Towards Ab Initio Prediction of Protein Conformations – The Multicanonical Approach

Ulrich H.E. Hansmann [1,2]

Abstract

We demonstrate the effectiveness of the multicanonical algorithm for the tertiary structure prediction of peptides and proteins. Unlike to simulated annealing the relationship to the canonical ensemble remains exactly controlled. Hence, the new method allows not only the prediction of the lowest-energy conformation, but also the calculation of thermodynamic quantities at various temperature from one run.

1 INTRODUCTION

The prediction of tertiary structures of proteins from their primary sequences remains one of the long-standing unsolved problems (for a recent review, see, for example, Ref. [1]). It is widely believed that this structure corresponds to the global minimum in the energy. So the problem amounts to finding the global minimum energy out of a huge number of local minima separated by high tunneling barriers. Within the presently available computer resources, the traditional methods such as molecular dynamics and Monte Carlo simulations at relevant temperatures tend to get trapped in local minima. This is one of the so called NP complete optimization problems where the number of computing steps required to solve the problem increases faster than any power of the size of the system. A now almost classical way to alleviate this kind of optimization problems is *simulated annealing*[2] The method is based on the "crystal forming" process; during simulation temperature is lowered very slowly from a sufficiently high temperature to a "freezing" temperature. However, simulated annealing is not without problems. There is no established protocol for annealing and a certain number (which is not known *a priori*) of runs are necessary to evaluate the performance. Worse, due to the in praxis finite annealing steps the relationship of the obtained conformations to the equilibrium canonical ensemble at a fixed temperature remains unclear.

[1] *Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), Berlin, GERMANY.*

[2] Talk, presented at the workshop "*Bioinformatik – Computereinsatz in den Biowissenschaften*", 5. - 7. Sept. 1994, Jena, Germany.

Recently another approach was proposed: to apply the multicanonical algorithm[3] to the protein folding problem.[4] This new method was already successfully tested for systems with first order phase transitions [3, 5] and spin glasses[6] where one has to deal with similiar problems. The core of this method is to perform Monte Carlo simulations in a *multicanonical* ensemble [3] instead of the usual (canonical) Gibbs-ensemble. In this new ensemble the energy is forced onto an one dimensional random walk and a simulation may overcome the barriers between local minima by connecting back to the high temperature states. The canonical distribution for *any* temperature can then be obtained by the re-weighting techniques[7].

Here, we review the multicanonical ansatz and demonstrate its effectiveness. For Met-enkephalin we show that the lowest-energy conformation obtained agrees with that determined by other methods. As another example we study α-helix propensities of some non polar amino acids. The results are shown to be in agreement with recent experimental results.

2 THE MULTICANONICAL APPROACH

In the canonical ensemble, configurations at an inverse temperature $\hat{\beta} \equiv 1/RT$ are weighted with the Boltzmann factor $w_B(E) = \exp\left(-\hat{\beta}E\right)$. The resulting probability distribution is given by

$$P_B(E) \propto n(E)w_B(E) , \tag{1}$$

where $n(E)$ is the spectral density. In the *multicanonical* ensemble,[3] on the other hand, the probability distribution is defined in such a way that a configuration with any energy enters with equal probability:

$$P_{mu}(E) \propto n(E)w_{mu}(E) = \text{const.} \tag{2}$$

Then it follows that the multicanonical weight factor should have the form

$$w_{mu}(E) \propto n^{-1}(E) . \tag{3}$$

In order to define a explicit form of this weight factor, one can introduce two parameters $\alpha(E)$ and $\beta(E)$ as follows:[3]

$$w_{mu}(E) = \exp\Big\{-(\hat{\beta} + \beta(E))E - \alpha(E)\Big\}. \tag{4}$$

For any fixed $\beta(E)$ and $\alpha(E)$ this leads to the canonical weight factor with the inverse temperature $\beta = \hat{\beta} + \beta(E)$, hence the name "multicanonical".

Unlike to the canonical ensemble the multicanonical weight factor are not *a priori* known. Hence, the multicanonical ansatz consists of three steps. First multicanonical weight factors are constructed in a recursive way.[4] This allows to simulate the multicanonical ensemble[3] in which all energies enter with equal probability. With respect to this ensemble equilibrium configurations are generated by the standard Monte Carlo

procedure. Since the energy is forced onto a $1d$ random walk by performing a simulation in this new ensemble, one avoids getting trapped in a local minimum and the probability of finding the global minimum is increased. In the last step canonical expectation values are calculated by re-weighting[7] over a wide range of temperatures using the relation: [4, 6]

$$P_B(\hat{\beta}, E) \propto P_{mu} w_{mu}^{-1}(E) e^{-\hat{\beta}E} \ . \tag{5}$$

3 POTENTIAL ENERGY FUNCTION

The potential energy function we used for our simulations is given by the sum of the electrostatic term E_{es}, the van der Waals energy E_{vdW}, and hydrogen-bond term E_{hb} for all pairs of atoms in the peptide together with the torsion term E_{tors} for all torsion angles:

$$\begin{aligned} E_{tot} &= E_{es} + E_{vdW} + E_{hb} + E_{tors} && (6)\\ E_{es} &= \sum_{(i,j)} \frac{332 q_i q_j}{\epsilon r_{ij}}, && (7)\\ E_{vdW} &= \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right), && (8)\\ E_{hb} &= \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right), && (9)\\ E_{tors} &= \sum_{l} U_l \left(1 \pm \cos(n_l \alpha_l)\right). && (10) \end{aligned}$$

r_{ij} is the distance between the atoms i and j, and α_l is the torsion angle for the chemical bond l. The parameters ($q_i, A_{ij}, B_{ij}, C_{ij}, D_{ij}, U_l$ and n_l) for the energy function were adopted from ECEPP/2,[8]. The effect of surrounding atoms of water is neglected and the dielectric constant c is set equal to 2. The computer code KONF90,[9] was modified to accomodate the multicanonical method. The peptide-bond dihedral angles ω were fixed at the value 180° for simplicity, which leaves dihedral angles ϕ_i, Ψ_i and χ_i as independent variables.

4 RESULTS FOR MET-ENKEPHALIN

To test the effectiveness of the algorithm for the protein folding problem, we have studied one of the simplest peptide, Met-enkephalin.[4] The lowest-energy conformation for the potential energy function ECEPP/2 [8] is known[10] and analyzes with Monte Carlo simulated annealing with ECEPP/2 also exist.[11, 12] Met-enkephalin has the amino-acid sequence Tyr-Gly-Gly-Phe-Met. During the production run, which consisted of 10^5 Monte Carlo steps and followed $4 \cdot 10^4$ steps for calculating the multicanonical weight factors, the system reached the global-energy minimum region six times, at Monte Carlo steps 20128,

Table 1: Lowest-energy conformations of Met-enkephalin obtained by a multicanonical run. Conformation A is the lowest-energy conformation obtained by simulated annealing (taken from Ref. 8).

Conformation	A	1	2	3	4	5	6
E - kcal/mol -	−11.9	−11.9	−12.0	−12.0	−12.1	−12.0	−11.9
ϕ_1	98	90	91	90	97	96	98
ψ_1	154	153	152	154	151	153	156
ϕ_2	−161	−160	−157	−161	−158	−161	−163
ψ_2	69	72	64	71	71	68	65
ϕ_3	65	64	66	63	64	64	66
ψ_3	−93	−95	−92	−95	−94	−89	−92
ϕ_4	−85	−82	−80	−77	−83	−85	−80
ψ_4	−27	−26	−29	−32	−30	−31	−29
ϕ_5	−83	−81	−82	−78	−80	−82	−86
ψ_5	142	142	138	137	145	151	147
χ_1^1	−179	179	−177	179	179	−178	−176
χ_1^2	−112	−110	−117	−109	−111	−115	−114
χ_1^3	149	144	146	143	149	145	142
χ_4^1	180	−176	178	177	180	−178	180
χ_4^2	73	79	81	86	79	78	78
χ_5^1	−65	−64	−67	−67	−66	−67	−66
χ_5^2	180	−179	180	180	−176	180	176
χ_5^3	179	178	179	−179	−179	−178	−178
χ_5^4	−55	−66	−59	−62	−61	−60	−57

39521, 44462, 65412, 89413, and 95143. The lowest-energy conformation within each visit is listed in Table I together with the global-minimum energy conformation (Conformation A in Table I) obtained by simulated annealing.[11] The energies are almost all equal, and the lowest-energy value in the present work (−12.1 kcal/mol) is slightly less than the previous result (−11.9 kcal/mol) by simulated annealing.[11] Most of the dihedral angles of the six conformations also agree with the corresponding ones of Conformation A within $\approx 5^\circ$. In a recent study [13] we found that the multicanonical method is about two to three times faster than simulated annealing in predicting groundstate configurations.

In contrast to other methods we could not only reproduce the groundstate configuration, but also calculate thermodynamic quantities like energy and specific heat over a wide temperature range from just one Monte Carlo run. From this property follows one of the major advantages of the new method for studying the protein folding problem: unlike other methods it allows to investigate the relation between the global minimum in the potential energy function and the native conformation around room temperature. As an example we have calculated the fraction in which the lowest-energy conformation exists at various temperatures. The results are shown in Fig. 1. As expected, at $T = 50$ K the peptide is almost always in the ground state. As the temperature rises, the conformation is thermally excited and the fraction decreases. However, at $T = 300$ K the peptide still

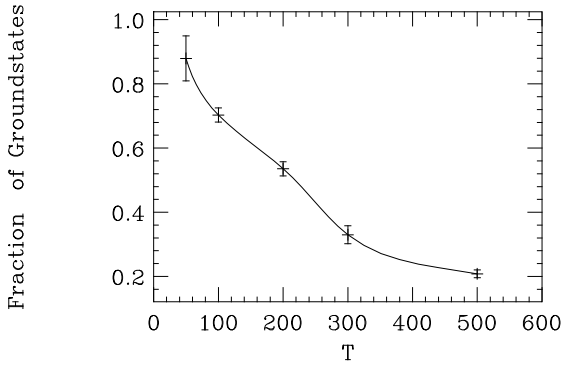

Figure 1: Fraction of groundstate configurations as a function of temperature T.

stays close to the groundstate for a substantial amount of time (≈ 35 %).

5 SIMULATION NONPOLAR AMINO ACIDS

As another example we studied α-helix propensities of homo-oligomers of nonpolar amino acids.[14] Recent experimental measurements [15, 16] suggest large differences in helix propensities among the amino acids while the older host-guest method[17] indicated small differences (for a review see Ref. [15]). Our aim is to reproduce these experimental results in a numerical simulation for three characteristic amino acids: Alanine (helix former), Glycine (helix breaker), and Valine (helix indifferent).

For our simulation we considered homo-oligomers of 10 amino acids. The criterion we adopt for α-helix formation is as follows: We consider that a residue is in the α-helix state when the dihedral angles (ϕ, ψ) fall in the range $(-70 \pm 20°, -37 \pm 20°)$. The length ℓ of a helical segment is then defined by the number of successive residues which are in the α-helix state. The number n of helical residues in a conformation is defined by the sum of ℓ over all helical segments in the conformation. In Fig. 2 we show the average % helix per residue $\frac{<n>}{N}$ $(N = 10)$ as a function of temperature for each homo-oligomer. $(\mathrm{Ala})_{10}$ is a strong helix former with % helix varying from ~ 80 % at $T = 200$ K to ~ 50 % at $T = 400$ K, and $(\mathrm{Gly})_{10}$ is a strong helix breaker with % helix varying from ~ 10 % at $T = 200$ K to ~ 7 % at $T = 400$ K, while $(\mathrm{Val})_{10}$ comes in between the two with % helix varying from ~ 35 % at $T = 200$ K to ~ 17 % at $T = 400$ K. This is in accord with the experimental results. [15, 16] From the average of n and ℓ one can calculate the helix propagation parameter s of the Zimm-Bragg model [18]

$$s = \frac{(< l > -1)(1 - \frac{<n>}{N})}{< l > (1 - \frac{<n>}{N}) - \frac{<n>}{N}} \ . \tag{11}$$

This parameter was also obtained by experiments.[15, 16] We found $s(\mathrm{Ala}) = 1.5 \sim 1.6$, $s(\mathrm{Val}) = 0.37 \sim 0.45$, and $s(\mathrm{Gly}) = 0.13 \sim 0.16$ around the experimentally relevant temperature ($\sim 0°$ C). These values are in remarkable agreement with the experiments,[16]

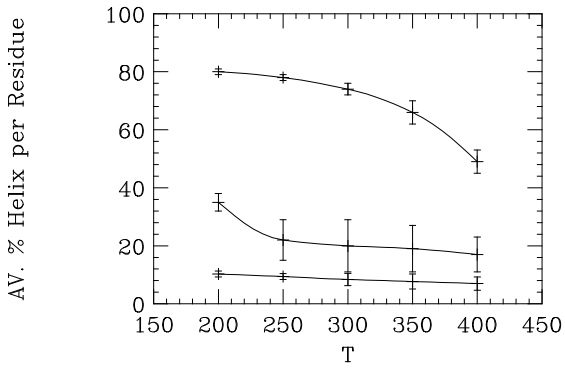

Figure 2: Average % helix per residue $\frac{<n>}{N}$ ($N = 10$) as a function of temperature for the three homo-oligomers, $(\mathrm{Ala})_{10}$, $(\mathrm{Val})_{10}$, and $(\mathrm{Gly})_{10}$.

where they give $s(\mathrm{Ala}) = 1.99 \sim 2.19$, $s(\mathrm{Val}) = 0.20 \sim 0.93$, and $s(\mathrm{Gly}) = 0.02 \sim 0.57$.

In Table 2 we present the free energy differences $\Delta G \equiv G_H - G_C$, enthalpy differences ΔH, and entropy differences $T\Delta S$ between helix (H) and non-helix (C) states. Here, a conformation is considered as in the helix state if it has a segment with helix length $\ell \geq 3$. Note that $\ell = 3$ corresponds to roughly one turn of the α-helix . The free energy differences were calculated from $\Delta G = -RT \ln \frac{N_H}{N_C}$, where N_H and N_C are average numbers of conformations in helix and non-helix states, respectively. The enthalpy differences were obtained from $\Delta H = E_H - E_C$, where E_H and E_C are average potential energies in helix and non-helix states, respectively. Finally, the entropy differences were derived from ΔG and ΔH by the relation $T\Delta S = \Delta H - \Delta G$. It is clear from the table that around temperatures near 0° C $(\mathrm{Ala})_{10}$ favors helix state with $\Delta G = -3 \sim -4$ kcal/mol and $(\mathrm{Gly})_{10}$ favors non-helix state with $\Delta G = 2.7 \sim 3$ kcal/mol, while $(\mathrm{Val})_{10}$ slightly favors non-helix state with $\Delta G = 0.4 \sim 0.8$ kcal/mol. These results again support the experimental fact that Alanine is a helix former and Glycine is a helix breaker, while Valine comes in between the two. Note that for each homo-oligomer the entropy contribution $-T\Delta S$ monotonically increases with temperature as it should because of the increased thermal fluctuations. Note also that ΔH is large negative for $(\mathrm{Ala})_{10}$, whereas it is small for $(\mathrm{Val})_{10}$ and $(\mathrm{Gly})_{10}$, suggesting that ΔH is a key factor for helix stability. This can be understood by the fact that helical conformations is one of the ideal conformations that minimize the Lennard-Jones term E_{LJ}. Fig. 3 shows for Alanine that the behavior of the energy as a function of temperature is indeed dominated by this term. The relative dominance of the Lennard-Jones energy depends on the geometry of the side chains. However, we found for all three homo-oligomers studied in the present work that the lowest energy configuration is helical.

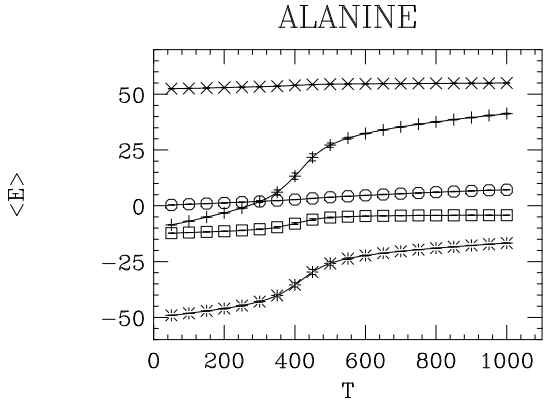

Figure 3: Total Energy E_{tot} (+),Coulomb term $E_{es}(x)$,Lennard-Jones term $E_{LJ}(*)$, Hydrogen-bond term $E_{HB}(\Box)$ and Torsion energy $E_{tors}(o)$ as a function of temperature.

Peptide	T	ΔG	ΔH	$T\Delta S$
$(\text{Ala})_{10}$	250	-4.3(1.0)	-10.6(1.1)	-6.5(1.5)
	300	-3.0(8)	-10.1(2.5)	-7.2(2.6)
	350	-1.9(5)	-10.1(2.8)	-8.3(2.9)
$(\text{Val})_{10}$	250	0.41(28)	-2.1(1.8)	-1.0(1.8)
	300	0.79(54)	-0.94(53)	-1.6(8)
	350	1.1(5)	-1.8(1.3)	-3.8(1.1)
$(\text{Gly})_{10}$	250	2.7(1.2)	1.3(2.1)	-0.4(2.4)
	300	3.1(1.0)	0.28(1.7)	-1.9(2.0)
	350	3.6(9)	-0.29(2.0)	-4.1(2.2)

Table 2: Free energy differences ΔG, enthalpy differences ΔH, and entropy differences $T\Delta S$ (in kcal/mol) between helix and non-helix states.

6 CONCLUSION

We reviewed the the multicanonical approach to the *ab initio* prediction of peptide and protein conformations. This ansatz allows not only to find the lowest energy conformation but also to calculate thermodynamic quantities over a wide range of temperatures from just one simulation. We applied the new method to some simple peptides. For Met-enkephalin the known groundstate configuration could be reproduced. Simulating homo-oligomers of nonpolar amino acids we observe direct *folding* of helices from completely *random* initial conformations. Our numerical results could qualitatively and quantitatively reproduce recent experimental results.

Acknowledgements:

This work was supported by the U.S. Department of Energy under contract DE-FC05-85ER250000. Our simulations were performed on the cluster of RISC workstations at SCRI, The Florida State University, Tallahassee, USA.

References

[1] M. Levitt, *Curr. Opin. Struct. Biol.*, **1**, 224 (1991).

[2] Kirkpatrick, S., Gelatt, C.D., Jr. & Vecchi, M.P. *Science* **220**, 671–680 (1983).

[3] Berg, B.A. & Neuhaus, T. *Phys. Lett.* **B267**, 249–253 (1991); Berg, B.A. & Neuhaus, T. *Phys. Rev. Lett.* **68**, 9–12 (1992).

[4] Hansmann, U. & Okamoto, Y. *J. Comp. Chem.* **14**, 1333–1338 (1993).

[5] B. Berg, U. Hansmann and T. Neuhaus, *Phys. Rev.* **B 47** (1993) 497; B. Berg, U. Hansmann and T. Neuhaus, *Z. Phys.* **B 90** (1993) 229;

[6] Berg, B.A. & Celik, T. *Phys. Rev. Lett.* **69**, 2292–2295 (1992); Berg, B.A., Celik, T. & Hansmann, U. *Europhysics Lett.* **22**, 63–68 (1993); Celik, T., Hansmann, U. & Katoot, M. *J. Stat. Phys.* **73** (1993) 775-788.

[7] A.M. Ferrenberg and R.H. Swendsen, *Phys. Rev. Lett.*, **61**, 2635 (1988); **63** , 1658(E) (1989), and references given in the erratum.

[8] Momany, F.A., McGuire, R.F., Burgess, A.W. & Scheraga, H.A. *J. Phys. Chem.* **79**, 2361–2381 (1975); Némethy, G., Pottle, M.S. & Scheraga, H.A. *J. Phys. Chem.* **87**, 1883–1887 (1983); Sipple, M.J., Némethy, G. & Scheraga, H.A. *J. Phys. Chem.* **88**, 6231–6233 (1984).

[9] Kawai, H., Okamoto, Y., Fukugita, M., Nakazawa, T. & Kikuchi, T. *Chem. Lett.* **1991**, 213–216 (1991); Okamoto, Y., Fukugita, M., Nakazawa, T. & Kawai, H. *Protein Engng* **4**, 639–647 (1991).

[10] Z. Li and H.A. Scheraga, *Proc. Natl. Aca. Sci., U.S.A.*, **84**, 6611 (1987); A. Nayeem, J. Vila, and H.A. Scheraga, *J. Comp. Chem.*, **12**, 594 (1991).

[11] Y. Okamoto, T. Kikuchi, and H. Kawai, *Chem. Lett.*, **1992**, 1275.

[12] B. von Freyberg and W. Braun, *J. Comp. Chem.*, **12**, 1065 (1991).

[13] U.H.E. Hansmann & Y. Okamoto, *Sampling Groundstate Configurations of a Peptide by Multicanonical Annealing*, preprint NWU 2/94, submitted for publication.

[14] Okamoto, Y. & Hansmann, U.,*Predicting α-helix propensities of nonpolar amino acids by multicanonical algorithm*, preprint NWU 1/94, submitted for publication.

[15] Chakrabartty, A. & Baldwin, R.L. in *Protein Folding: In Vivo and In Vitro* (eds Cleland, J. & King, J.) 166–177 (ACS Press, Washington, D.C., 1993).

[16] Padamanabhan, S., Marqusee, S., Ridgeway, T., Laue, T.M. & Baldwin, R.L. *Nature* **344**, 268–270 (1990); Lyu, P.C., Liff, M.I., Marky, L.A. & Kallenbach, N.R. *Science* **250**, 669–673 (1990); Chakrabartty, A., Schellman, J.A. & Baldwin, R.L. *Nature* **351**, 586–588 (1991); Gans, P.J., Lyu, P.C., Manning, M.C., Woody, R.W. & Kallenback, N.R. *Biopolymers* **31**, 1605–1614 (1991); Stellwagen, E., Park, S.-H., Shalango, W. & Jain, A. *Biopolymers* **32**, 1193–1200 (1992).

[17] Sueki, M.,Lee, S.,Power, S.P.,Denton, J.B.,Konishi, Y. and Scheraga, H.A. *Macromolecules* **17**, 148-155 (1984).

[18] Zimm, B.H. & Bragg, J.K. *J. Chem. Phys.* **31**, 526–535 (1959).