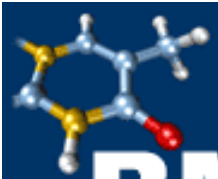


BMO

Extracting Markov models of peptide conformational dynamics from simulation data

Verena Schultheis, Thomas Hirschberger, and Paul Tavan

Motivation

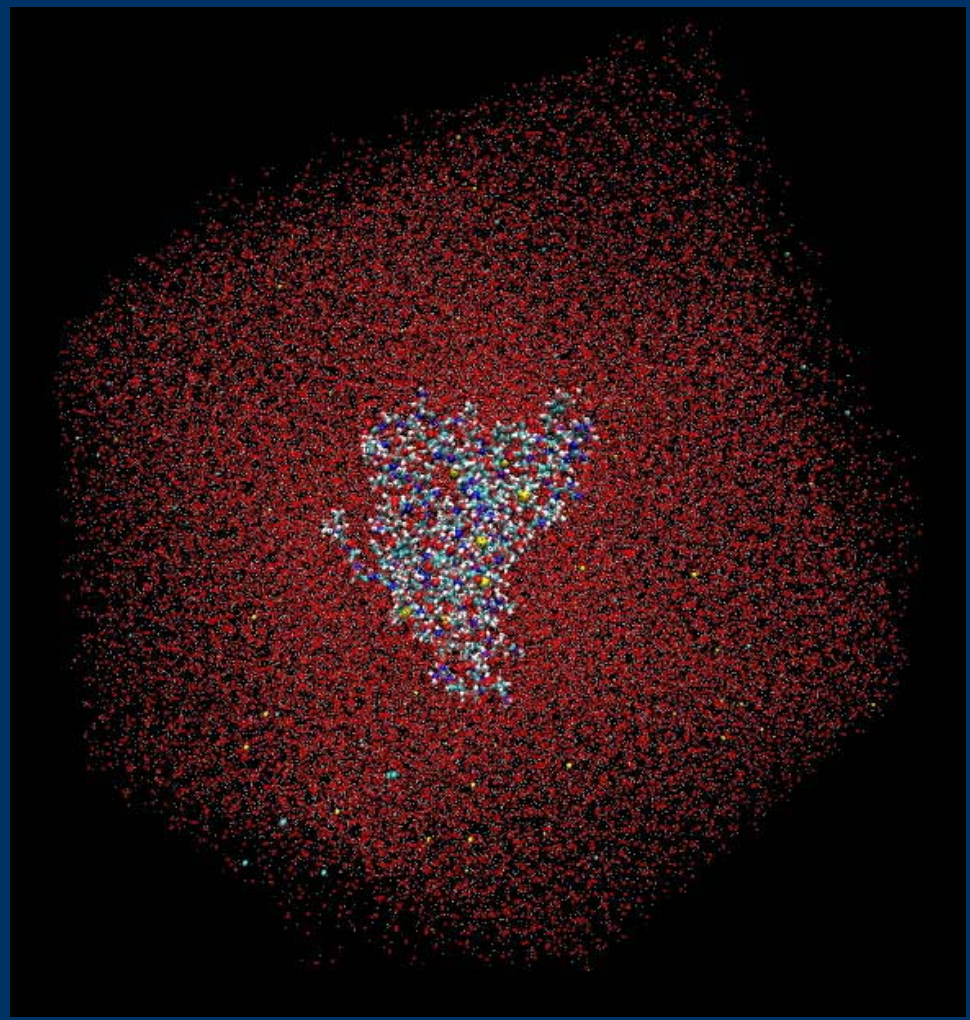
A complex example for

peptide conformational dynamics:

The cellular prion protein PrP^C in solution

Simulation system

- periodic orthorhombic dodecahedron
  (inner radius r = 52Å)
- PrP^C (125-228)
  - molecular mechanics (MM) force field: CHARMM22
  - M205R mutant obtained by remodeling the PrP^C structure
- ~25800 H_2O molecules
  (MM force field: TIP3P)
- ~150 Na^+ and Cl^- ions
  (165 mM NaCl)



PrP^C-structure: **Zahn et al., PNAS 97, 145 (2000)**

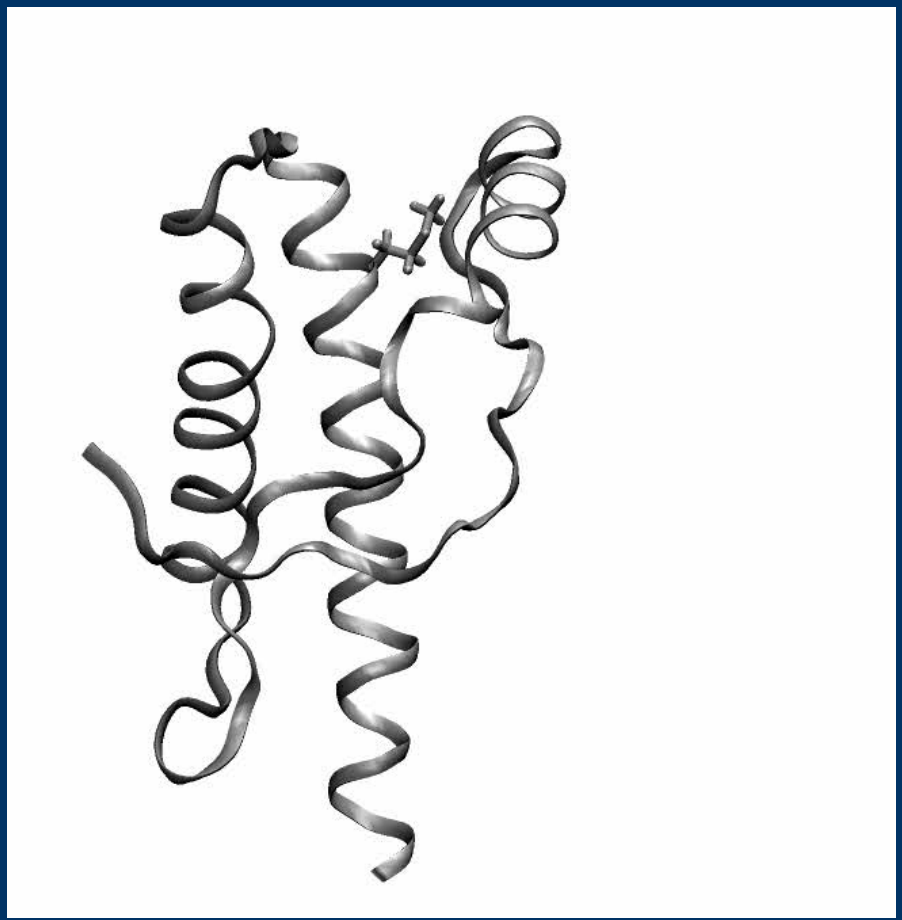
MD methods

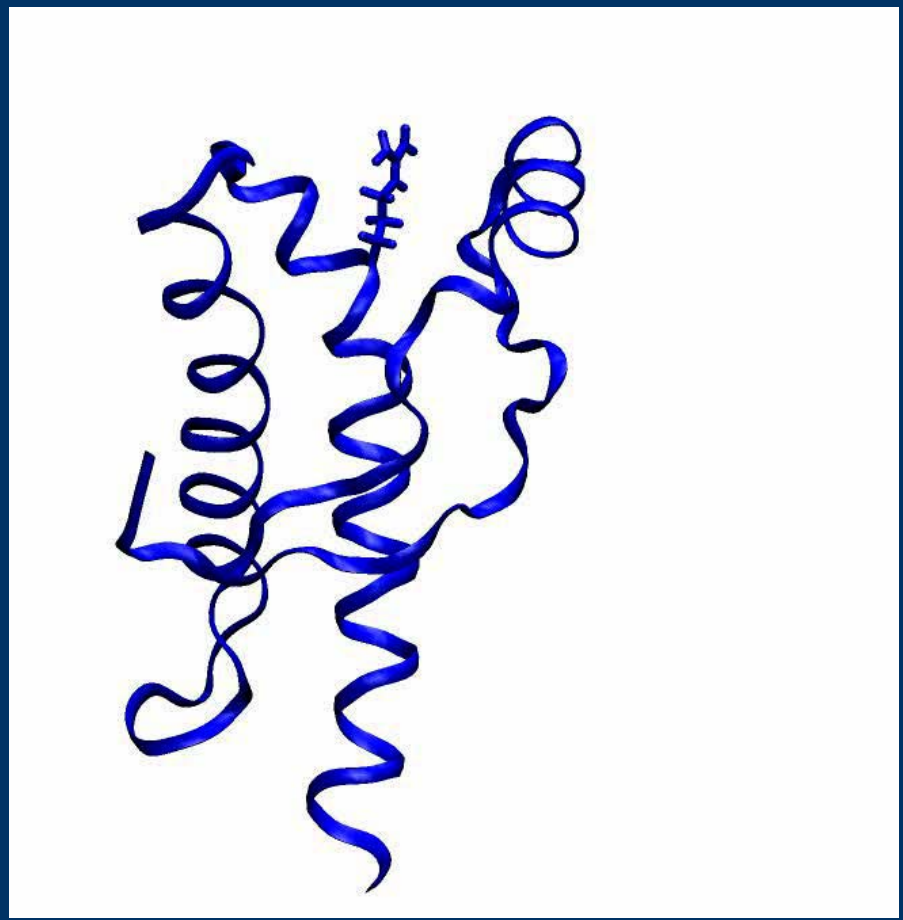
MD program EGO-MMII used for

- 10 ns simulations at T = 300 K, p = 1 atm
- time step 1fs
- computer time using six 1.6 GHz processors in parallel: ~ 20 weeks

Niedermeier, C, and P Tavan (1994). *J. Chem. Phys*. **101**: 734-748.
Niedermeier, C, and P Tavan (1996), *Mol. Simul*. **17**: 57-66.
Eichinger, M, H Grubmüller, H Heller, and P Tavan (1997). *J. Comp. Chem*. **18**: 1729-1749.
Mathias, G, B Egwolf, M Nonella, and P Tavan (2003). *J. Chem. Phys.* **118**, 10847-10860.
Mathias, G, and P Tavan (2004). *J. Chem. Phys*. **120**, 4393-4403.

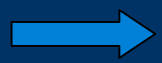
Trajectories





General Problem

How can one gain insight into such processes beyond showing movies?

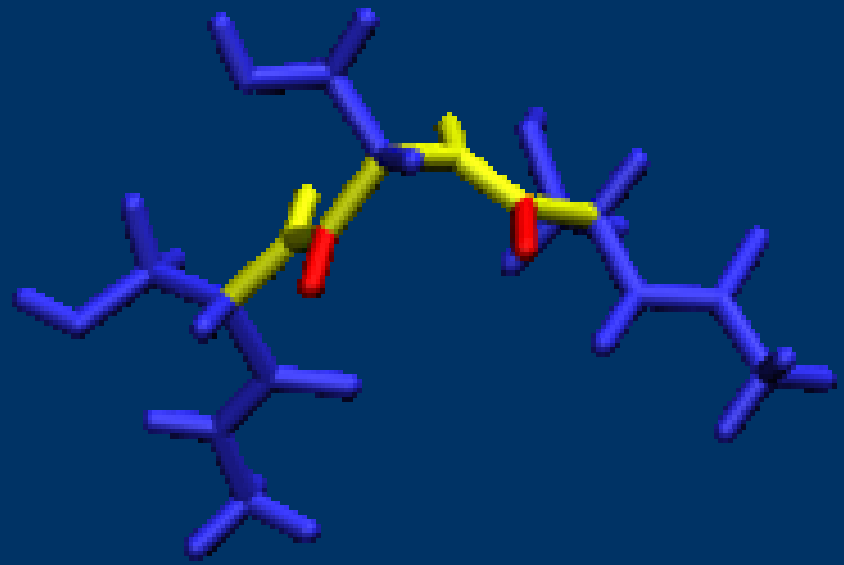
complex virtual reality ?  simplified models

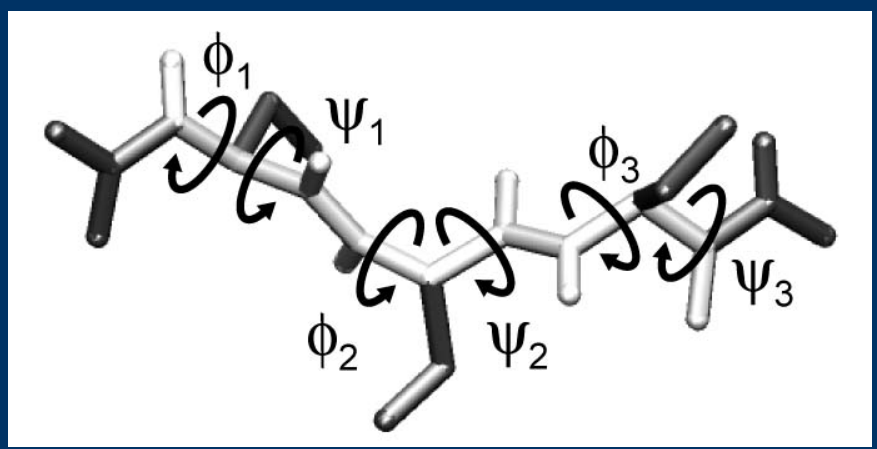
A much more simple example

50 ns backbone dynamics of the

tripeptide Ac-Ser-Ser-Ser-NH_2

fluctuating in water





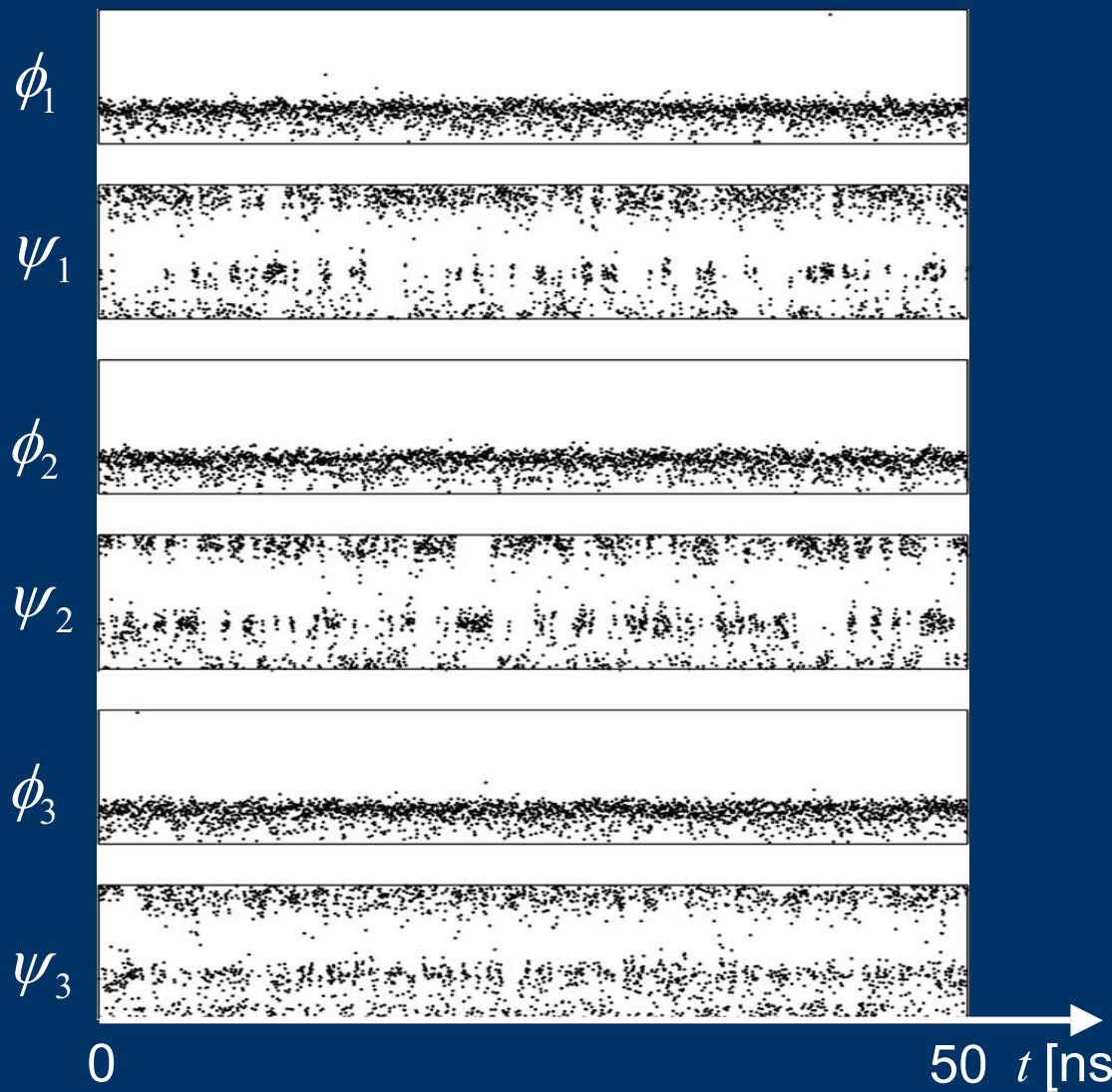
is described by a time series

$$\left\{\vec{x}_t \mid t = 1,2,\ldots,5\cdot 10^4\ ps\right\} \subset [0,2\pi[^6$$

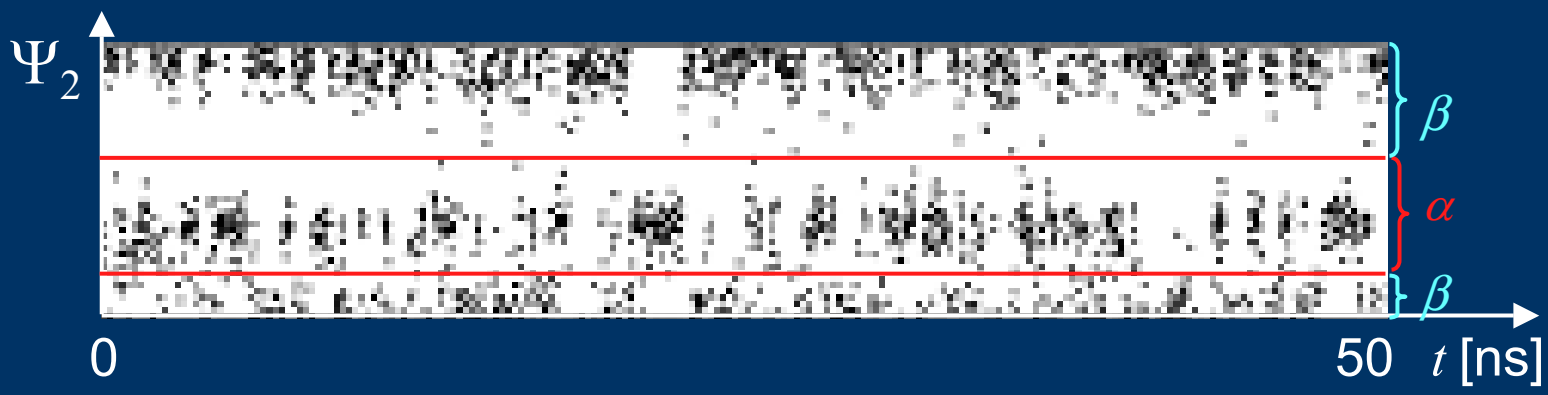
of six dihedral angles

Trajectory

$$\vec{x}_t = (\phi_1, \psi_1, \phi_2, \psi_2, \phi_3, \psi_3)_t$$



Look at Ψ_2



Ψ_2 randomly switches between two ranges α and β of values

3 Ψ angles $\Rightarrow$ $2^3 = 8$ combinations: $r = \alpha\alpha\alpha, \alpha\alpha\beta, ..., \beta\beta\beta$

$\Rightarrow$ discretization of the configuration space

into 8 disjoint partial volumes $V_r \subset [0,\ 2\pi[^6$

Conformations

The average peptide structures in the volumes V_r

$$\langle \vec{x} \rangle_r = \sum_{\vec{x}_t \in V_r} \vec{x}_t \Big/ \sum_{\vec{x}_t \in V_r} 1$$

represent coarse-grained meta-stable states, the so-called

peptide conformations $r = 1, ..., 8$

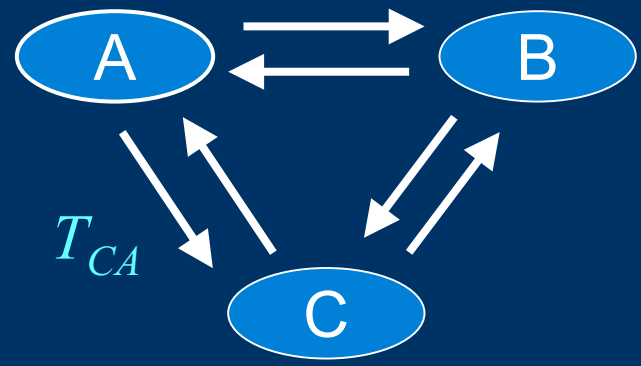
$\Rightarrow$ the x_t define a conformational dynamics

of transitions among the volumes V_r characterized by

the transfer operator $T_{rr'} = \dfrac{transitions\ r' \to r}{all\ transitions\ from\ r'}$

Markov processes

- System switching randomly between R states within τ



- Transition probabilities $T_{rr'}$ depend only on present state r‘, not on the past

- Time discrete Markov process of occupation probabilities $p_r(t)$

$$p_r(t+\tau) = \sum_{r'=1}^{R} T_{rr'} p_{r'}(t)$$

The question

The definition of conformations was natural and trivial for our tripeptide:

Suitable discretization from simple inspection of $\psi_i(t)$

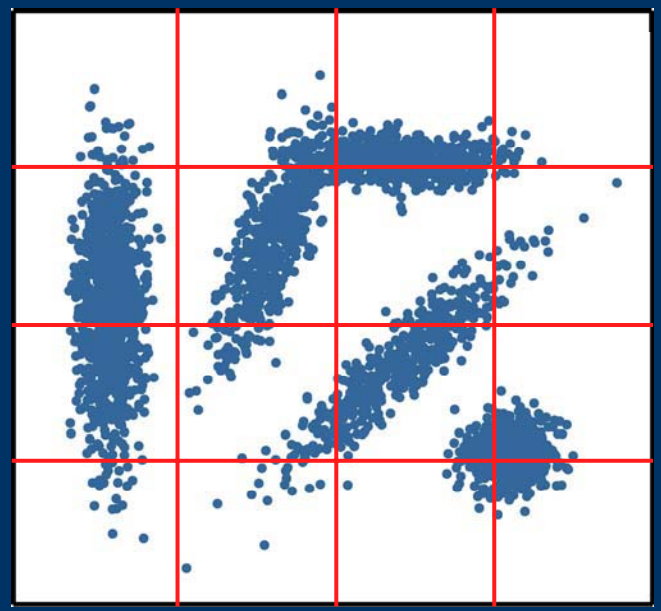
Can a suitable discretization be determined

- automatically
- without inspection
- for arbitrary proteins sampled by MD?

Naive construction of the transfer operator

Step 1: Grid partition

Step 2: Count transitions between R grid cells

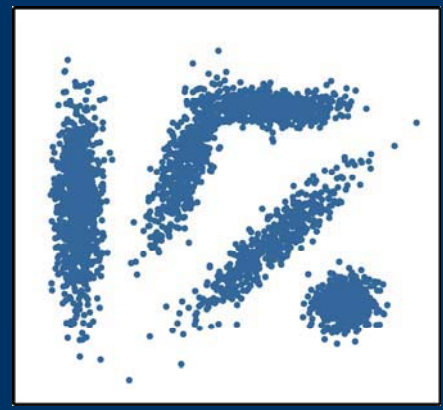


$$T_{rr'} = \frac{transitions \;\; r' \to r}{all \;\; transitions \;\; from \;\; r'}$$

Problem: $R \sim \exp(D)$ "curse of dimensionality„

bad statistics for many V_r

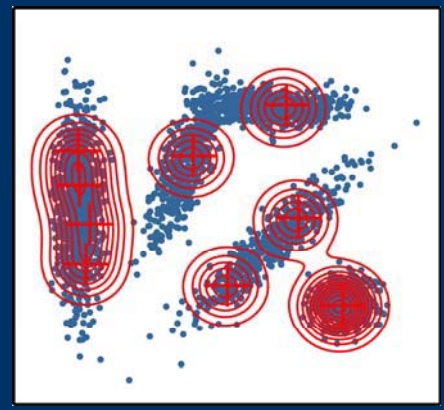
Alternative discretization



Alternative discretization

Density $p(\vec{x})$ of data points $\vec{x}_t$ modeled as a mixture

$$\hat{p}(\vec{x} \mid \vec{w}_1, \ldots, \vec{w}_R, \sigma) = \frac{1}{R} \sum_{r=1}^{R} g\left(\vec{x} \mid \vec{w}_r, \sigma\right)$$



of univariate normal distributions $g\left(\vec{x} \mid \vec{w}_r, \sigma\right)$

centered at points $\vec{w}_r$ with identical widths σ and weights $1/R$.

Maximum likelihood principle ⇨ parameters $\sigma, \left\{\vec{w}_r \mid r = 1, \ldots, R\right\}$

Algorithms:

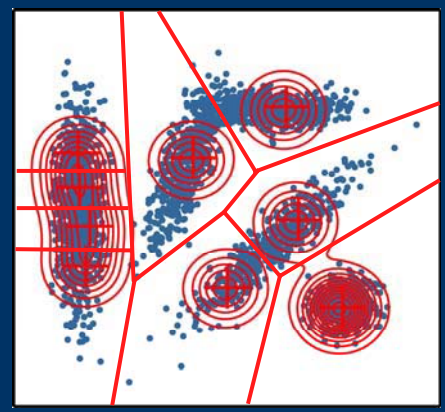
Kloppenburg & Tavan (1997). *Phys. Rev. E* **55**, 2089-2092.

Albrecht et al. (2000), *Neural Networks* **13**, 1075-1093.

Alternative discretization

Bayesian association probabilities

$$\hat{P}(r \mid \vec{x}) = \frac{\frac{1}{R} g(\vec{x} \mid \vec{w}_r, \sigma)}{\hat{p}(\vec{x} \mid \vec{w}_1, \ldots, \vec{w}_R, \sigma)}$$



define fuzzy volumes V_r discretizing configuration space.

Advantages of this fuzzy partition:

- number of V_r independent of D
- load balance:

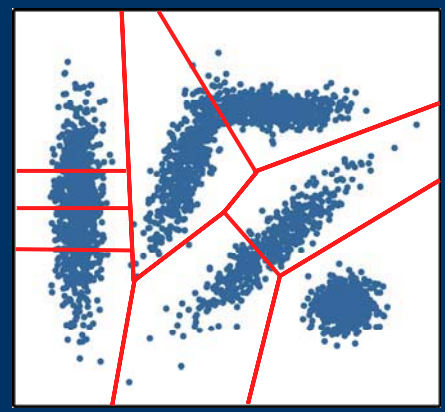
$$\forall r: \quad \left\langle \hat{P}(r \mid \vec{x}_t, \vec{w}_1, \ldots, \vec{w}_R, \sigma) \right\rangle \approx \frac{1}{R}$$

$\Rightarrow$ same statistics for all V_r

Alternative discretization

But otherwise it is still an abritrary discretization

How to construct a „natural“ discretization ?

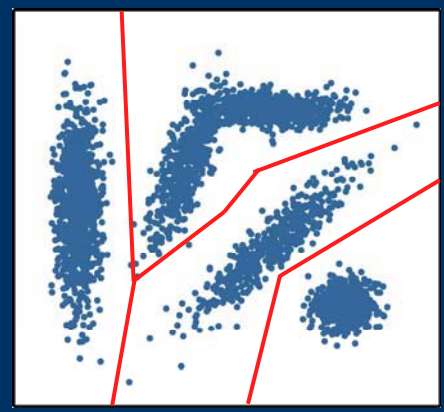


Alternative discretization

But otherwise it is still an abritrary discretization

How to construct a „natural“ discretization ?

like this:



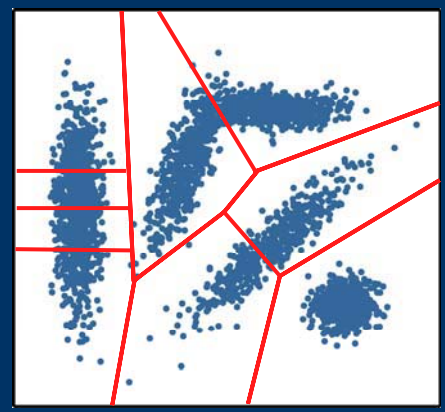
Idea:

Construct „natural“ coarse-grained discretization by

- successive unification of originial fuzzy sets V_r
- guided by an analysis of the Markovian dynamics given by the V_r

Transfer operator

Trajectory $\Rightarrow$ R-dim Markovian transfer matrix



$$T_{rr'} = \frac{\left\langle P(r \mid \vec{x}_{t+1}) P(r' \mid \vec{x}_t) \right\rangle}{\left\langle P(r' \mid \vec{x}_t) \right\rangle}$$

Choice of R dictated by statistics ($R << $ number of points $\vec{x}_t$)

Criteria for sequentially unifying the V_r

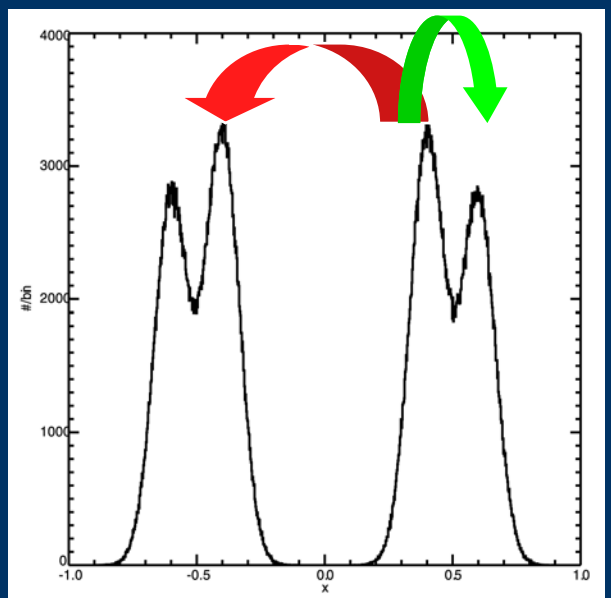
$\rightarrow$ look at most simple 1-dim example

1-dim sample data

define a 4-state Markov matrix

$$T^{ex} = \begin{pmatrix} 80 & 17 & 0 & 0 \\ 20 & 80 & 3 & 0 \\ 0 & 3 & 80 & 20 \\ 0 & 0 & 17 & 80 \end{pmatrix} \%$$

3% 17%



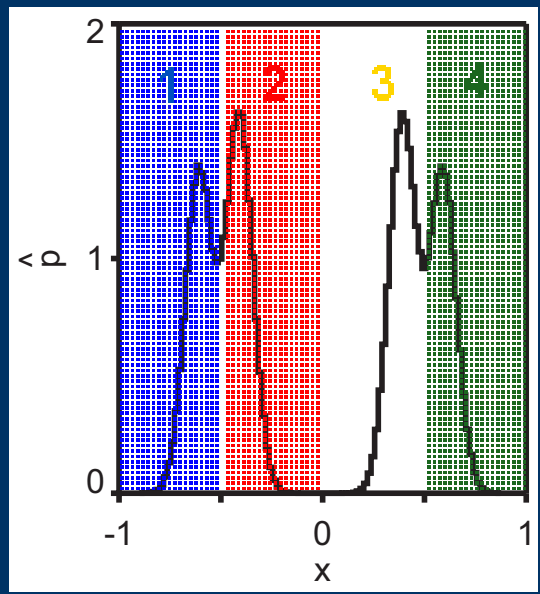
- associate a 1-dim normal distribution to each state
- generate a 1-dim trajectory x_t, $t = 1, 2, ...,$ in a two-stage stochastic process

$\Leftarrow$ invariant density $p_{inv}(x)$ of the process estimated by R-bin histogram of $\{x_t\}$ or R-component Gaussian mixture

Bayesian classification of the data

Model of $p_{inv}(x)$:

- coarse grained states are obvious
- and are obtained by Bayesian classification



Counting $\Rightarrow$

$$T^{Bayes}{}_{nn'} = \frac{\text{transitions } n' \rightarrow n}{\text{all transitions from } n'} = \begin{pmatrix} 72 & 24 & 0 & 0 \\ 28 & 73 & 3 & 0 \\ 0 & 3 & 73 & 28 \\ 0 & 0 & 24 & 72 \end{pmatrix} \%$$

Bayesian classification of the data

but otherwise coarse graining is trivial in 1-dim

$$T^{ex} = \begin{pmatrix} 80 & 17 & 0 & 0 \\ 20 & 80 & 3 & 0 \\ 0 & 3 & 80 & 20 \\ 0 & 0 & 17 & 80 \end{pmatrix} \%$$

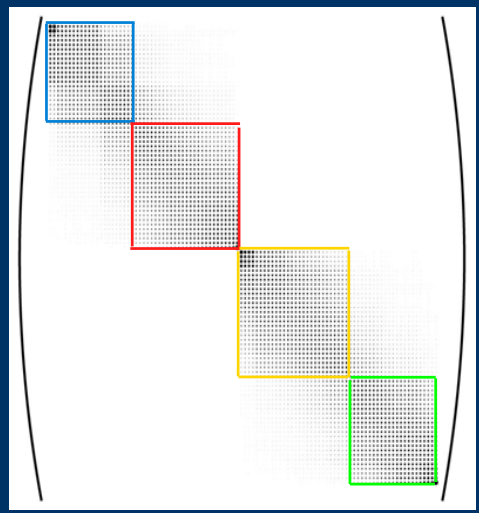
and we get the typical Bayesian decision errors for overlapping classes (optimal result)

$$= \begin{pmatrix} 72 & 24 & 0 & 0 \\ 28 & 73 & 3 & 0 \\ 0 & 3 & 73 & 28 \\ 0 & 0 & 24 & 72 \end{pmatrix} \%$$

Transfer operator

Original discretization $\Rightarrow$ R-dim Markovian transfer matrix

$$T_{rr'} = \frac{\left\langle P(r \mid \vec{x}_{t+1}) P(r' \mid \vec{x}_t) \right\rangle}{\left\langle P(r' \mid \vec{x}_t) \right\rangle} =$$



for 1-dim example

Alternative to Bayes:

Sequentially unify the fastest mixing V_r

until the four long-lived states remain!

Unification algorithm

Join sequentially states i, j with fastest transitions $i \leftrightarrow j$

Assume detailed balance: $T_{rr'}\left\langle P\left(r' \mid \vec{x}_t\right)\right\rangle = T_{r'r}\left\langle P\left(r \mid \vec{x}_t\right)\right\rangle$

Unification algorithm

Join sequentially states i, j with fastest transitions $i \leftrightarrow j$

Assume detailed balance:

$$\frac{T_{rr'}}{\left\langle P\left(r \mid \vec{x}_t\right)\right\rangle} = \frac{T_{r'r}}{\left\langle P\left(r' \mid \vec{x}_t\right)\right\rangle}$$

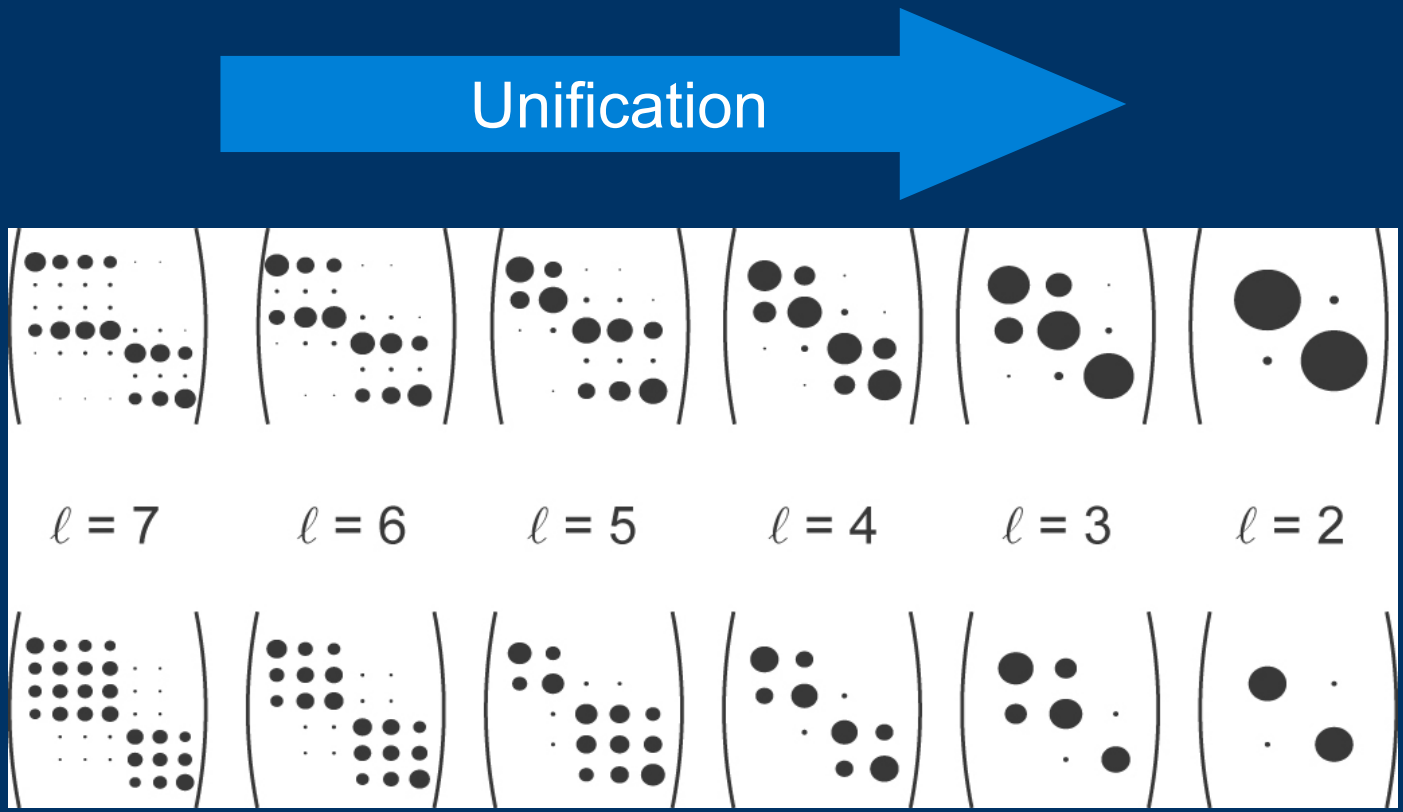
Look at $D_{rr'} \equiv \frac{T_{rr'}}{\left\langle P\left(r \mid \vec{x}_t\right)\right\rangle}$ and choose i, j : $\max_{r, r' \neq r} D_{rr'}$

Unify partition functions $P\left(n \mid \vec{x}_t\right) = \sum_{r \in I_n} P\left(r \mid \vec{x}_t\right), \quad n = 1, \ldots, R-1$

and the transfer matrix $T_{n,n'} \leftarrow T_{r,r'}$

V. Schultheis, T. Hirschberger, H. Carstens, P. Tavan (2005) *J. Chem. Theory Comput.* **1**, 515-526.

Unification algorithm

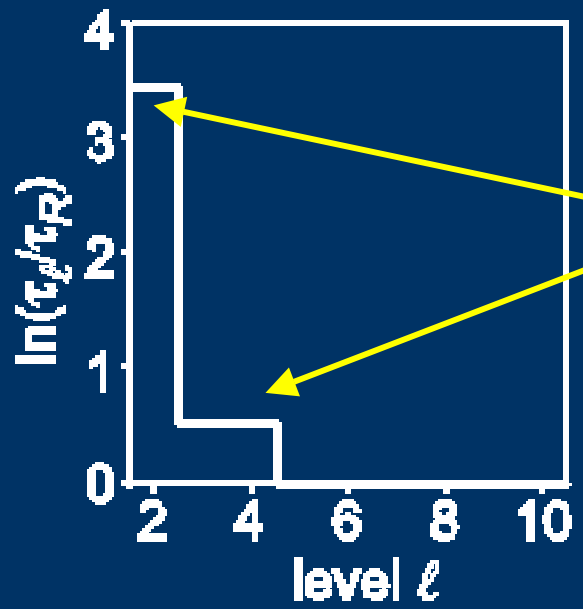


eigenvalues λ^ℓ_r

Unification algorithm

from smallest eigenvalue λ_r^ℓ

get fastest time scale τ^ℓ at each level: $\tau_\ell \equiv \frac{1}{1-\lambda_{\min}^\ell}$



4 and 2 state models are clearly distingushed by jumps to slower time scales of dynamics

„natural“ models

Unification algorithm

Unified transfer matrix at level $\ell = 4$:

$$\mathbf{T} = \begin{pmatrix} 69 & 26 & 0 & 0 \\ 31 & 71 & 3 & 0 \\ 0 & 3 & 71 & 32 \\ 0 & 0 & 26 & 68 \end{pmatrix} \%$$

approximtely reconstructs optimal decision:

$$\mathbf{T}^{Bayes} = \begin{pmatrix} 72 & 24 & 0 & 0 \\ 28 & 73 & 3 & 0 \\ 0 & 3 & 73 & 28 \\ 0 & 0 & 24 & 72 \end{pmatrix} \%$$

and the unified partition functions yield a classification of the data by

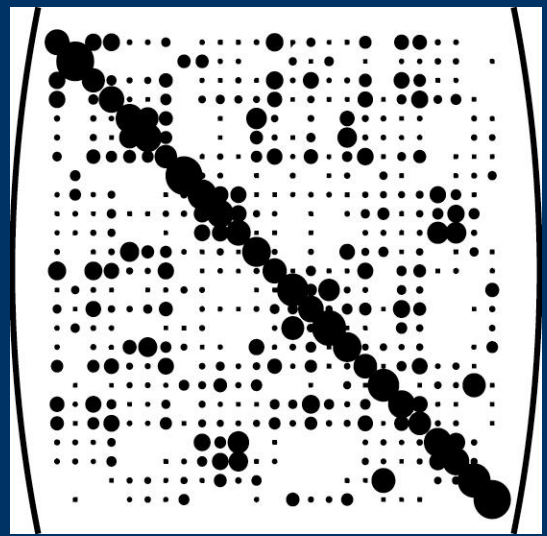
$$\vec{x}_t \rightarrow n \quad \text{if} \quad n = \max_{n'} P\left(n' \mid \vec{x}_t\right)$$

Application to tripeptide trajectory

Discretize data by 25-component Gaussian mixture

Transfer operator:

$$T_{rr'} = \frac{\left\langle \hat{P}(r \mid \vec{x}_{t+1}) \hat{P}(r' \mid \vec{x}_t) \right\rangle}{\left\langle \hat{P}(r' \mid \vec{x}_t) \right\rangle} =$$

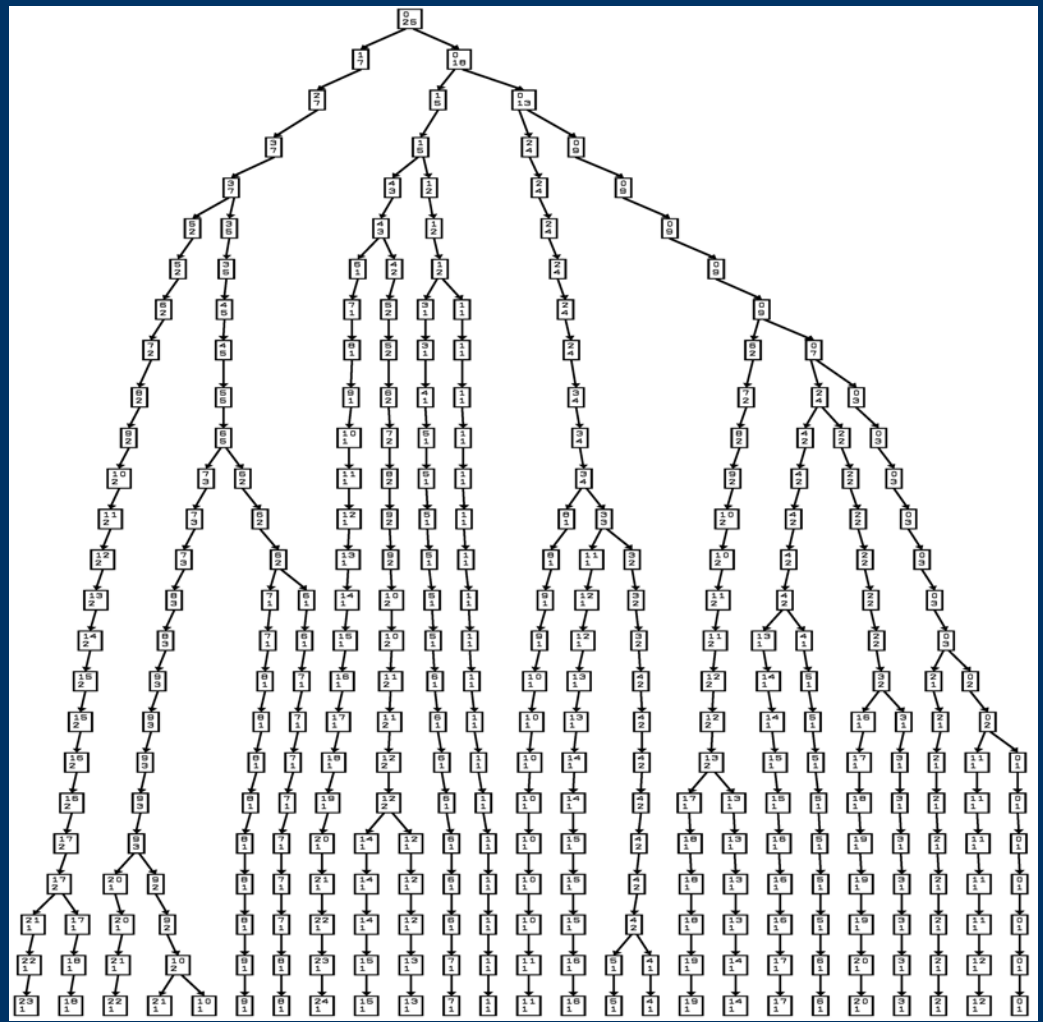


Hierarchy of Markov models

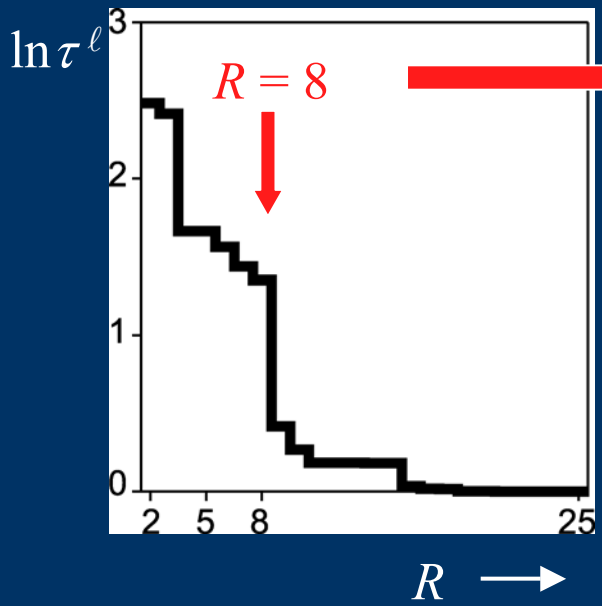
$R = 1$

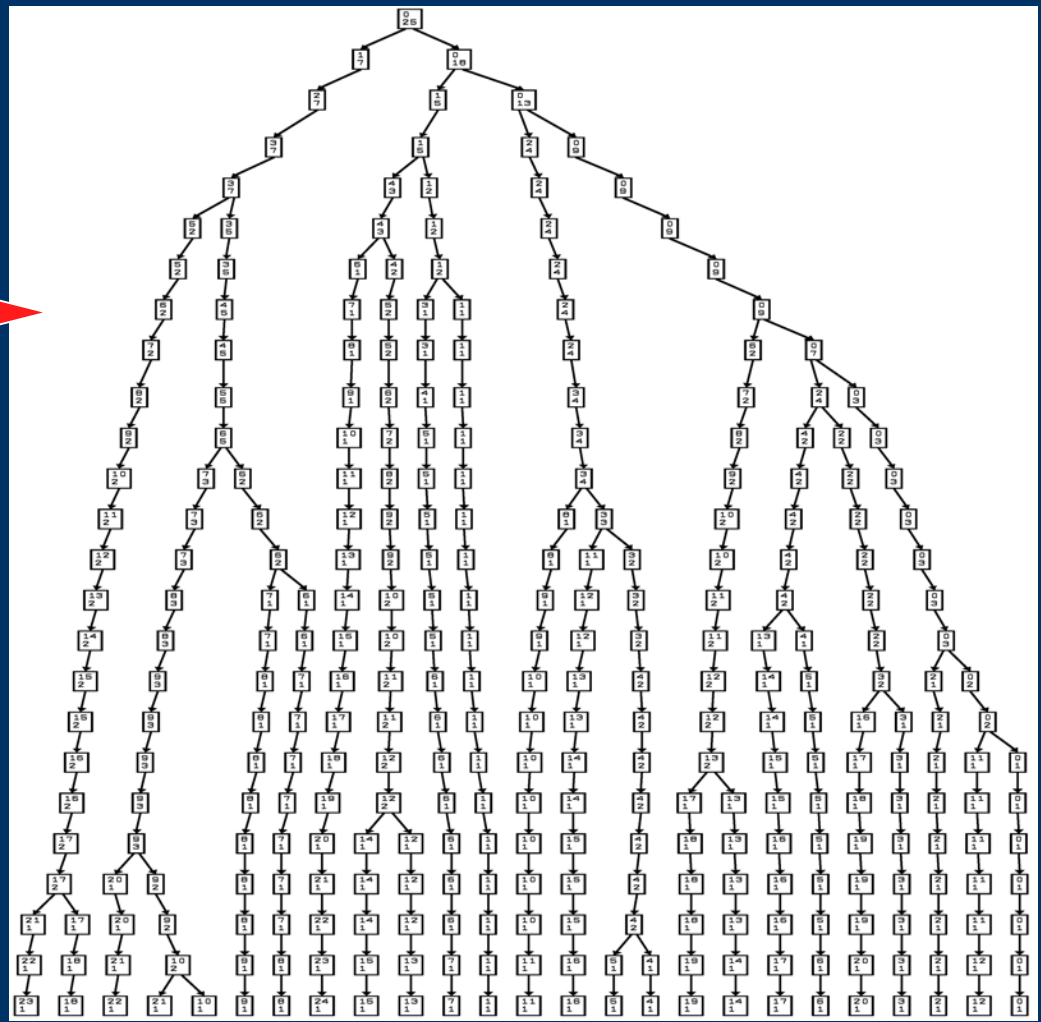
unification

$R = 25$



Hierarchy of Markov models

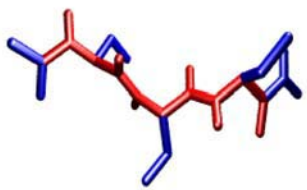




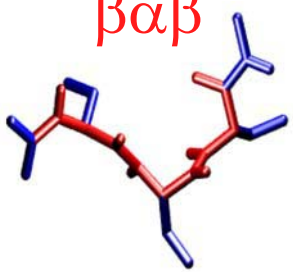
8-state model

Associated conformations $\langle \vec{x} \rangle_r$, $r = 1, \ldots, 8$

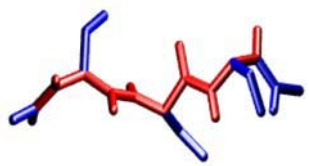
βββ



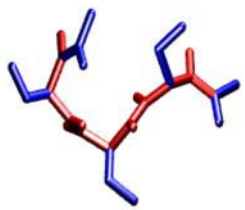
βαβ



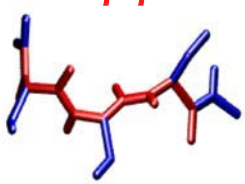
ββα



βαα

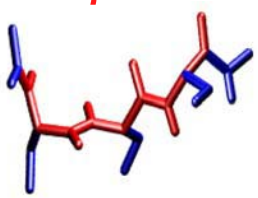


αββ

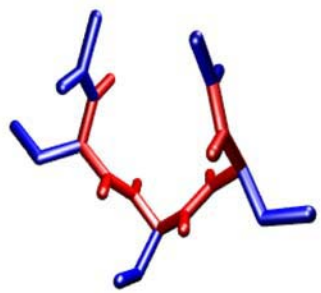


ααβ

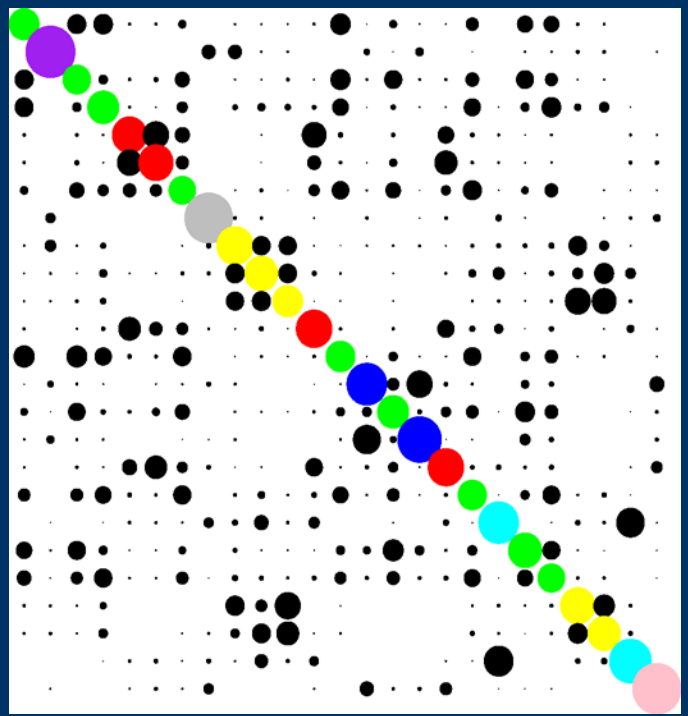
αβα

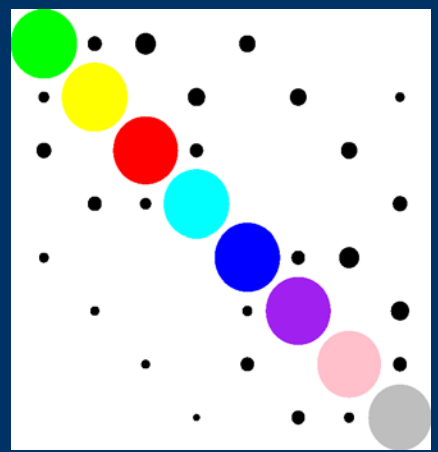


ααα



Dimension reduction





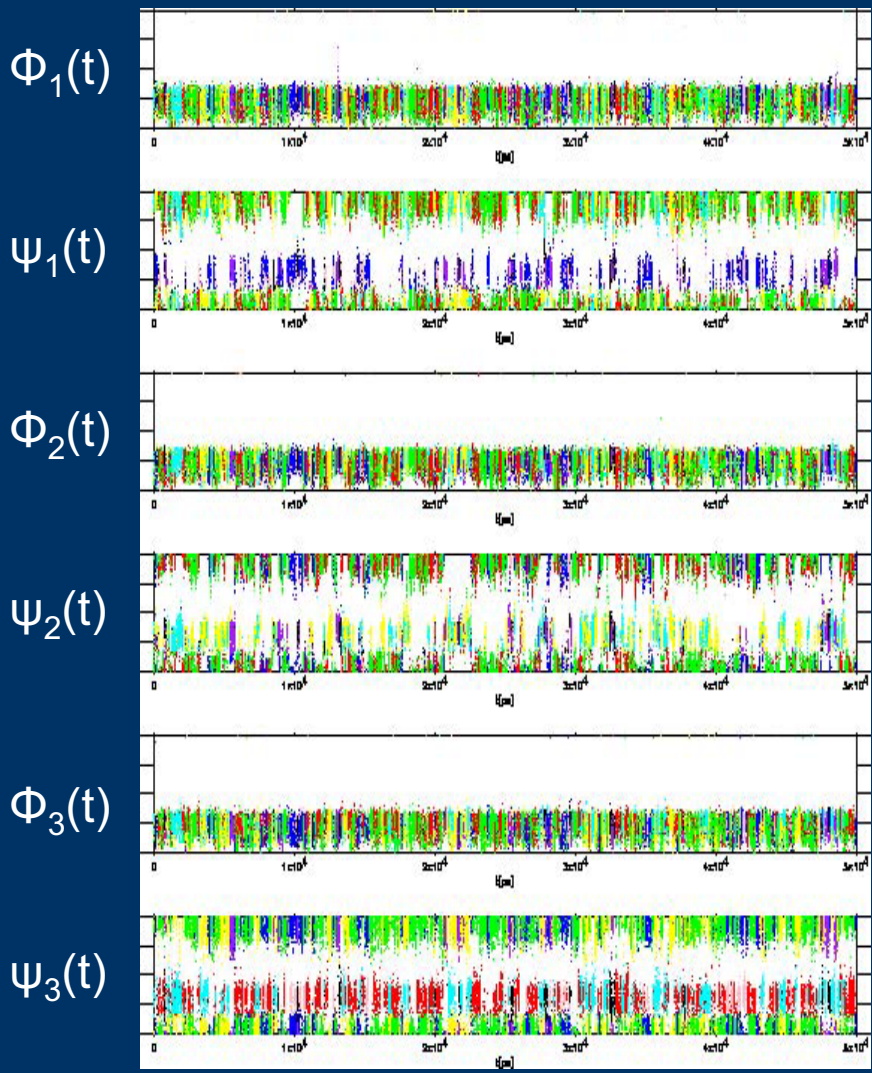
$R = 25$

$R = 8$

8-state model

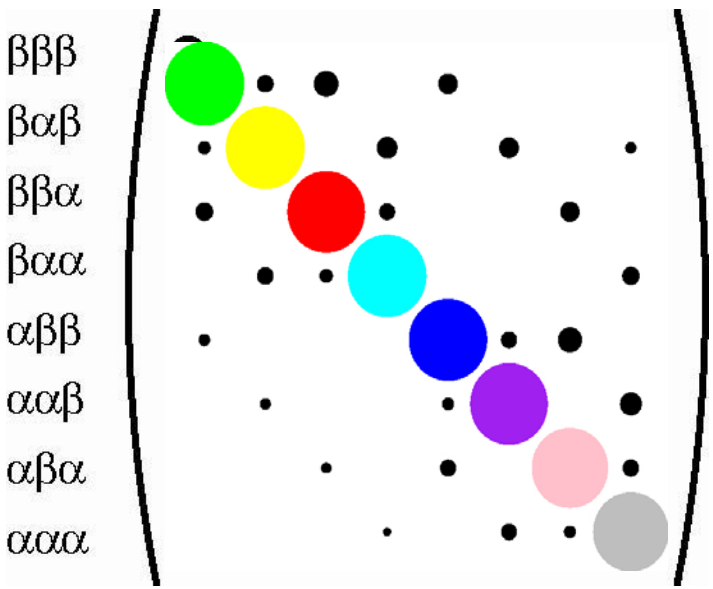
classify $\vec{x}(t)$ to

8 conformational states



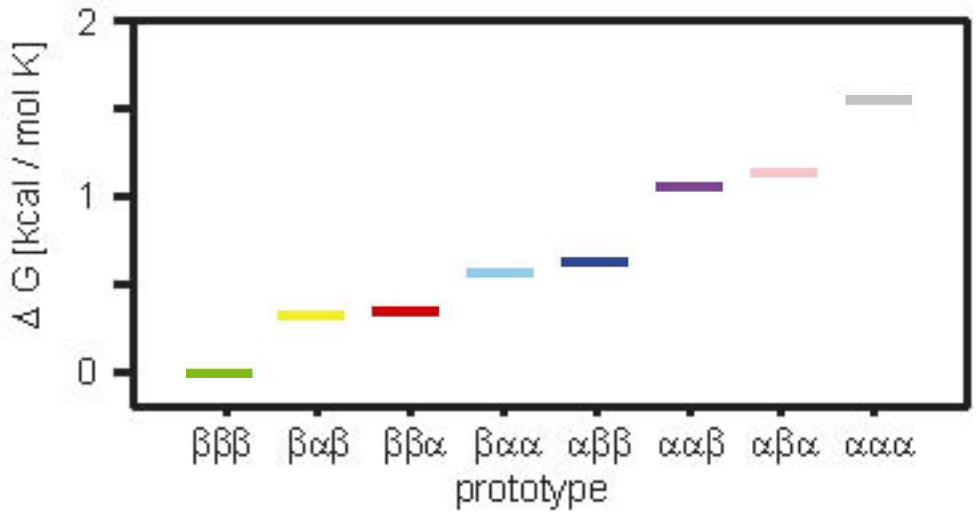
Results

Markov model



Gibbs free energy

$$G_n = -k_B T \ln \tilde{P}_n$$



Summary

Tools for the analysis of simulation data:

- fuzzy partition $\Rightarrow$ Transfer operator at good statistics and moderate dimension
- coarse graining $\Rightarrow$ hierarchy of Markov models
- most plausible model selected by certain observables

Simplified models $\Rightarrow$

insights into the structures and conformational dynamics of high-dimensional systems

BMO