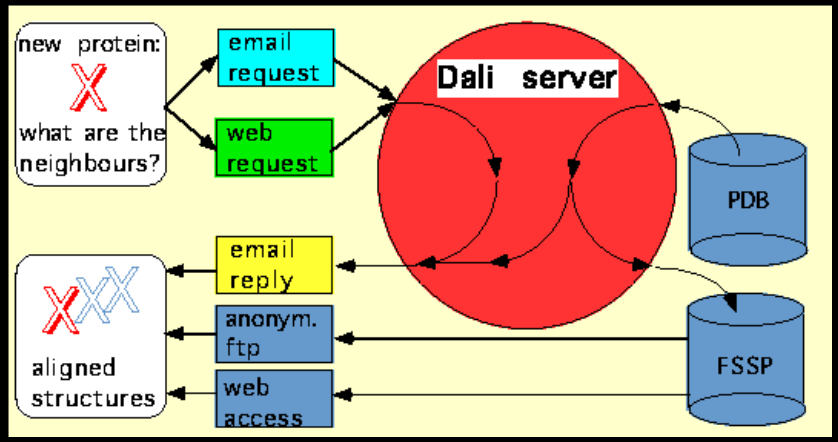
- Below 20 % ***Midnight zone***

Structural Homology

- Dali Server (Sander-Holm)

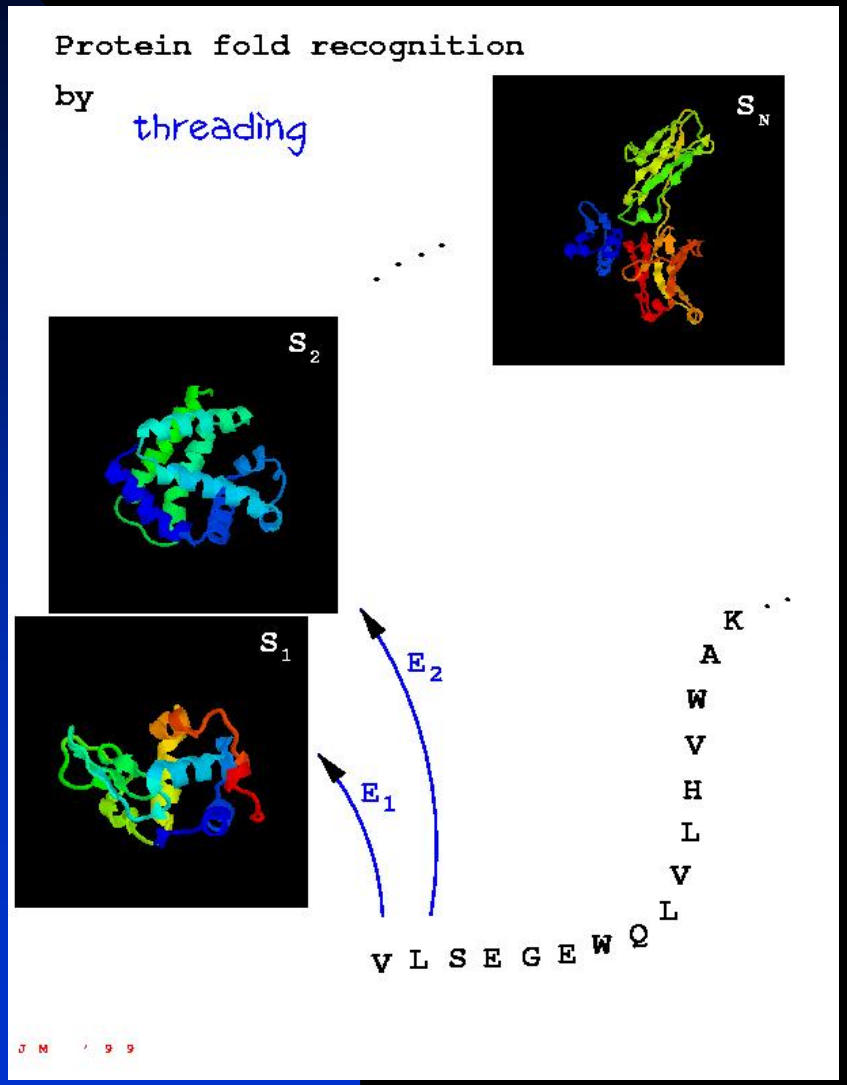
http://www2.ebi.ac.uk/dali/

The Dali server is a network service for **comparing** protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the PDB, with or without sequence constraints



L. Holm and C. Sander (1996) Mapping the protein universe. Science 273:595-602.

Threading (Fold recognition)



Loopp (Elber)
Threader (Jones)

Ab initio simulations

- Protarch (Scheraga's group)
- Rosetta (Baker's lab)
- Touchstone (Skolnick)

Molecular dynamics (MD) simulations

- A deterministic method based on the solution of Newton's equation of motion

$$\mathbf{F}_i = \mathbf{m}_i \mathbf{a}_i$$

for the ith particle; the acceleration at each step is calculated from the negative gradient of the overall potential, using

$$F_i = - \text{grad } V_i - = - \nabla \mathbf{v}_i$$

$V_i = \Sigma_k$(energies of interactions between i and all other residues k located within a cutoff distance of R_c from i)

∇V_i = Gradient of potential?

Derivative of V with respect to position vector $r_i = (x_i, y_i, z_i)^T$

$$a_{xi} \sim -\partial V/\partial x_i$$
$$a_{yi} \sim -\partial V/\partial y_i$$
$$a_{zi} \sim -\partial V/\partial z_i$$

Interaction potentials include;

Non-Bonded Interaction Potentials

- Electrostatic interactions of the form $E_{ik}(es) = q_i q_k / r_{ik}$
- Van de Waals interactions $E_{ij}(vdW) = - a_{ik}/r_{ik}^{6} + b_{ik}/r_{ik}^{12}$

Bonded Interaction Potentials

- Bond stretching $E_i(bs) = (k_{bs}/2)\ (l_i - l_i^{0})^2$
- Bond angle distortion $E_i(bad) = (k_\theta/2)\ (\theta_i - \theta_i^{0})^2$
- Bond torsional rotation $E_i(tor) = (k_\phi/2)\ f(\cos\phi_i)$

Example 1: gradient of vdW interaction with residue k

- $\mathbf{E_{ik}(vdW) = - a_{ik}/r_{ik}^{6} + b_{ik}/r_{ik}^{12}}$
- $r_{ik} = r_k - r_i$
  - $x_{ik} = x_k - x_i$
  - $y_{ik} = y_k - y_i$
  - $z_{ik} = z_k - z_i$
  - $r_{ik} = [\ (x_k - x_i)^2 + (y_k - y_i)^2 + (z_k - z_i)^2\ ]^{1/2}$
- $\partial V/\partial x_i = \partial\ [- a_{ik}/r_{ik}^{6} + b_{ik}/r_{ik}^{12}] / \partial x_i$

  where $r_{ik}^{6} = [\ (x_k - x_i)^2 + (y_k - y_i)^2 + (z_k - z_i)^2\ ]^3$

Example 2: gradient of bond stretching potential with respect to r_i

- **$E_i(bs) = (k_{bs}/2)\,(l_i - l_i^0)^2$**
- $l_i = r_{i+1} - r_i$
  - $l_{ix} = x_{i+1} - x_i$
  - $l_{iy} = y_{i+1} - y_i$
  - $l_{iz} = z_{i+1} - z_i$
  - $l_i = [\,(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2\,]^{1/2}$
- $\partial\, E_i(bs)/\partial x_i = -\, m_i a_{ix}(bs)$ *(induced by deforming bond l_i)*

$= (k_{bs}/2)\,\partial\,\{[\,(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2\,]^{1/2} - l_i^0\}^2/\partial x_i$
$= k_{bs}\,(l_i - l_i^0)\,\partial\,\{[\,(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2\,]^{1/2} - l_i^0\}/\partial x_i$
$= k_{bs}\,(l_i - l_i^0)\,(1/2)\,(l_i^{-1})\,\partial\,(x_{i+1} - x_i)^2/\partial x_i = -\,k_{bs}\,(1 - l_i^0/l_i)\,(x_{i+1} - x_i)$

The Verlet algorithm

The most widely used method of integrating the equations of motion.

$$r(t+\delta t) = 2r(t)-r(t-\delta t)+ \delta t^2 a(t)$$

The velocities are eliminated by adding the Taylor expansions

$$r(t+\delta t) = r(t) + \delta t\, v(t) + (1/2)\, \delta t^2\, a(t)+ \ldots$$

$$r(t-\delta t) = r(t) - \delta t\, v(t) + (1/2)\, \delta t^2\, a(t)-$$

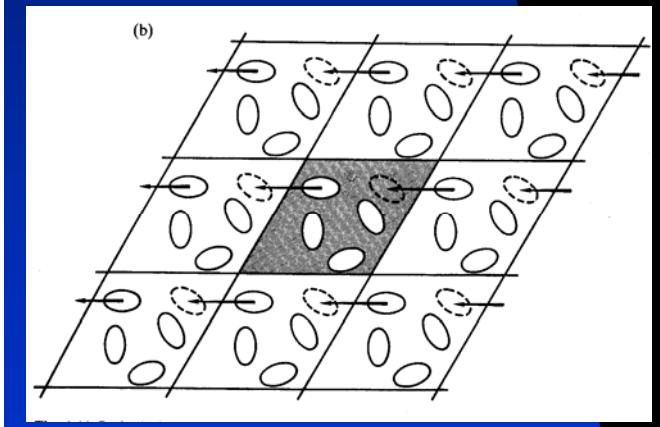
The velocities may be obtained from $v(t)= [r(t+\delta t)-r(t-\delta t)]/2\delta t$

Initial velocities (v_i)

$$v_i = (m_i/2\pi kT)^{1/2} \exp(- m_i v_i^2/2kT)$$

(Boltzmann distribution at the given temperature)

Periodic boundary conditions



How to generate MD trajectories?

- Known initial conformation, i.e. $r_i(0)$ for all atom i
- Assign $v_i(0)$, based on Boltzmann distribution at given T
- Calculate $r_i(\delta t) = r_i(0) + \delta t\, v_i(0)$
- Using new $r_i(\delta t)$ evaluate the total potential V_i on atom I
- Calculate negative gradient of V_i to find $a_i(\delta t) = -\nabla V_i / m_i$
- Start Verlet algorithm using $r_i(0)$, $r_i(\delta t)$ and $a_i(\delta t)$
- Repeat for all atoms (including solvent, if any)
- Repeat the last three steps ~ 10^6 times (MD steps)

Limitations of MD simulations

- Full atomic representation → noise
- Empirical force fields → limited by the accuracy of the potentials
- Time steps constrained by the fastest motion (bond stretching of the order of femptoseconds
- Inefficient sampling of the complete space of conformations
- Limited to small proteins (100s of residues) and short times (subnanoseconds)

Need for Low Resolution Approaches

Coarse-grained Models

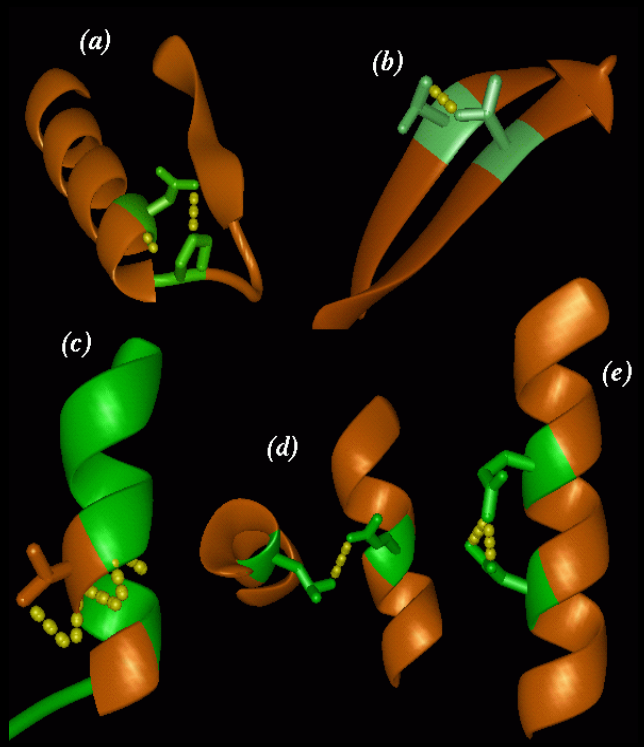
with

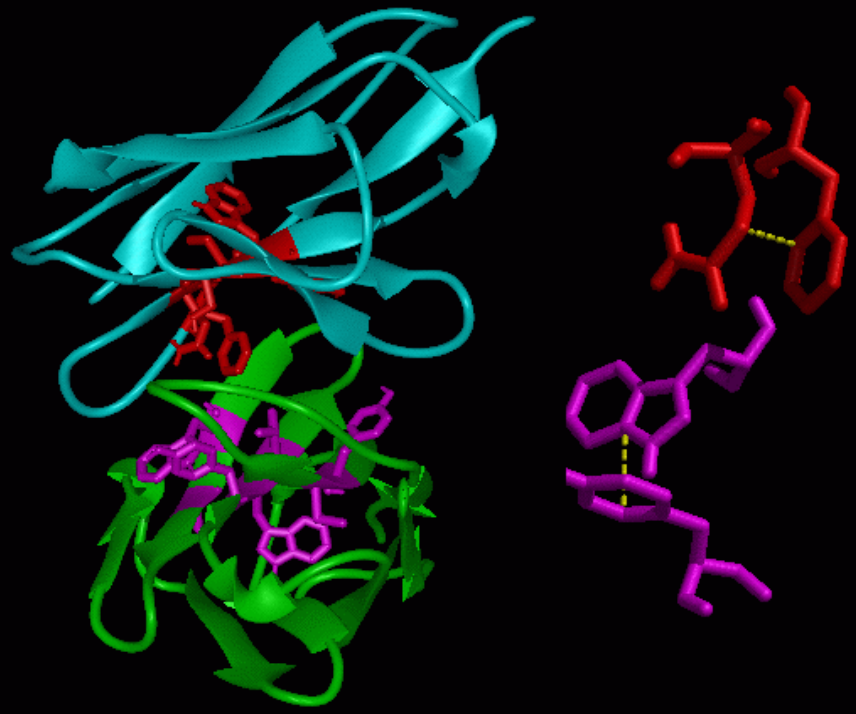
Empirical Force Fields

are the most tractable - if not the only possible –
computational tools for investigating large systems,
and complex biological processes

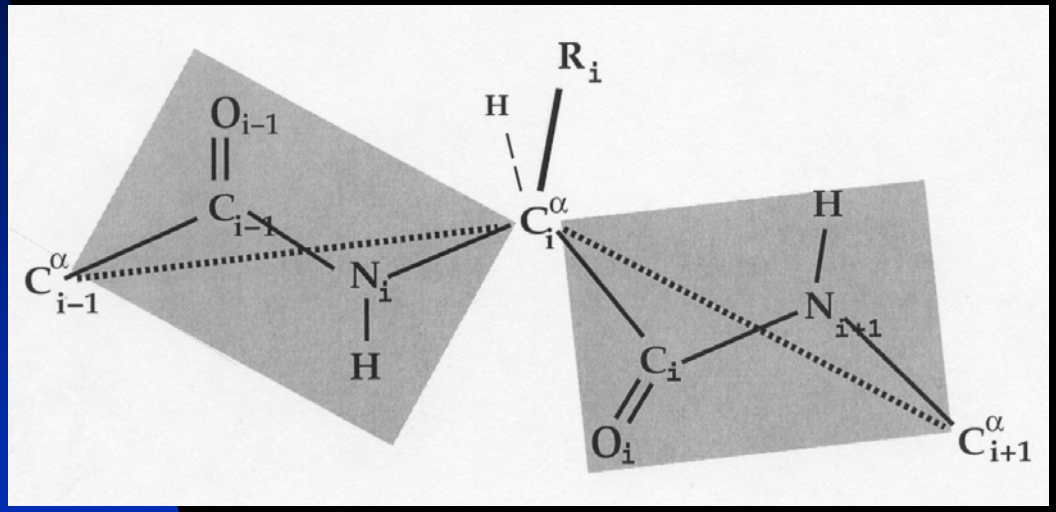
Knowledge-based studies

Exploiting PDB structures...





Virtual bond model



1. **Single interaction site per residue, identified by the α- or β-carbon**
2. **Need for empirical potentials for inter-residue interactions**

ANOTHER LOW-RESOLUTION MODEL

Two sites per residue: one at sidechain centroid, and the other and the peptide bond center (Scheraga and co-workers)

- The UNRES force field

- 1. Liwo, A., Oldziej, S., Pincus, M.R., Wawak, R.J., Rackovsky, S., Scheraga, H.A. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J. Comput. Chem., 1997, 18, 849-873.
- 2. Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S., Oldziej, S. Scheraga, H.A. A united-residue force field for off-lattice protein-structure simulations. II: Parameterization of local interactions and determination of the weights of energy terms by Z-score optimization. J. Comput. Chem., 1997, 18, 874-887.
- 3. Liwo, A., Kazmierkiewicz, R., Czaplewski, C., Groth, M., Oldziej, S., Wawak, R.J., Rackovsky, S., Pincus, M.R., Scheraga, H.A. United-residue force field for off-lattice protein-structure simulations; III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. J. Comput. Chem. 1998, 19, 259-276.

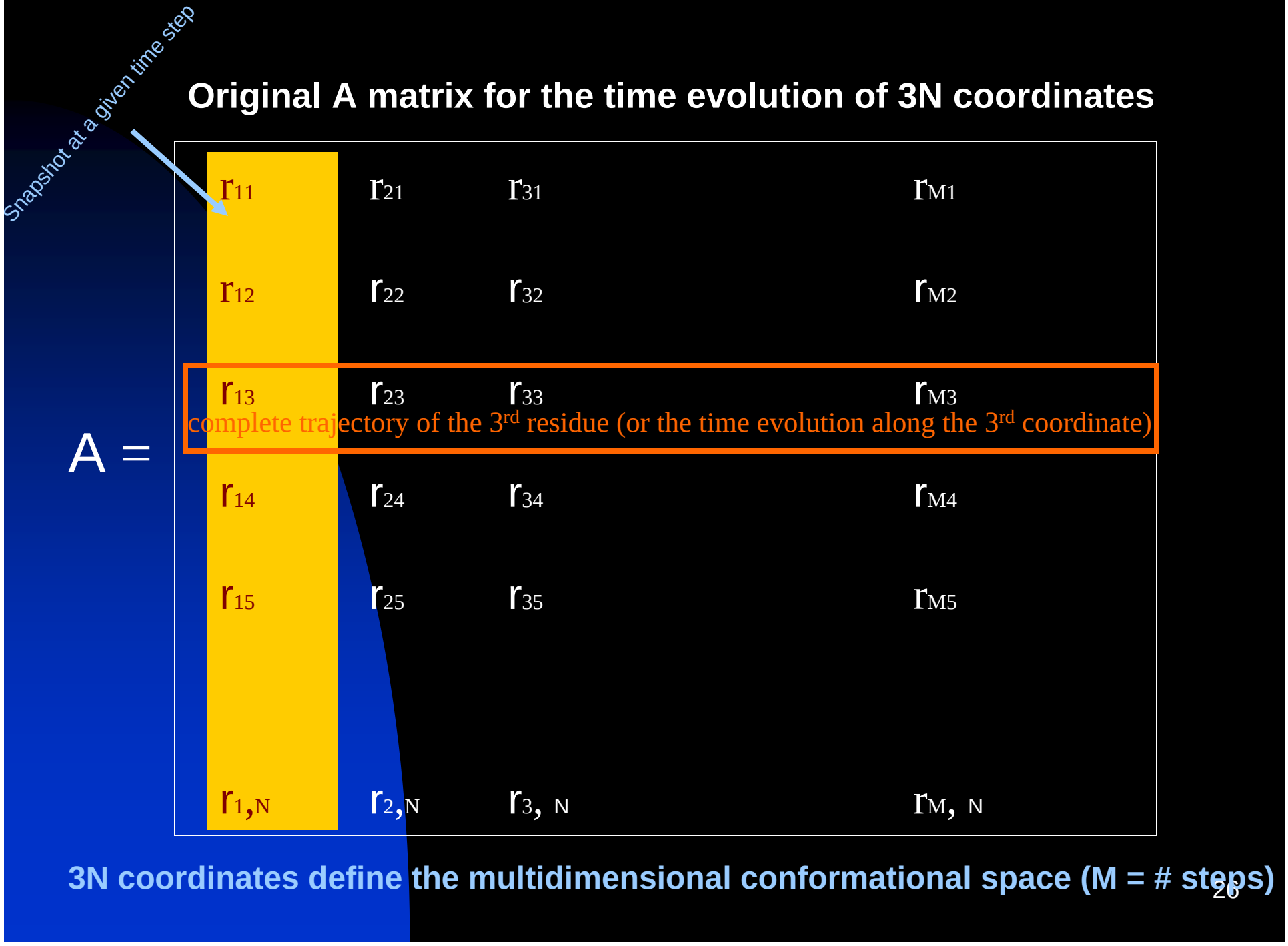
Aim: to understand the long-time dynamics, to remove the ‘uninteresting’ fast modes

Method: to map the trajectory onto a new multidimensional space, the axes of which refer to motions along principal coordinates

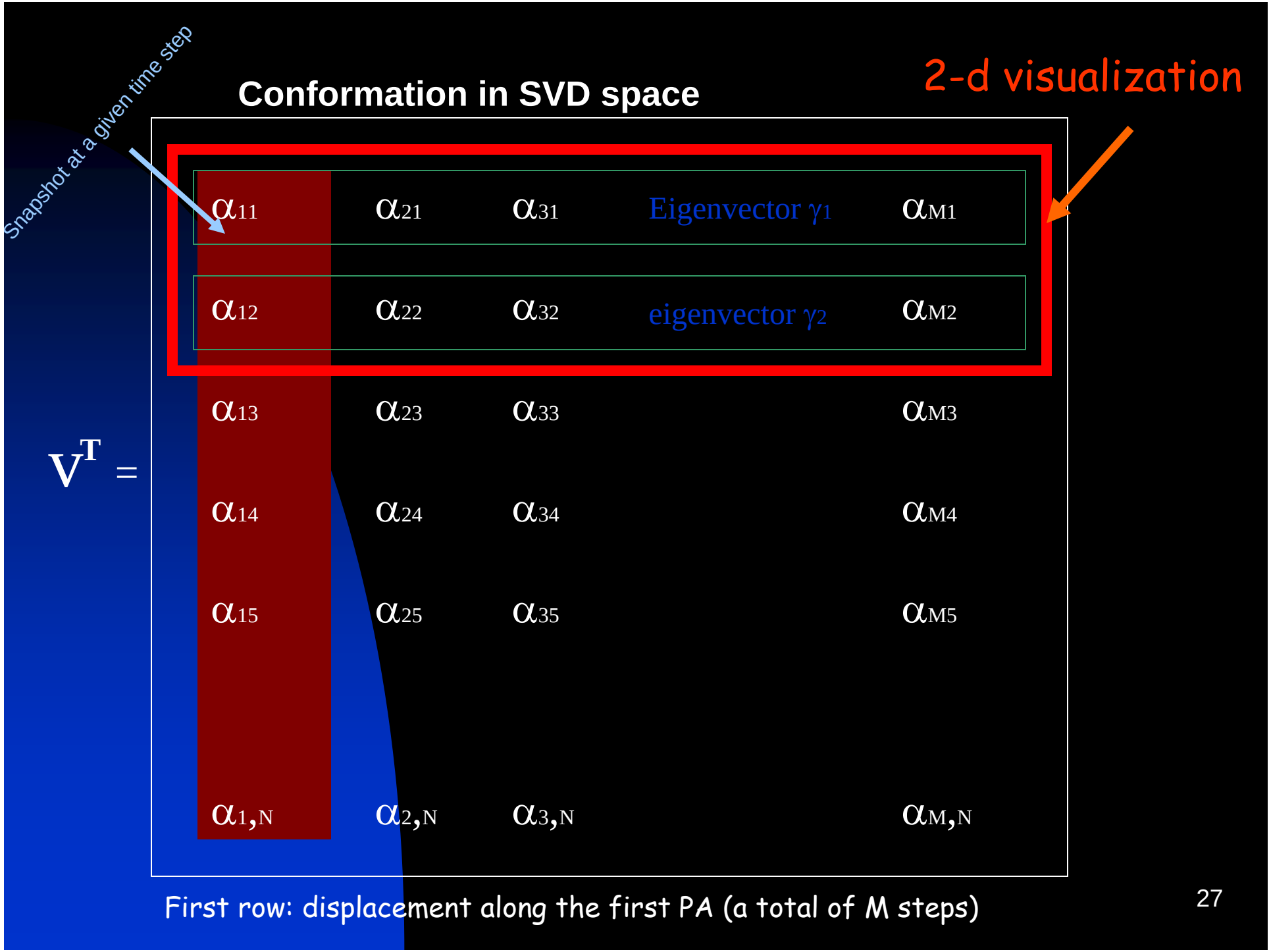
Frame transformation: From the 3N-dimensional space defining ‘conformations’ in Cartesian coordinates to the 3N-6 dimensional space of conformations in collective coordinates

Excellent review: Kitao & Go, Curr Opin Struct Biol 9, 164, 1999.

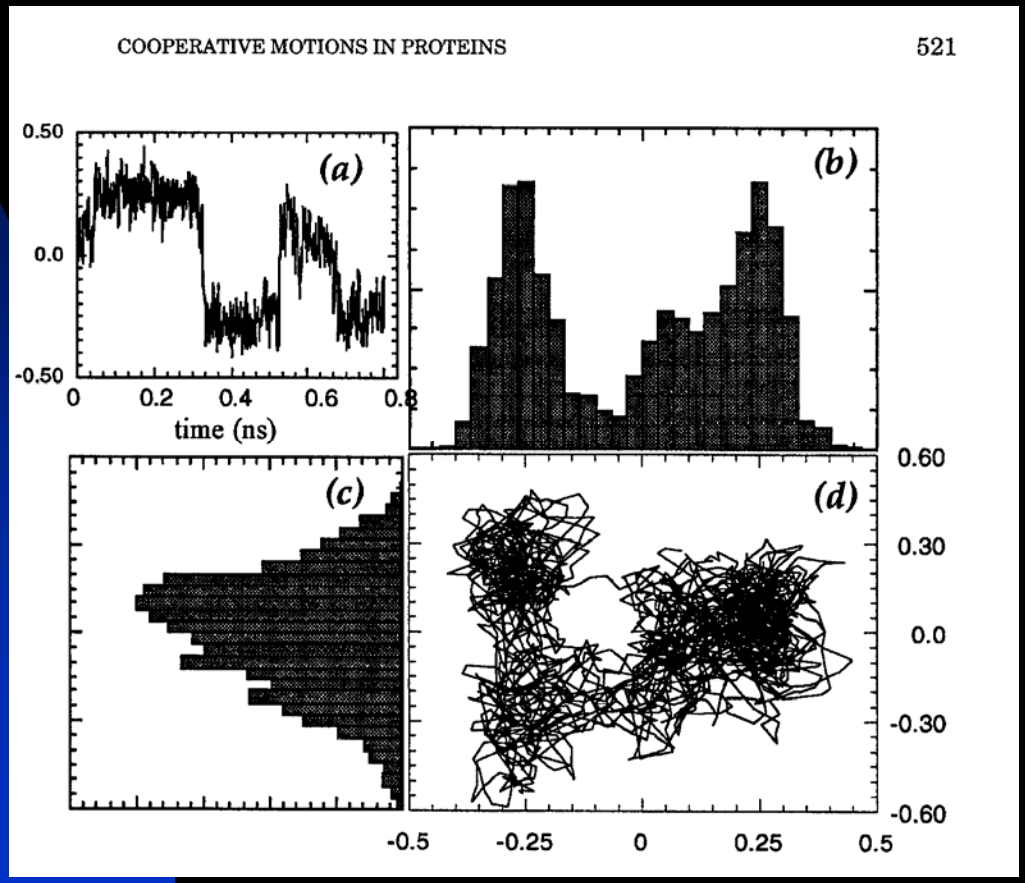
Original A matrix for the time evolution of 3N coordinates



3N coordinates define the multidimensional conformational space (M = # steps)


Snapshot at a given time step
Conformation in SVD space
2-d visualization
$V^T =$
α_{11} α_{21} α_{31} Eigenvector γ_1 α_{M1}
α_{12} α_{22} α_{32} eigenvector γ_2 α_{M2}
α_{13} α_{23} α_{33} α_{M3}
α_{14} α_{24} α_{34} α_{M4}
α_{15} α_{25} α_{35} α_{M5}
$\alpha_{1,N}$ $\alpha_{2,N}$ $\alpha_{3,N}$ $\alpha_{M,N}$
First row: displacement along the first PA (a total of M steps)

Projection of the motion onto the space of the two first principal axes



Comparison with essential modes from MD

Doruker, Atilgan & Bahar, *Proteins* 40, 512, 2000

- What is the **optimal** (realistic, but computationally efficient) model **for a given scale** (length and time) of representation?

- Which level of **details** is needed for representing **global** (collective) motions?

- How much **specificity** we need for modeling **large** scale systems and/or motions?

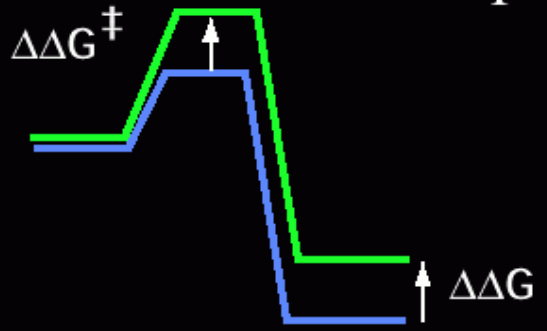
- What should be the **minimal** **ingredients** of a **simplified** (reductionist) model?

Φ Value Analysis *

$\Delta G = -RT \ln K$

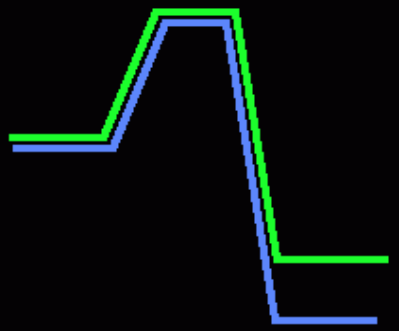
$\Delta G^{\ddagger} = -RT \ln k_f$

$$\Phi = \frac{\Delta\Delta G^{\ddagger}}{\Delta\Delta G}$$



$\Phi = 1$

At mutation site:
TS has Native-like structure



$\Phi = 0$

TS has Denatured-like structure

* A Fersht, Structure and Mechanism in Protein Science. Freeman (1999)

Protein folding kinetics examined by a Go-like model

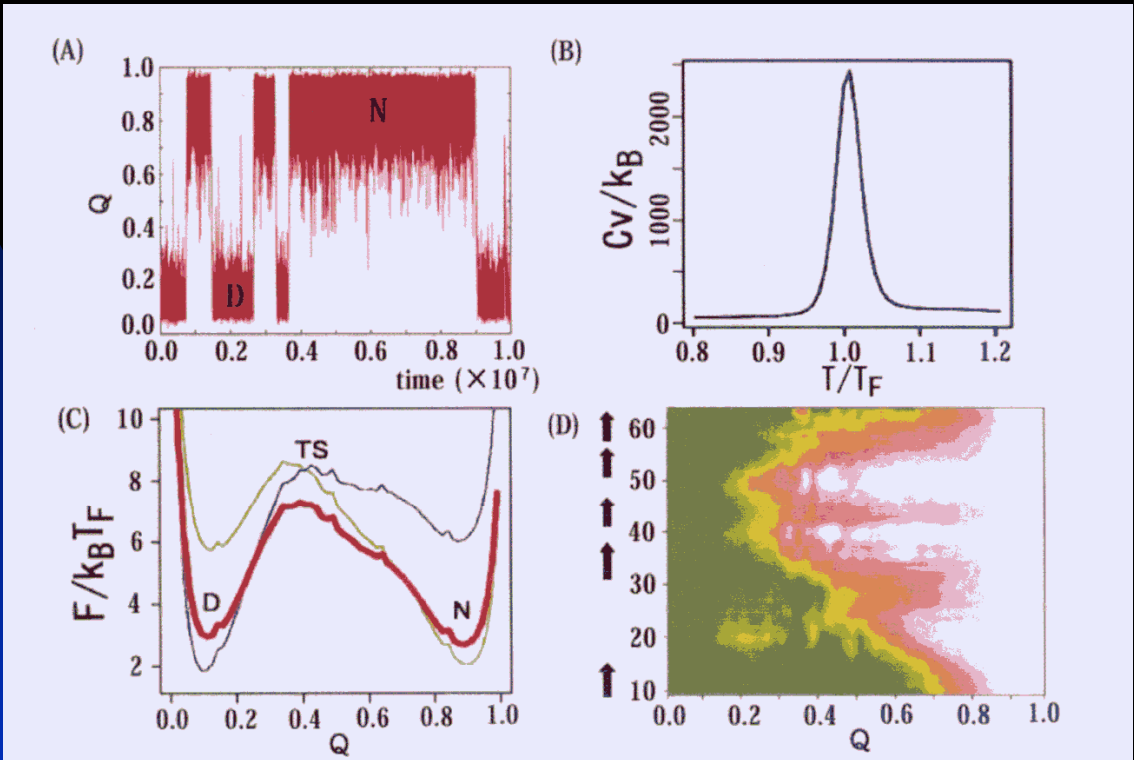
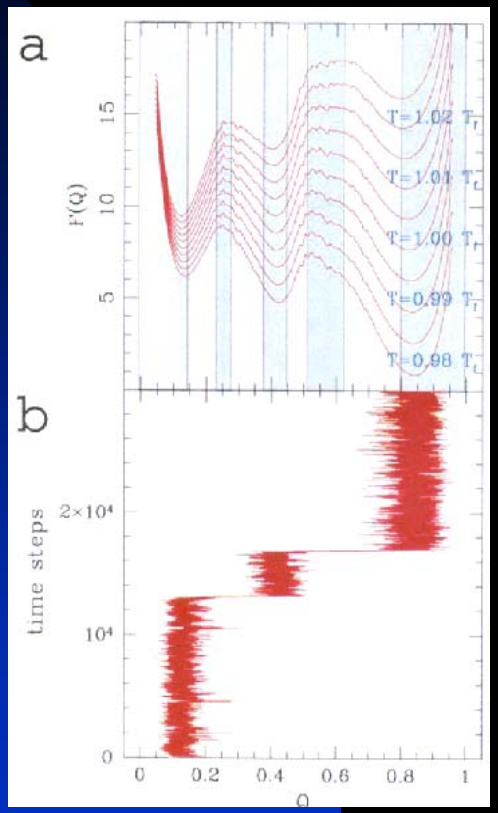


Figure 1. Example of folding time course and statistical mechanical analysis for the case of src SH3 domain. (a) The native-ness Q-value as a function of time t near the folding transition temperature T_F. This shows apparent two-state transition between native ($Q \sim 0.9$) and denatured ($Q \sim 0.15$) states. (b) C_v as a function of temperature T, which exhibits a clear peak at the folding transition temperature T_F. (c) The free energy profiles $F(Q)$ near the folding transition temperature T_F, $T = 0.96T_F$ (green), $T = T_F$ (the thick red curve), and $T = 1.04T_F$ (blue). The denatured and native states are separated by a free energy barrier around $Q \sim 0.4$. (d) The site-resolved folding pathways $q_i(Q)$ (see the text for the explicit definition) are plotted along the reaction coordinate Q. $q_i(Q)$ is unity (zero) when the local environment of site i is native-like (denatured-like) and the value of $q_i(Q)$ is represented by the color; green, yellow, and white correspond to 0, 0.5, and 1, respectively. Positions of β-strands are illustrated as arrows in the left side.

Koga, N. & Takada, S. J Mol. Biol. 2001, 313, 171-180

Topological and Energetic Factors: What determines the transition state ensemble, and folding intermediates?



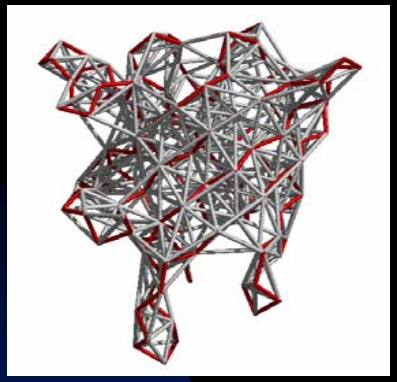
Simulations with Go-like potential

$$E(\Gamma,\ \Gamma_0) = \sum_{\text{bonds}} K_r(r - r_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedral}} K_\phi^{(n)}[1 + \cos(n \times (\phi - \phi_0))] + \sum_{i<j-3} \left\{ \varepsilon(i,j) \left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} \right] + \varepsilon_2(i,j)\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \right\} \quad \text{(A1)}$$

Applied to CI2, SH3 (2-state folders) and barnase, RNase H and CheY (have intermediates)

“ Topology plays a central role in determining folding mechanisms”

Clementi, C. Nyemeyer, H. & Onuchic, J. N. *J Mol. Biol* 2000, 298, 937.



Topology-based models

- **Near-native fluctuations**
  (springs acting on effective centroids, usually Cα atoms)

- Ben-Avraham (1993)
- Tirion (1996)
- Bahar et al. (1997)
- Hinsen (1998)
- Sanejouand, Tama (2000)
- Wriggers, Brooks (2001)
- Ma (2002)

Folding/unfolding
(folding ← loss of configurational entropy)

- Micheletti et al, *PRL* (1999)
- Cecconi et al. *Proteins* (2001)
- Go & Scheraga *Macromolecules* (1976)
- Galzitskaya & Finkelstein, PNAS (1999)
- Munoz et al. PNAS (1999)
- Alm & Baker, PNAS (1999)
- Klimov & Thirumalai, PNAS (2000)
- Clementi et al (Onuchic), JMB (2000)

"Native topology determines force-induced unfolding pathways"