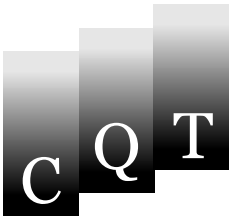
Folding Lattice Proteins: Multicanonical Chain Growth and Exact Enumerations

Michael Bachmann, Reinhard Schiemann,
Thomas Vogel and Wolfhard Janke

Institut für Theoretische Physik
Universität Leipzig

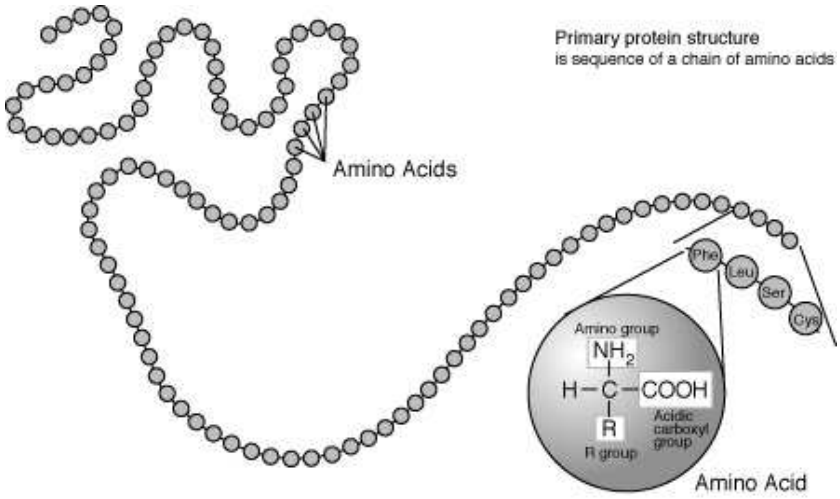
Computational
Quantum Field
Theory



Statistik-Seminar
Universität Göttingen
10 November 2004

Motivation

Proteins = chains of amino acids



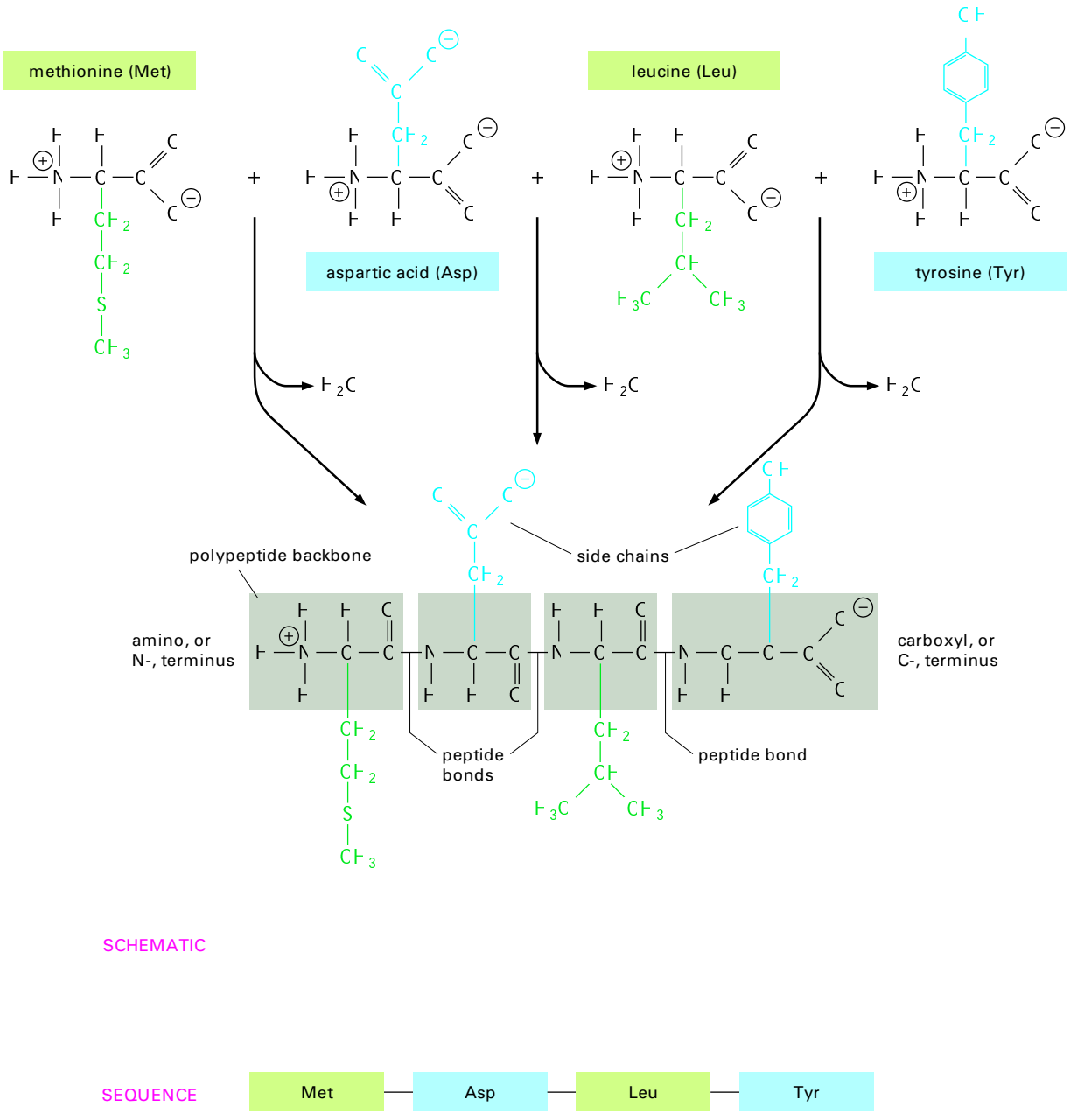
Amino acids: amino group NH_2
carboxyl group COOH
side chain R

Side chain R: distinguishing component
20 different amino acids in proteins

Primary structure = sequence of amino acids

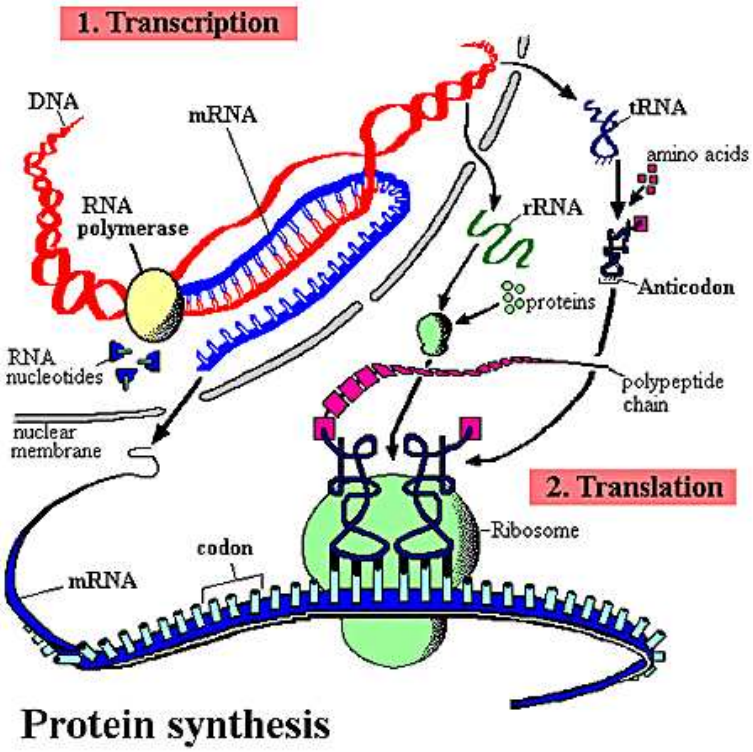
Petide bonds

Protein Structure



Typical proteins consist of $N = 50 - 4000$ residues

Protein with N residues: 20^N possibilities



Protein synthesis

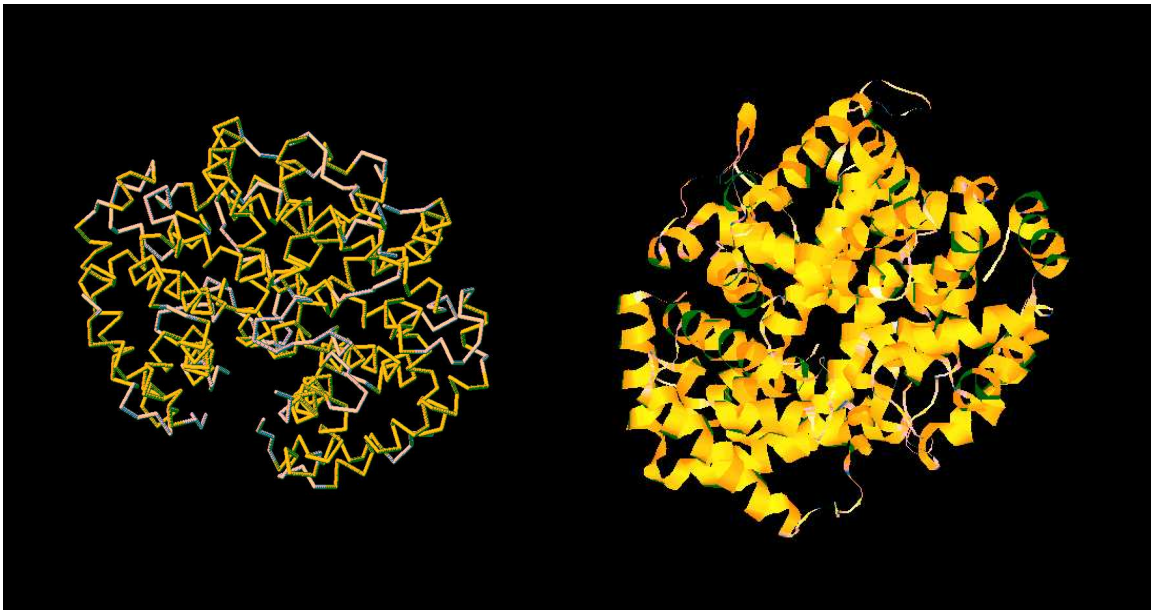
Today more than $100\,000$ sequences identified

First complete sequence: insulin (1953)

Human body: $\approx 100\,000$ different proteins

Anfinsen's experiment:

Sequence determines 3D folded structure



Hemoglobin

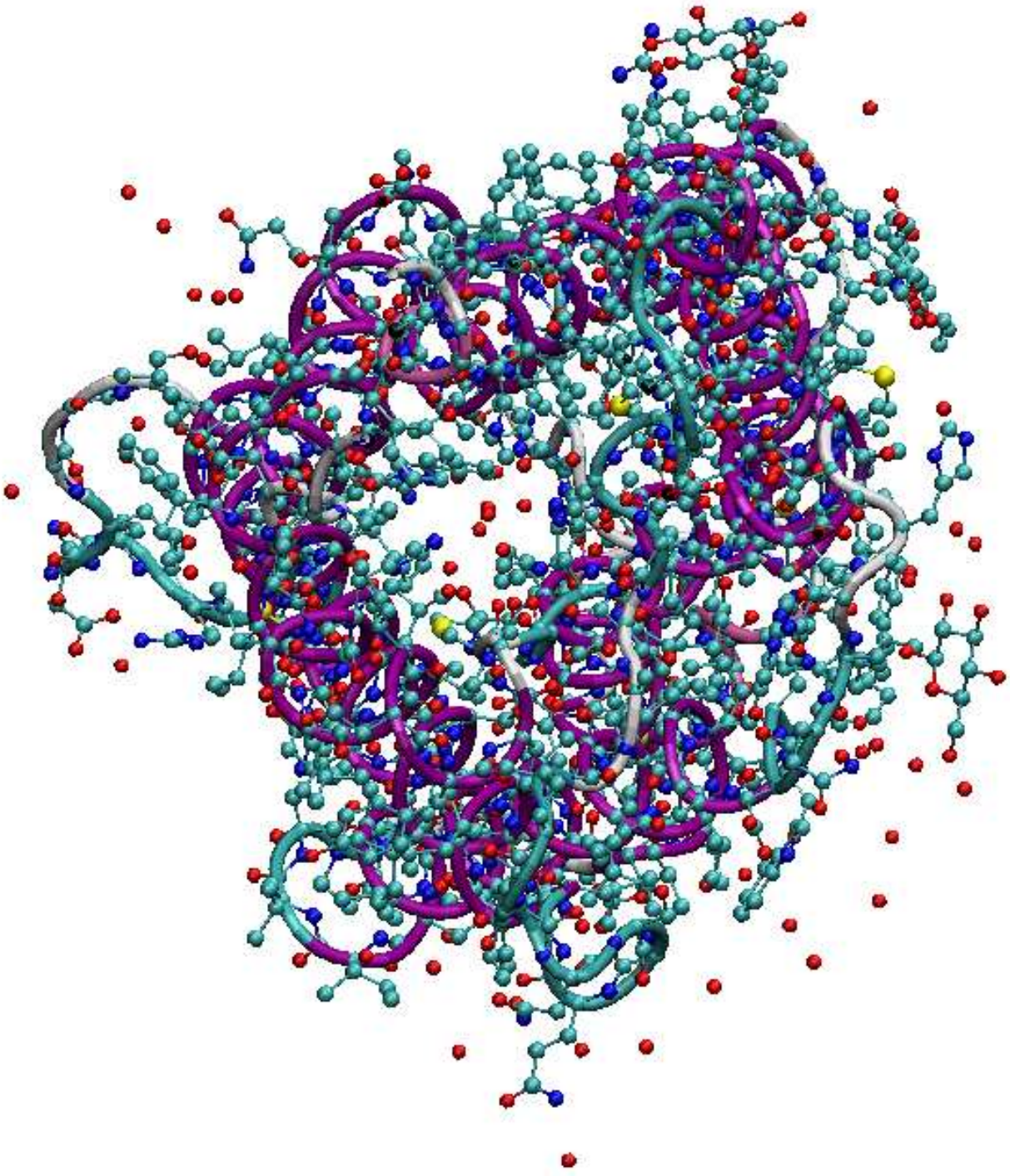
→ **Protein Data Bank, plotted with RasMol**

Secondary structures:

- α **helices**
- β **sheets**

Tertiary, quarternary structures

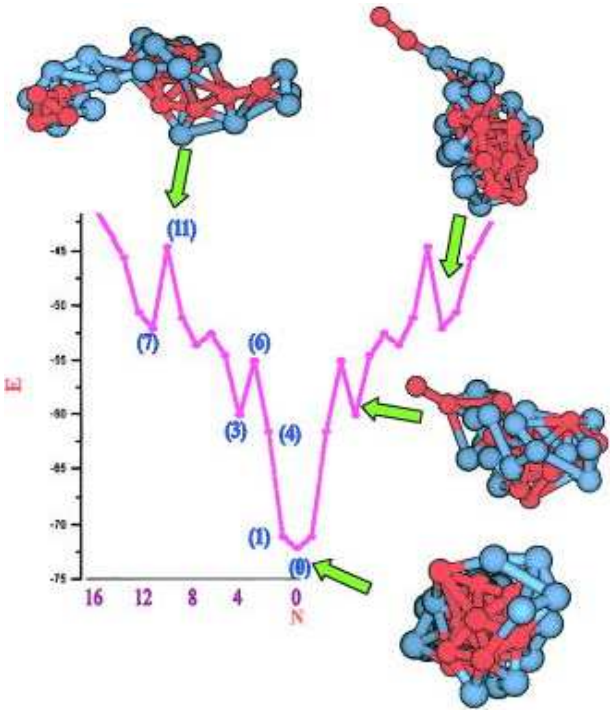
3D structure determines biological function



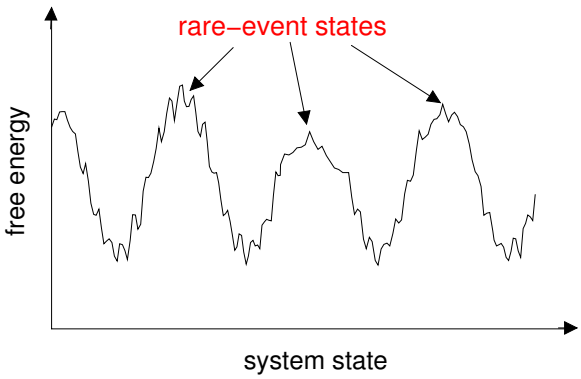
Aquaporin

Native 3D structure: Minimum in funnel-like rugged free energy landscape

Proteins



Spin Glasses



Protein sketch taken from: G. Srinivas and B. Bagchia, J. Chem. Phys. **116** (2002) 8579

Main objectives:

Direct problem:	**Given sequence $\longrightarrow$ predict 3D structure**
Inverse problem:	**Given 3D structure $\longrightarrow$ find associated sequence**

Levels of abstraction:

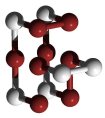
- **All-atom models**
  force fields (CHARMM, AMBER, ECEPP, GROMOS, FANTOM, . . .) from quantum chemistry and experiments, explicit/implicit solvent models
- **United residues models**
  atomistic, but coarse-grained effective interactions
- **Off-lattice heteropolymers (AB models)**
  flexible chain models with long-range Lennard-Jones interactions, A and B type of residues only
- **Lattice heteropolymers (HP models)**
  short-range interacting self-avoiding random walks, two types of monomers: hydrophobic (H) and polar (P)

  Ising model of biophysics ?

Overview

I. HP model

- 3D HP lattice proteins
- Density of states: energetic quantities in thermodynamics
- Results from exact enumeration in the space of sequences of 14mers and 18mers



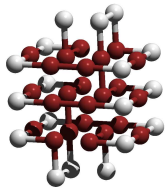
- Stochastic search strategies for longer chains
- Updates: move sets vs chain growth

II. Multicanonical chain growth

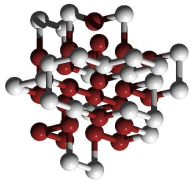
- Idea
- Realisation

III. Results

- Simulating a 42mer: model for parallel β-helix of *pectate lyase C*



- Density of states of 48mers with different ground-state degeneracies



3D HP Lattice Proteins

Lattice heteropolymers in 3D with sequence of two types of monomers:


hydrophobic (H) and


polar (P)

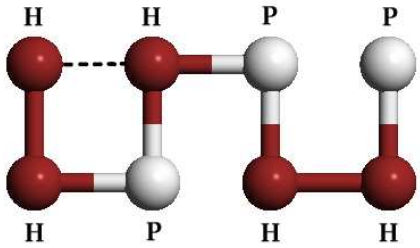
amino acids

HP protein folding principle: Screening of the hydrophobic core from the (fictitious) aqueous environment by the polar residues

HP model (Dill, 1985); simplest form: only regard to hydrophobic interaction:

$$E = -\sum_{\langle i,j<i-1\rangle} \sigma_i\sigma_j, \qquad \sigma_i = \begin{cases} 1 & \text{hydrophobic} \\ 0 & \text{polar} \end{cases}$$

$\Longrightarrow$ attraction between next-neighbored hydrophobic monomers nonadjacent along the chain, forming *HH contacts*:



Density of States

HP model – goals:

- Search for sequences with unique (native) ground state
- Analysis of relation sequence $\longleftrightarrow$ conformation: "secondary structures"
- Investigation of thermodynamic properties, e.g. conformational transitions

Density of states: degeneracy of states with energy E

Partition sum of heteropolymer with given sequence:

$$\begin{aligned} Z &= \sum_{\{\mathbf{x}\}} \exp(-E(\{\mathbf{x}\})/k_B T) \\ &= \sum_i g(E_i) \exp(-E_i/k_B T) \end{aligned}$$

with $g_0 \equiv g(E_0)$ denoting the ground-state degeneracy

Determination of $g(E)$ for lattice proteins with given sequence:

- Exact enumeration (for short chains only)
- Stochastic search methods
- g_0 via H-core construction methods (Yue, Dill, 1993/95)

Indicators of temperature-dependent transitions:

Mean energy:

$$\langle E \rangle(T) = \frac{1}{Z} \sum_i E_i \, g(E_i) \exp(-E_i/k_B T)$$

Specific heat:

$$C_V(T) = \frac{1}{k_B T^2} \left(\langle E^2 \rangle - \langle E \rangle^2 \right)$$

Helmholtz free energy:

$$F(T) \quad = \quad -k_B T \ln \sum_i g(E_i) \exp(-E_i/k_B T)$$

Entropy:

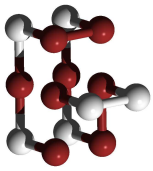
$$\begin{aligned} S(T) &= \frac{1}{T} \left(\langle E \rangle - F \right), \\ S_0 &= k_B \ln \, g_0 \end{aligned}$$

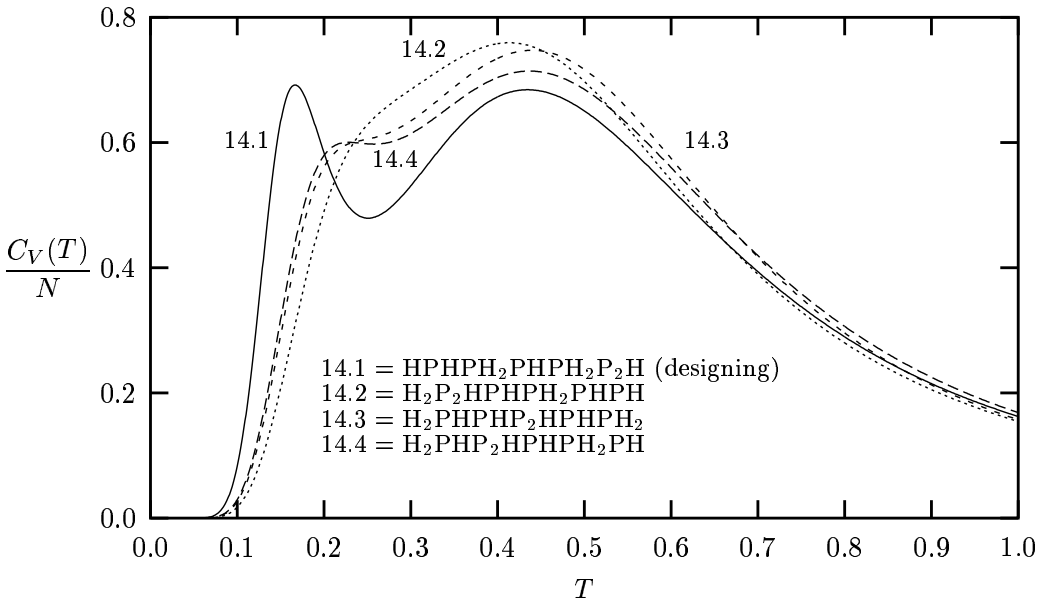
Conformational quantities: mean end-to-end distance, radius of gyration

Exact Enumeration for 14mers

- Comparison of thermodynamics for a subset of 14mers with
  - the same hydrophobicity ($n_H = 8$) and
  - identical lowest energy ($E = -8$),

  but different sequence
- On 3D s.c. lattice $\Longrightarrow$ one designing sequence only (up to a reflection symmetry): $HPHPH_2PHPH_2P_2H$

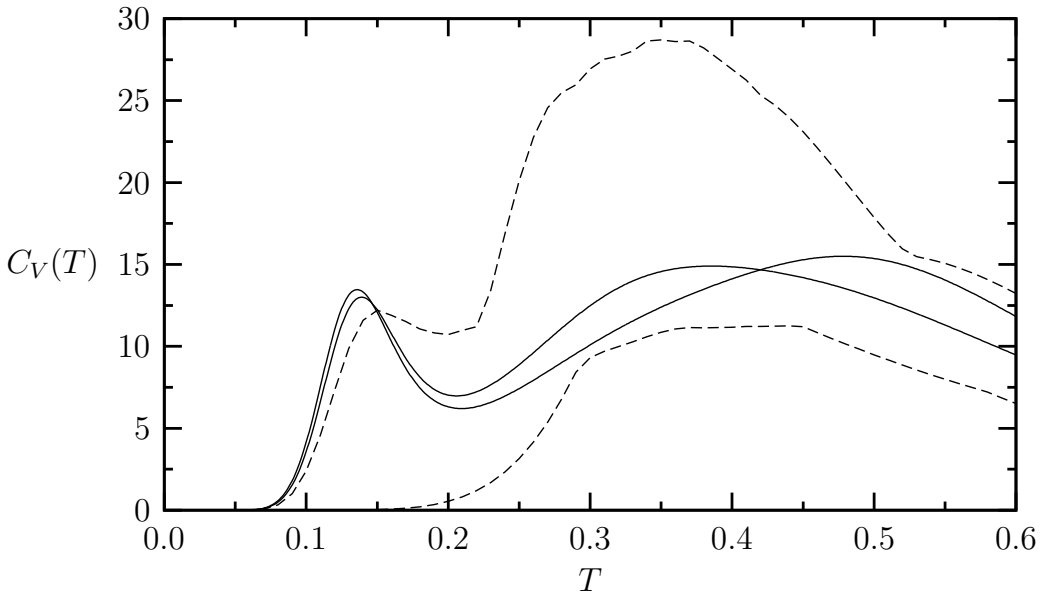




Specific heat: pronounced low-temperature peak for the designing 14mer $\Longrightarrow$ ground state–globule transition

. . . and for 18mers

- Enumeration of all $78\,955\,042\,017 \approx 8 \times 10^{10}$ conformations for each of the $2^{18} \approx 2.5 \times 10^5$ sequences
- Comparison of specific heat for a subset of 18mers with
  - the same hydrophobicity ($n_H = 8$) and
  - identical ground-state energy ($E = -9$)



Solid lines: 2 designing sequences (unique ground states)
Dashed lines: Envelope of 525 non-designing sequences

R. Schiemann, M. Bachmann, WJ, q-bio.BM/0405009, J. Chem. Phys. (in print); and Comp. Phys. Comm. (in print).

Stochastic Search Algorithms

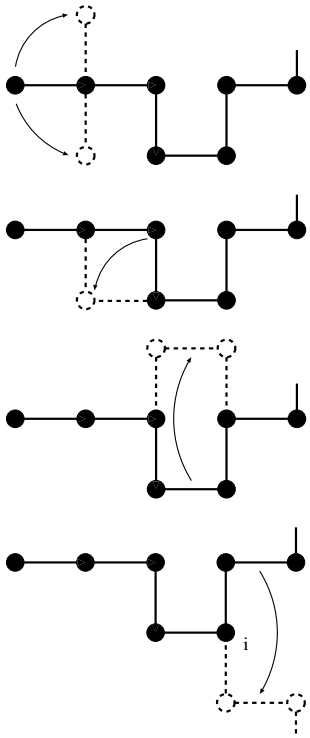
- Number of self-avoiding conformations grows exponentially with chain length N ($\sim 4.8^N$) $\Rightarrow$ enumeration exhausting for chains with more than 20 monomers
- Efficient approximate search strategies are required for statistical sampling, e.g.
  - Histogram reweighting Monte Carlo algorithms (e.g. multicanonical sampling, Wang-Landau method, multi-self-overlap ensemble)
  - Simulated and parallel tempering
  - Rosenbluth chain growth algorithms, such as PERM and nPERM$_{\rm is}^{\rm ss}$ (Pruned Enriched Rosenbluth Method)

General problem: Updating the conformation under the constraint of self-avoidance ($\Longrightarrow$ single occupation of a lattice site)

- Examples for possible strategies:
  - Move sets (e.g. consisting of end and corner flips, crankshaft moves, pivot rotations)
  - Coded updates (embedded into fixed coordinate system; F=forward, B=backward, U=up, D=down, L=left, R=right): e.g. FULBBDR $\longrightarrow$ FDLBBDR
  - Bond fluctuations (but: non-conserved bond length between adjacent monomers)
  - Chain growth (with population control)

Move Sets vs Chain Growth

Move set (frequently used and hopefully ergodic):



end flip

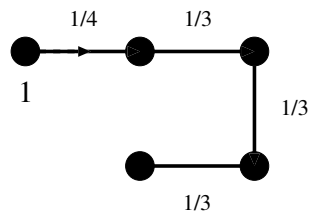
corner flip

crankshaft

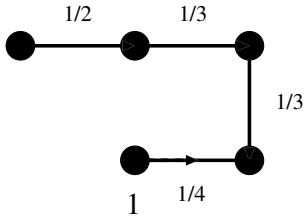
pivot rotation about any axis through ith monomer

Rosenbluth chain growth (self-avoidingly attaching a new monomer at the end of the chain until total chain length is reached): more advantageous to avoid conformational barriers!

Peculiarity (2D example, all monomers hydrophobic):



$$E = -1 \qquad p = \frac{1}{108}$$



$$E = -1 \qquad p = \frac{1}{72}$$

Necessary: bias correction by Rosenbluth weights $\sim p^{-1}$

Multicanonical Chain Growth

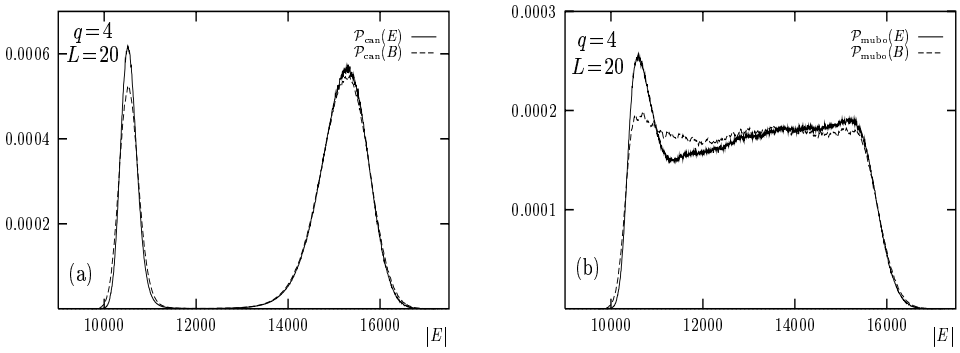
PERM (Grassberger, 1997), **nPERM**$^{\mathbf{ss}}_{\mathbf{is}}$ (Hsu, Mehra, Nadler, Grassberger, 2002):

- chain growth is controlled by comparing present weight W_n (or W_n^{pred}) with optimally chosen bounds $W_n^{>}$ and $W_n^{<}$
- $W_n^{(\mathrm{pred})} > W_n^{>}$: enrich the sample (make copies of the present chain)
  $W_n^{(\mathrm{pred})} < W_n^{<}$: prune the present chain with some probability

Multicanonical histogram (e.g. Berg, Neuhaus, 1992):

- flattening of the energy histogram by introducing a weight factor proportional to the inverse canonical distribution $\Rightarrow$ random walk in energy space

canonical multicanonical



Example from first-order transition in 3D 4-state Potts model

Idea of Multicanonical Chain Growth

Introducing additional (energy-dependent) weight into the partition sum (M. Bachmann, WJ, Phys. Rev. Lett. **91** (2003) 208105):

$$Z_n \sim \sum_{\{\mathbf{x}\}} W_n^{\mathrm{PERM}}(\{\mathbf{x}\}) W_n^{\mathrm{flat}}(E_n(\{\mathbf{x}\})) \left[W_n^{\mathrm{flat}}(E_n(\{\mathbf{x}\}))\right]^{-1},$$

where (as usual)

$$W_n^{\mathrm{PERM}} = \prod_{l=2}^{n} m_l e^{-(E_l - E_{l-1})/k_B T}, \quad 2 \le n \le N,$$

$$W_1^{\mathrm{PERM}} = 1$$

m_l : number of possible (free) sites of the lth monomer
E_l : total energy of a partial chain of length l ($E_1 = 0$)

Chain growth requires product form of the weight factors, e.g.

$$Z_n \sim \sum_{\{\mathbf{x}\}} \left[W_n^{\mathrm{flat}}(E_n)\right]^{-1} \prod_{l=2}^{n} m_l e^{-(E_l - E_{l-1})/k_B T} \frac{W_l^{\mathrm{flat}}(E_l)}{W_{l-1}^{\mathrm{flat}}(E_{l-1})}$$

with $W_1^{\mathrm{flat}} = 1$

PERM chain growth at certain temperature T:

$p_n \sim W_n^{\mathrm{PERM}} \quad \Longrightarrow \quad$ canonical distribution $P_n^{\mathrm{can},T}(E)$ of chains with length n

Density of states $g_n(E) = P_n^{\mathrm{can},\infty}(E)$?

$\longrightarrow$ e.g. through multi-histogram reweighting, requires simulations for different temperatures.

Multicanonical chain growth at $\beta \sim 1/T = 0$:

$p_n \sim W_n^{\mathrm{PERM}} W_n^{\mathrm{flat}} \quad \Longrightarrow \quad$ flat distribution $P_n^{\mathrm{flat}}(E)$ of chains with length n

$\longrightarrow$ direct simulation of $g(E)$; $P_n^{\mathrm{can},T}(E)$ by reweighting to temperature T.

Since $W_n^{\mathrm{flat}} \sim 1/g_n(E_n)$:

$$Z_n \sim \sum_{\{\mathbf{x}\}} g_n(E_n(\{\mathbf{x}\}))W_n(\{\mathbf{x}\})$$

with the combined weight $W_n = W_n^{\mathrm{PERM}} W_n^{\mathrm{flat}}$:

$$W_n(\{\mathbf{x}\}) = \prod_{l=2}^{n} m_l \frac{g_l^{-1}(E_l)}{g_{l-1}^{-1}(E_{l-1})}, \quad W_1 = 1, \quad g_1 = 1,$$

or recursively $W_n = W_{n-1} m_n g_n^{-1}(E_n)/g_{n-1}^{-1}(E_{n-1})$.

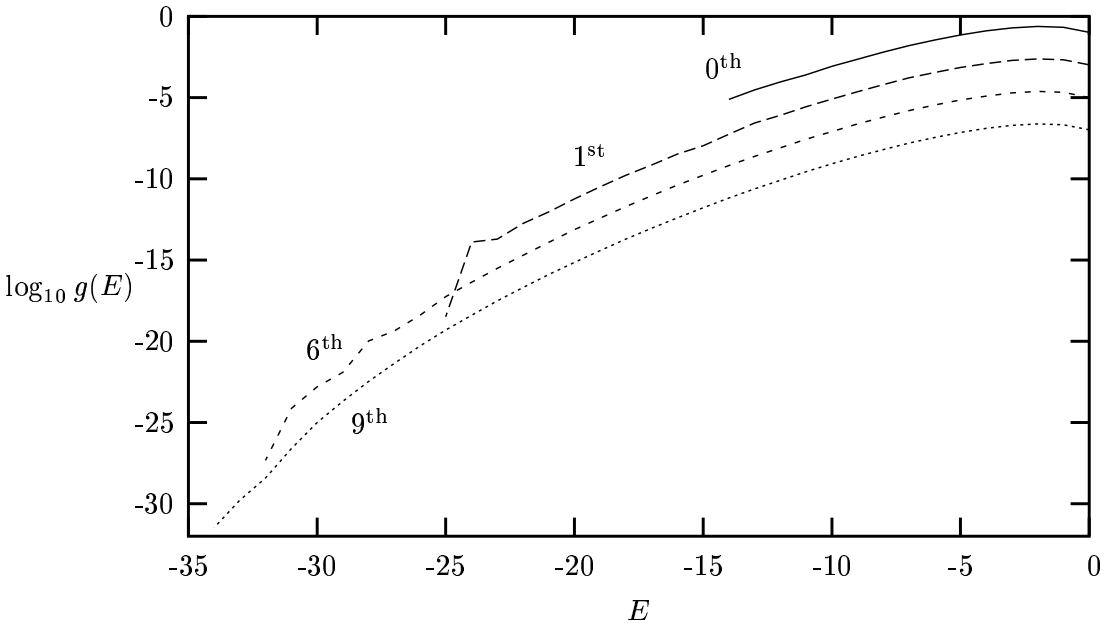
After K^{th} weight iteration: final estimate of the density of states:

$$g_n^{(K)}(E) = \frac{h_n^{(K)}(E)}{W_n^{\text{flat},(K)}(E)}$$

Remarks:

- Terminating iterations after 10^5–10^6 (iterations 0 to $(K-1)$) or 10^7–10^8 chains of length N (K^{th} iteration: measuring run)
- Number of iterations: 20–30
- Resetting Z_n, c_n, $W_n^>$, and $W_n^<$ to zero after each iteration

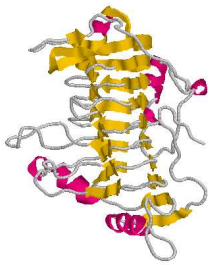
Example 42mer (curves offset by a constant):



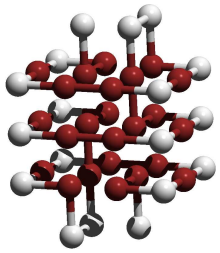
0^{th} – 8^{th} iteration: 4×10^5 chains each
9^{th} iteration (measuring run): 2×10^7 chains

≈ 25 orders of magnitude

Simulation of a 42mer



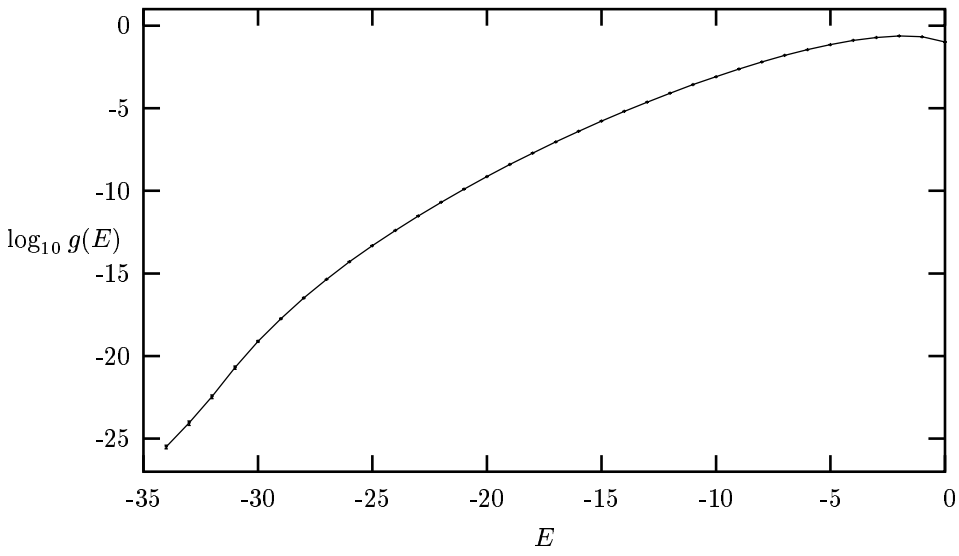
Lattice model for parallel β-helix of *pectate lyase C*



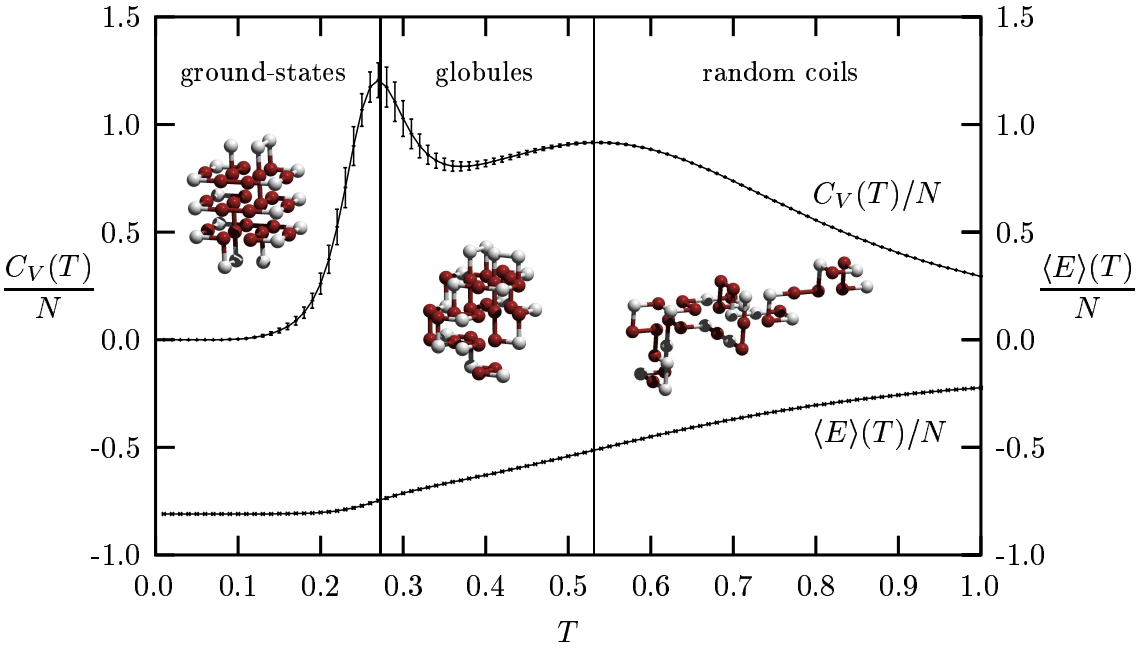
Properties of the lattice model:

- ground-state with energy -34 has 4-fold degeneracy (Yue, Dill, 1995)
- conformational transitions at $T \sim 0.27$ (ground state – globule) and $T \sim 0.53$ (globule – random coil) (see also Chikenji, Kikuchi, Iba, 1999)

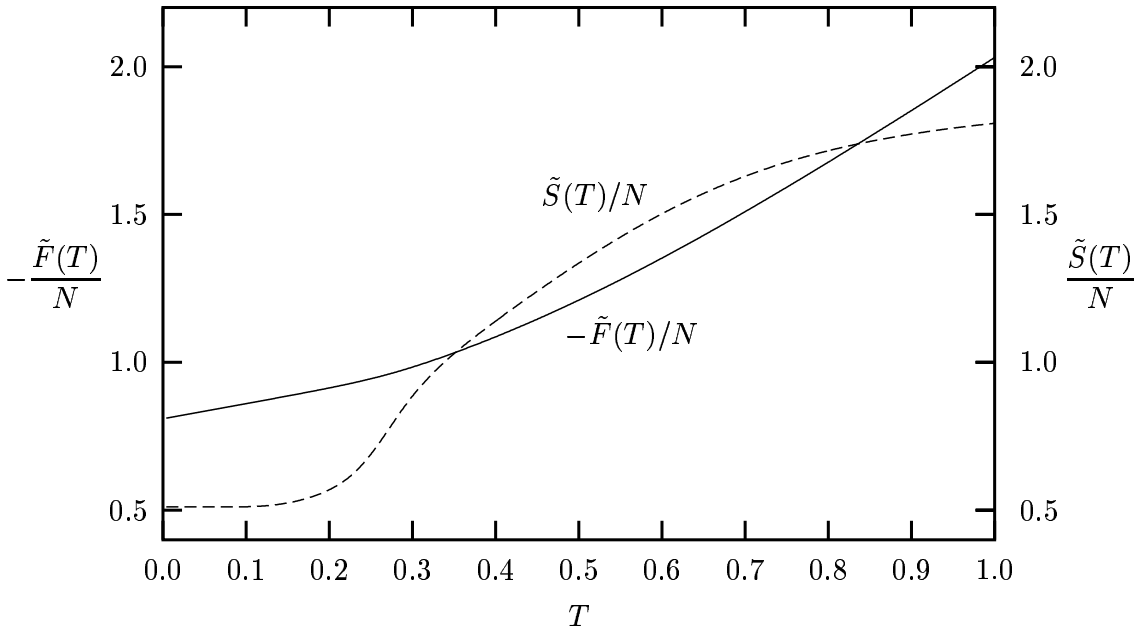
Density of states (normalized to unity):



Mean energy and specific heat:



Free energy and entropy:

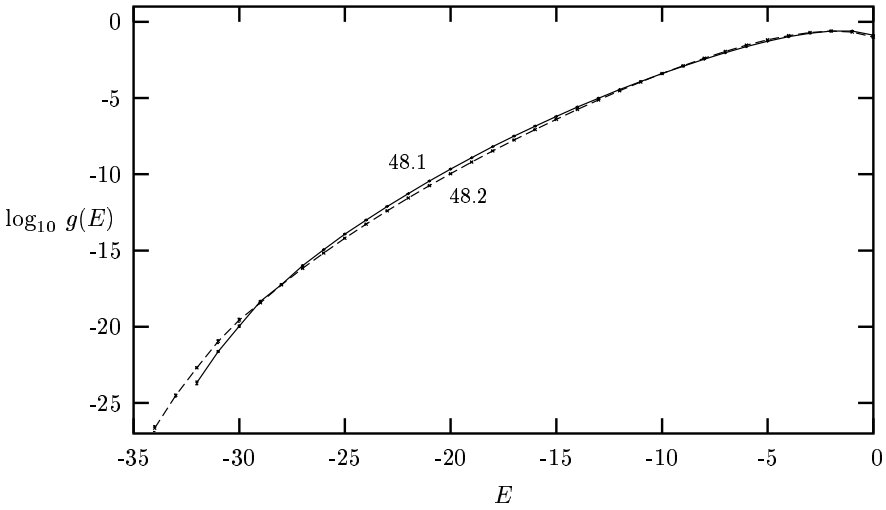


Density of States of 48mers

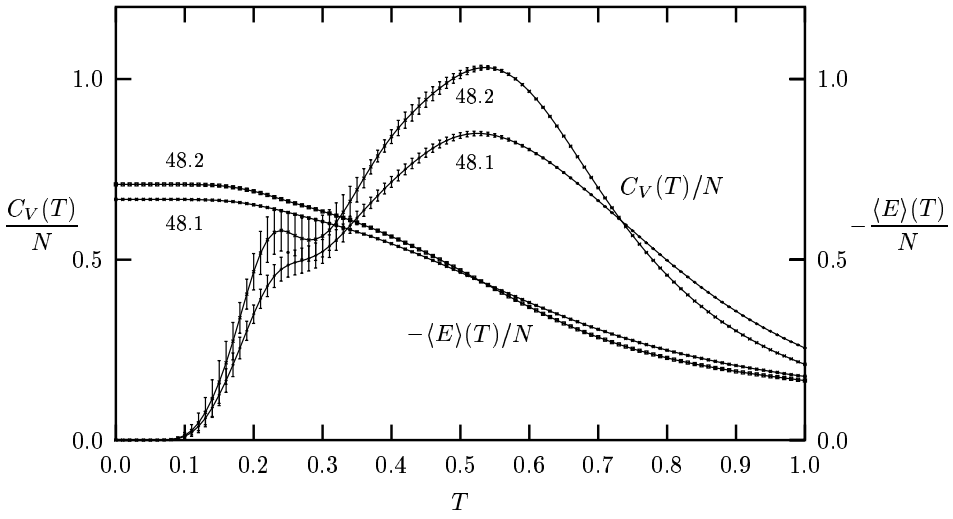
Two specific 48mers with different ground-state degeneracy:

48.1	$E_0 = -32$	$g_0 = 1.5 \times 10^6$
48.2	$E_0 = -34$	$g_0 = 5 \times 10^3$

Density of states (normalized to unity):



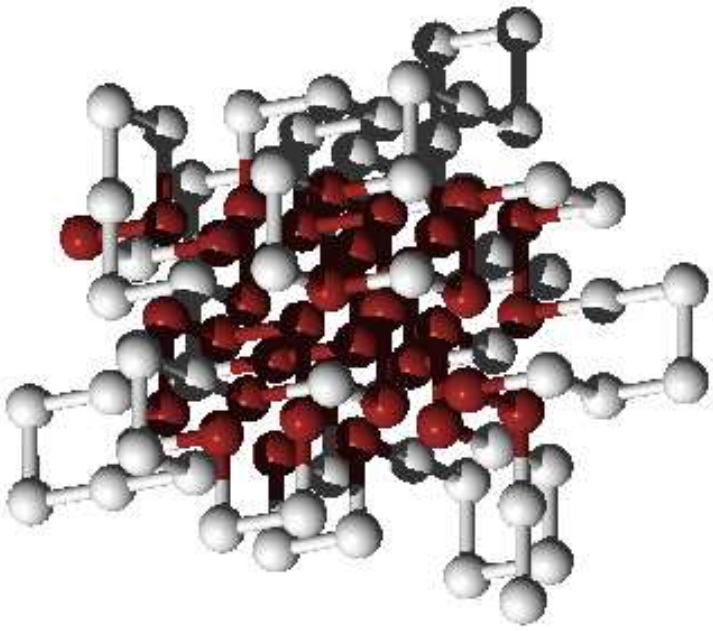
Mean energy and **specific heat:**



A 103mer

Standard test case by Lattman *et al.* (1994):

$E_{\min}$	authors
-49	Toma & Toma (1996)
-54	Hsu, Mehra, Nadler, Grassberger (2003)
-55	Hsu, Mehra, Nadler, Grassberger (2003)
-56	M. Bachmann, WJ (J. Chem. Phys. **120** (2004) 6779)



but ground-state degeneracy

$$g_0 \approx 10^{16}$$

Summary

Simulation of the density of states:

- Density of states contains all energetic thermodynamic informations
- Goal: Direct simulation of the density of states
- Updates: Chain growth more appropriate than move sets because of avoidance of conformational barriers
- Combination of chain growth algorithm (n)PERM with flat histogram techniques: sampling of the total energy space
- Accurate study of the conformational transitions, in particular the low-temperature ground state – globule transition
- Estimation of the free energy and entropy

Perspectives:

- Different lattice structures (FCC, BCC, . . .)
- Off-lattice AB models
  ⋮
- All-atom formulations