united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
**abdus salam**
international centre for theoretical physics

SMR/1499 - 25

**INTERNATIONAL WORKSHOP ON PROTEOMICS:**
**PROTEIN STRUCTURE, FUNCTION AND INTERACTIONS**
(5 - 16 May 2003)

---

# " Multicanonical Method and Generalized Ensemble Simulations of Short Peptides"

presented by:

**T. Celik**
Hacettepe University
Turkey

# Multicanonical Method and Generalized Ensemble Simulations of Short Peptides

**Tarık Çelik**
*Department of Physics Engineering*
*Hacettepe University, Ankara, Turkey*
*e-mail: tcelik@hacettepe.edu.tr*

**Abstract**

The conformation space of proteins and peptides presents a complex energy profile consisting of a tremendous number of local minima separated by energy barriers. Because of energy barriers, conventional simulations in the canonical ensemble are of little use, they tend to get trapped in states of these energy local minima. Multicanonical method overcomes this difficulty by performing a random walk in energy space and samples much wider phase space than by conventional methods . Following a brief discussion of the powerful multicanonical simulation method, the intermediate steps of the simulation starting from a given sequence as the input leading to the folded three dimensional structure and the minimization procedure are summarized. Attempt to design a hybrid generalized-ensemble algorithm, determination of the topographic structure of energy landscape and conformational coverage of the low energy region are presented.

Biological macromolecules such as proteins have a well defined 3D structure which is essential for their biological activity. Therefore, predicting the protein's structure by theoretical/computational methods is an important goal in structural biology. [1] The configuration space of peptide's and protein's presents a complex energy profile consisting of tremendous number of local minima; their basins of attraction were called localized microstates. The energy profile also contains larger potential energy wells defined over wide microstates (e.g., the protein's fluctuations around its averaged structure), each including many localized ones. [2] Proteins are expected to populate low energy wide microstates even at room temperature, while peptides might also populate relatively higher energy microstates. The most stable wide microstate corresponds to the native structure.

Because of energy barriers, the commonly used thermodynamic simulation techniques, such as the Metropolis Monte Carlo (MC) [3] and molecular dynamics (MD) [4] are not very efficient in sampling while the system may occur to be trapped in a basin. Therefore, developing simulation methods that lead to an efficient crossing of the energy barriers has been a long standing challenge.

The trapping problem of the MC and MD methods can be alleviated to a large extent, by the multicanonical MC method (MUCA) of Berg and collaborators, [5, 6, 7] which was applied initially to lattice spin models and its relevance for complex systems was first noticed in Ref.[6]. Application of MUCA to peptides was pioneered by Hansmann and Okamoto [8] and followed by others; [9] simulations of protein folding with MUCA and related generalized ensemble methods are reviewed in Refs.[10] and [11].

Details about the implementation of MUCA are given in Ref.[12]. Here we only provide a very brief description of the pro cess. The MUCA weights are a step function of the energy

$$w_i(\mathbf{x}) = \exp(-b_i E_{\mathbf{x}} + a_i) \quad \text{for} \quad E_{i-1} < E_{\mathbf{x}} \leq E_i \tag{1}$$

where the $b_i$ are inverse microcanonical temperatures, $b_i = (k_B T_i)^{-1}$, and the $a_i$ are related to microcanonical free energies. The $a_i$ are not independent, but follow from the $b_i$. For the determination of the $b_i$ we use the recursion in its extension to continuum peptides [12]. It relies on $m = 1, 2, \ldots$ short runs with weights determined by $b_i^{m-1}$ and the iteration from $m-1$ to $m$ is

$$b_{i-1}^m = b_{i-1}^{m-1} + \hat{g}_{i-1}^m \ln[H_{i-1}^m / H_i^m] / \triangle E_i \ . \tag{2}$$

Here the $H_i^m$ are (not yet used) energy histograms for the range $E_{i-1} \leq E \leq E_i$ and the statistical factor $\hat{g}_{i-1}^m$ incorporates information about all runs up to $m$. In particular, $\hat{g}_{i-1}^m$ is zero if either $H_{i-1}^m$ or $H_i^m$ is zero, such that the proper limit of $\hat{g}_{i-1}^m \ln[H_{i-1}^m / H_i^m]$ is also zero in that situation.

As a showcase, we have first modeled the pentapeptide Met-enkephalin (Tyr-Gly-Gly-Phe-Met) by the ECEPP/2 potential [13], which assumes a rigid geometry (i.e. constant bond lengths and angles), and is based on non-bonded, Lennard-Jones, torsional, hydrogen-bond, and electrostatic potentials, where the dielectric constant is $\epsilon = 2$.

Following the MUCA test runs at relatively high temperatures which enabled us to determine the required energy ranges, energy range was divided into 31 bins of 1 kcal/mol each, covering the range $[20, -11]$ kcal/mol. At each update step, a trial conformation was obtained by changing *one* dihedral angle at random within the range $[-180^o; 180^o]$, followed by the Metropolis test and an update of the suitable histogram. The weights were built after $m = 100$ recursions during a long *single* simulation, where the parameters $b_i$ and $a_i$ were iterated every 5000 sweeps. From the production run, canonical ensemble expectation values of thermodynamic quantities were obtained by re-weighting [14], e.g.

$$E(T) = \frac{\sum_t E_t \, \exp(-\beta E_t + b_i E_t - a_i)}{\sum_t \exp(-\beta E_t + b_i E_t - a_i)} \tag{3}$$

where each subscript is $i = i(t)$ such that $E_{i-1} \leq E_t < E_i$, gives the canonically re-weighted energy $E$ as a function of $T$.

The lowest energy conformation (our suspected GEM) was found at $E = -10.75$ kcal/mol. Here we define, following Hansmann et. al. [15], an order parameter (OP)

$$OP = 1 - \frac{1}{90 \ n_F} \sum_{i=1}^{n_F} |\alpha_i^{(t)} - \alpha_i^{(RS)}| \ , \tag{4}$$

where $\alpha_i^{(RS)}$ ve $\alpha_i^{(t)}$ are the dihedral angles of the reference state (which is taken as GEM) and of the considered configuration, respectively. The difference $\alpha_i^{(t)} - \alpha_i^{(RS)}$ is always in the interval $[-180°, 180°]$, which in turn gives for peptides

$$0 \leq \quad < OP >_T \quad \leq 1 \tag{5}$$

Figure.1 shows the energy surface obtained by the multicanonical simulation run of one million sweep plotted against energy and the order parameter [16]. Here, we would like to point out that the utilized data is obtained by sampling of the conformational space and no minimization procedure is applied. At high temperatures, where the peptide is in the random coil state, the energy surface looks as one gaussian-like peak centered around the value of the order parameter $OP \sim 0.3$. When the temperature is lowered, first a transition from the state of random coil to globular structure is expected. In Figure.2 we show the same energy surface of Fig.1(b) by grouping the conformations of 1 kcal/mol interval in energy. Curve A denotes the energy interval $-1$ kcal/mol $\leq E \leq 0$ kcal/mol, which corresponds after re-weighting to the temperature interval 315 K $\leq T_a \leq$ 330 K. At this temperature, the energy surface starts deviating from a smooth surface and develops a shoulder. We identify this temperature as the starting of forming a structure rather than a random coil. Further down in energy (temperature), the newly forming branch of the energy surface becomes more populated. At the temperature 215 K $\leq T_b \leq$ 230 K denoted by the curve B, the energy surface displays a typical structure bifurcating into two branches of almost equal height. From there on, the branch having larger values of the order parameter wins and more conformations populate that section of the conformational space. Our estimate of $T_a$ and $T_b$ from the topographic structure of the energy surface of Met-enkephalin are very close to the values of the collapse temperature $T_\theta = 295 \pm 20$ K and the folding temperature $T_f = 230 \pm 30$ K, respectively, determined by Hansmann et al [17]. We observe a third temperature denoted by the curve C in Fig.2 where the glassy behavior sets in and many valley structure of the energy surface become clearly pronounced. For our simulated peptide sample Met-enkephalin, this temperature is in the range 155 K $\leq T_c \leq$ 185 K. Below this temperature, the energy surface is made of valleys which are well separated. The valley at the far-out end of the order parameter scale having the conformations with the value of the order parameter in the range $0.98 \leq OP \leq 1$ contains the global energy minimum (GEM), respect to which the order parameter is evaluated. The temperature $T_c$ seems to correspond to the glass transition temperature estimate of $T_g = 180 \pm 30$ K, which value is based on the fractal dimension estimates. [18] In Fig.3 we plotted all the conformations found with energy $E \leq -10.5$ kcal/mol with respect to the order parameter. Their number is 3587 conformations in one production run of one million sweeps. As clearly seen from Fig.3 that the conformations in this energy range are localized in one of the four valleys, which are identified by the value of their order parameter $OP \sim 0.80, 0.87, 0.92$ and 0.98. The conformations in the neighborhood of the GEM take place within the same wide microstate of the GEM but they are grouped into local microstates, each of which are one of the above mentioned valleys. The small differences in values of OP comes from the differences in side-chain angles. We observe no conformation anywhere outside the definite valleys when the energy is less than about 1 kcal/mol above the GEM.

The number of conformations found in energy bins of 1 kcal/mol, which were plotted in Fig.2, appear in Table I. The lowest bin is 0.75 kcal/mol and includes the GEM. The table

displays the distribution of sampled conformations according to the order parameter values, namely the distribution with respect to how far they are in configuration space from the global energy minimum. We also included in Table I the same distribution obtained in our simulation of Met-enkephalin for the case of variable peptide-bond angles $\omega$.

Next we have utilized the energy landscape paving (ELP) algorithm recently introduced by Wille and Hansmann [19], which is designed to deform the energy surface to escape local minima as well as to direct the search towards the unexplored regions. ELP samples the significant local minima and the transition states without generating too many unimportant conformations.

The central feature of ELP is to perform Monte Carlo (MC) simulation with a modified energy expression which enables to keep the search away from the already explored regions. The weight for a state is taken as

$$w(\tilde{E}) = e^{-\tilde{E}/k_B T}, \tag{6}$$

where $T$ denotes temperature and $\tilde{E}$ is the following replacement of the energy $E$:

$$\tilde{E} = E + f(H(q,t)) \tag{7}$$

where $f(H(q,t))$ is a function of the histogram $H(q,t)$ in a chosen "order parameter" $q$. In order to test the efficiency of ELP, we adopted the simplest case and used the potential energy itself as an order parameter and the weight is generated by $\tilde{E} = E + H(E,t)$ where $H(E,t)$ is the histogram in energy. The histogram is updated at each MC step, hence the "time" dependence of $H(E,t)$, and normalized over the number of sweeps.

We have examined the performance of the ELP procedures in studying the low energy conformations by applying to a linear heptapeptide with bulky side chains, deltorphin (also known as dermenkephalin) (H-Tyr$^1$-D-Met$^2$-Phe$^3$-His$^4$-Leu$^5$-Met$^6$-Asp$^7$-$NH_2$) with 36 dihedral angles [20]. The characteristic behavior of ELP methods is shown in Fig.4 which is the time series of $5 \times 10^5$ sweeps for the ELP simulation of deltorphin at $T = 50K$. For comparison, the standart Monte Carlo simulation at $T = 50K$ shows a time series confined to rather narrow range of energy -33 kcal/mol $\leq E \leq$ -28 kcal/mol. The ELP time series has the typical time-dependent feature of continously extending the covered range of energy. After long enough time elapsed, the time series becomes like the one achieved by multicanonical simulation. Another important feature one sees from the time series is that the simulation gets trapped in a local minima in the energy landscape, spends some time there to built histogram, then escapes to search other regions. But when the search hits the same pre-visited minima, it does not get trapped, almost immediately leaves and freely searches till gets trapped in another basin with lower energy. In a stepwise fashion, the search tries to reach the global minima and afterwards the stepwise entrapments disappear and the time series looks like the typical time series of multicanonical simulation (e.g. after 400000 sweeps in Fig.4).

In Table II we have shown the number of conformations found in energy bins of 1 kcal/mol above the GEM with ELP and MUCA methods. First part of the table (part A) shows the result of the first $5 \times 10^5$ steps and the second part (part B) is for total of $10^6$ steps for both

methods. Because the efficiency of ELP strongly depends on the temperature, we carried out two different ELP simulations each of $10^6$ steps at temperatures $T = 50K$ and $250K$. From the Table II it is obvious that ELP sampled more low temperature conformations compared to MUCA, except for the second and the third lowest energy bins where the MUCA simulation encountered an entrapment in a wide macrostate. In searching the low-energy conformations, ELP search at $T = 50K$ is clearly more effective then the one at higher temperature. The lowest energy conformation (our suspected GEM) is $E = -44.1058\,\mathrm{kcal/mol}$.

A very good coverage of the lowest energy bins of deltorphin is provided by the energy landscape paving approach. Extensively long computer time would be needed in MUCA simulation and the probability weight factors have to be determined by iterations of trial simulations, while ELP simulation is found much simpler to implant and more effective in sampling the lowest energy region of the conformational space.

Further improvement of our simulational studies had been achived by our recently suggested fast and effective conformational search method [21], which combines the features of the ELP and the Monte Carlo Minimization (MCM) method developed by Li and Scheraga [22]. In order to design a search method especially effective at the low-energy part of the conformation space, the hope was while utilizing the ELP method to overcome the energy barriers, we simultaneously benefit from the MCM technique to lower the energy. Namely, we have implanted an MCM step in between the two updates of the dihedral angles in ELP algorithm, with the MCM protocal adopted from the Ref. [23] We have tested the performance of this procedure in studying the low energy conformations of deltorphin, again modeled by the ECEPP/2 potential.

Fig.5 displays the time sequence of the first 80000 sweeps for the simulation of deltorphin by utilizing our hybrid algorithm. The first 10000 sweeps is typical of the MCM algorithm, which is directed to find out the lowest energy state. Actually the search succeeds to visit the global energy minimum (GEM) with $E = -44.1058$ kcal/mol by the time it reaches to 10000 iterations. At that point, if we had utilized the standard MCM algorithm, the time sequence would continue as a straight horizontal line indicating that the lowest energy state had been captured and the further minimization would not introduce any change. The next step in Fig.5, namely the portion covering the sweeps from 10000 to 30000, seems to be the regime where the characteristic feature of ELP comes into play. By building up the histogram, it rescues the system from getting trapped at the lowest energy state and pushes the search to explore other regions of conformation space. After 30000 sweeps in Fig.5, we observe a third regime of an effective and fast search, confined to the low energy corner of the conformation space with energies between the GEM and about 5 kcal/mol above the GEM. With this algorithm, the simulation does not repeatedly spend time in searching the unwanted higher energy states. Our hybrid algorithm adopts the typical search patterns of ELP and the Multicanonical algorithms, but instead of visiting the whole range of available energies, it conducts an effective search of the desired low energy region.

In order to classify the microstates according to the potential wells they belong around thermodynamically stable different structures, we have adopted a variance criterion whereby two structures are considered different if at least two corresponding dihedral angles differ by $2^o$

or more. The lowest energy conformation found in our simulation with energy $E = -44.1058$ kcal/mol is taken as the GEM and the number of different structures found in energy bins of 1 kcal/mol above the GEM are listed in Table III. We see that the conformational coverage of the low energy region is quite good and all the bins upto about $E = -40.00$ kcal/mol are almost equally populated by around 12000 entries. The results show that this hybrid algorithm used here is more effective in sampling the lowest energy region of the conformational space, faster in reaching the global energy minimum and the significant low energy conformations pertaining the GEM and saves computer time.

In summary, we have shown here that the multicanonical method and the generalized-ensemble simulations of short peptides can provide a good coverage of the conformational space, especialy in the lowest energy region, a good sampling of the conformations pertaining the basin of GEM. Concerning computer time, MUCA simulations required a 12 h production run for Met-enkephalin and about a week for deltorphin on a DEC-Alpha 433 workstation, while the use of our proposed hybrid algorithm made the simulation of deltorphin possible in less than one day.

I would like to take this opportunity to thank the organizers of the International Workshop on Proteomics held at ICTP on May 2003, where this talk was presented.

# References

[1] M. Vásquez, G. Némethy, H.A. Scheraga, Chem Rev 94, 2183 (1994).

[2] H. Meirovitch, E. Meirovitch, J Phys Chem 100, 5123 (1996); C. Baysal, H. Meirovitch, Biopolymers 50, 329 (1999).

[3] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J Chem Phys 21, 1087 (1953).

[4] B.J. Alder, T.E. Wainwright, J Chem Phys 27, 1208 (1957); J.A. McCammon, B.R. Gelin, M. Karplus, Nature 267, 585 (1977).

[5] B.A. Berg, T. Neuhaus, Phys Lett. B 267, 249 (1991).

[6] B.A. Berg, T. Çelik, Phys Rev Lett 69, 2292 (1992).

[7] B.A. Berg, Fields Institute Communications 28, 1 (2000).

[8] U.H.E. Hansmann, Y. Okamoto, J Comput Chem 14, 1333 (1993).

[9] M.-H. Hao, H.A. Scheraga, J Phys Chem 98, 4940 (1994); J Phys Chem 98, 9882 (1994); A. Kolinski, W. Galazka, Skolnick, J. Proteins 26, 271 (1996); J. Higo, N. Nakajima, H. Shirai, A. Kidera, H. Nakamura, J Comput Chem 18 ,2086 (1997).

[10] U.H.E. Hansmann, Y. Okamoto, Ann. Rev. Comp. Physics 5, 129 (1999).

[11] A. Mitsutake, Y. Sugita, Y. Okamoto, Biopolymers (Peptide Science) 60, 96 (2001).

Table 1: Number of conformations in energy bins of 1 kcal/mol.

| ENERGY | OVERLAP | | | | TOTAL |
|---|---|---|---|---|---|
| Fix $\omega$ | 1.0-0.9 | 0.9-0.8 | 0.8-0.7 | 0.7-0.6 | CONF. |
| -10.75 to -10.0 | 3282 | 3935 | 3073 | 2779 | 15327 |
| -10.0 to -9.0 | 1001 | 3530 | 4925 | 4475 | 28088 |
| -9.0 to -8.0 | 467 | 2332 | 4003 | 3979 | 26220 |
| -8.0 to -7.0 | 190 | 1460 | 3150 | 3488 | 24139 |
| -7.0 to -6.0 | 90 | 897 | 2515 | 3290 | 22497 |
| Variable $\omega$ | | | | | |
| -12.21 to -12.0 | 23 | 25 | - | - | 48 |
| -12.0 to -11.0 | 6380 | 7568 | 302 | 197 | 14457 |
| -11.0 to -10.0 | 7600 | 21199 | 4775 | 2784 | 37107 |
| -10.0 to -9.0 | 2700 | 9956 | 3959 | 3456 | 28430 |
| -9.0 to -8.0 | 600 | 3107 | 2390 | 3137 | 2644 |


[12] F. Yaşar, T. Çelik, B.A. Berg, H. Meirovitch, J Comp Chem 21, 1251 (2000).

[13] F.A. Momany, R.F. McGuire, A.W. Burgess, H.A. Scheraga, J Phys Chem 79, 2361 (1975); M.J. Sippl, G. Némethy, H.A. Scheraga, J Phys Chem 88, 6231 (1984).

[14] A.M. Ferrenberg, R.H. Swendsen, Phys Rev Lett 61, 2635 (1988); Ibid 63, 1658 (1989).

[15] U.H.E. Hansmann, Y. Okamoto, J.N. Onuchic, Proteins 34, 472 (1999); U.H.E. Hansmann, J.N. Onuchic, J. Chem. Phys. 115, 1601 (2001).

[16] H. Arkın and T. Çelik, Int. J. Mod. Phys. C 14, 113 (2003).

[17] U.H.E. Hansmann, M. Masuya, Y. Okamoto, Proc. Natl. Acad. Sci. U.S.A. 94, 10652 (1997).

[18] D.A. Lidar, D. Thirumalai, R. Elber, R.B. Gerber, Phys. Rev. E 59, 2231 (1999); N.A. Alves and U.H.E. Hansmann, cond-mat/0001195.

[19] L.T. Wille and U.H.E. Hansmann, Phys. Rev. Lett. 88, 068105 (2002).

[20] H. Arkın and T. Çelik, Eur. Phys. J. B 30, 577 (2002).

[21] H. Arkın and T. Çelik, Int. J. Mod. Phys. C 14, xxx (2003).

[22] (a) Z. Li and H.A. Scheraga, Proc. Natl. Acad. Sci. USA, 84, 6611 (1987); (b) J. Mol. Struct. (Thechem.), 179, 333 (1988).

[23] H. Meirovitch, E. Meirovitch, J. Compt. Chem 18, 240 (1997).

Table 2: Number of conformations in energy bins of 1.0 kcal/mol above $E = -44.11$ kcal/mol as obtained by the MUCA and the ELP Methods. Part A shows the result of the first $5 \times 10^5$ sweeps and part B is for total of $10^6$ sweeps.

| | Energy(Kcal/mol) | MUCA | ELP T=50K | ELP T=250K |
|---|---|---|---|---|
| A | -44.11 to -43.11 | 1207 | 10770 | - |
| | -43.11 to -42.11 | 20116 | 13825 | 3 |
| | -42.11 to -41.11 | 25225 | 14023 | 2845 |
| | -41.11 to -40.11 | 7506 | 13863 | 9986 |
| | -40.11 to -39.11 | 3838 | 20362 | 12880 |
| | -39.11 to -38.11 | 4729 | 20449 | 14503 |
| B | -44.11 to -43.11 | 2674 | 27729 | 10843 |
| | -43.11 to -42.11 | 45350 | 30238 | 20189 |
| | -42.11 to -41.11 | 46091 | 32962 | 22656 |
| | -41.11 to -40.11 | 12188 | 33044 | 24125 |
| | -40.11 to -39.11 | 8485 | 33192 | 25245 |
| | -39.11 to -38.11 | 10983 | 33081 | 26039 |

Table 3: Number of significantly different structures in energy bins of 1 kcal/mol above $E = -44.11$ kcal/mol as obtained by the new designed hibrid algorithm. The number of different conformations are classified ac cording to the overlap parameter. The results of only 80000 sweeps are presented in the table.

| ENERGY | OVERLAP | | | | TOTAL |
|---|---|---|---|---|---|
| | 1.0 - 0.9 | 0.9 - 0.8 | 0.8 - 0.7 | 0.7 - 0.6 | CONF. |
| -44.11 to -44.00 | 1535 | 5101 | 5171 | 624 | 12431 |
| -44.00 to -43.00 | 588 | 5539 | 5210 | 921 | 12298 |
| -43.00 to -42.00 | 29 | 1284 | 5701 | 4425 | 12004 |
| -42.00 to -41.00 | - | 245 | 2041 | 4152 | 11462 |
| -41.00 to -40.00 | 1 | 116 | 1519 | 3944 | 11193 |
| -40.00 to -39.00 | - | 7 | 685 | 3204 | 10611 |
| -39.00 to -38.00 | - | 21 | 367 | 2905 | 9787 |

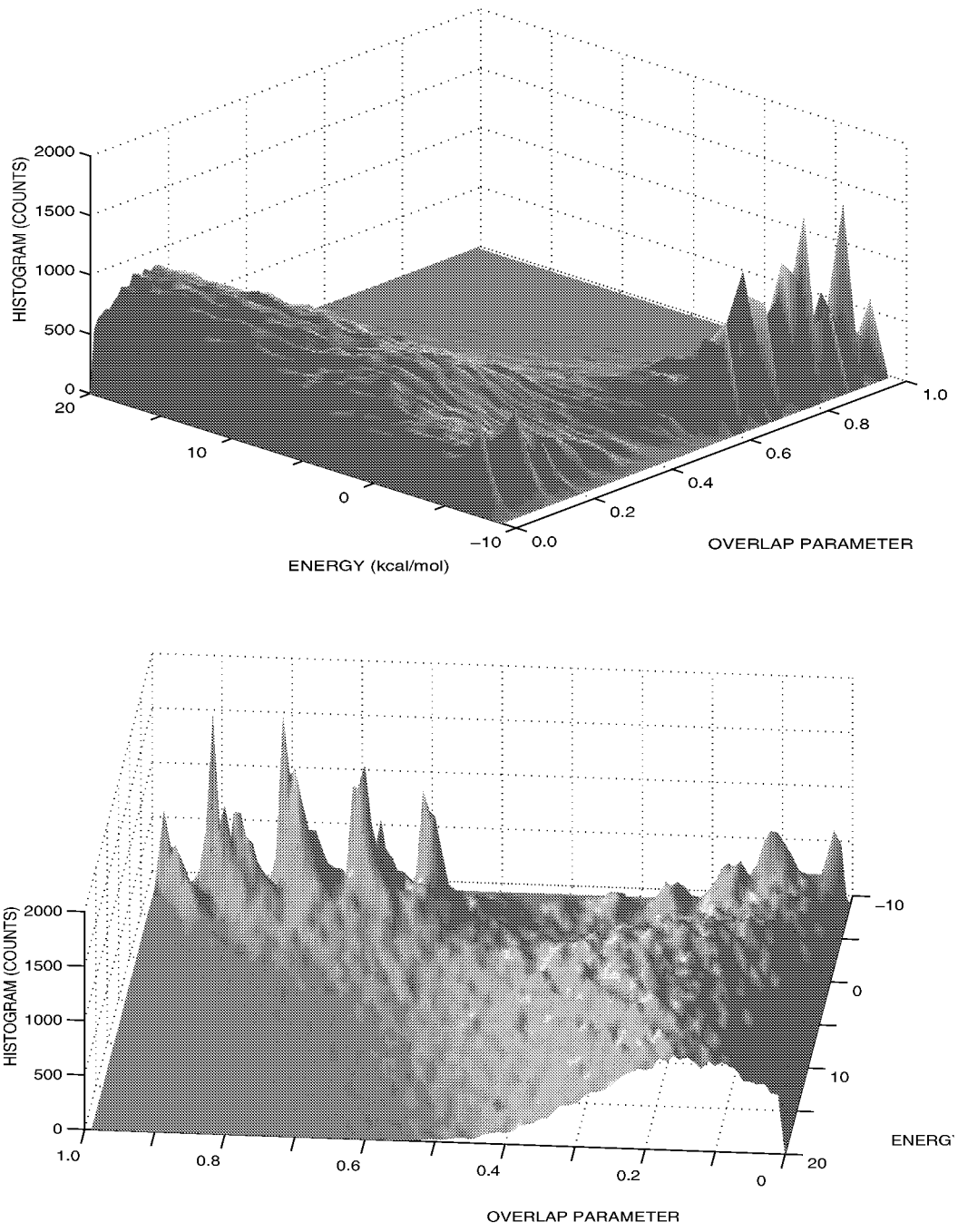


Figure 1: Energy surface in configuration space of Met-enkephalin viewed from different angles.

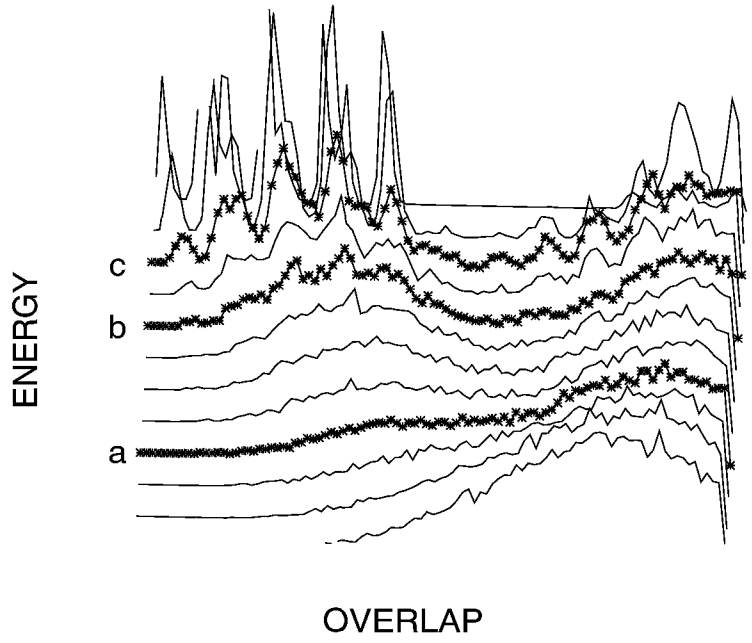


Figure 2: Same as Fig.1(b), plotted by grouping the conformations of 1 kcal/mol interval in energy.

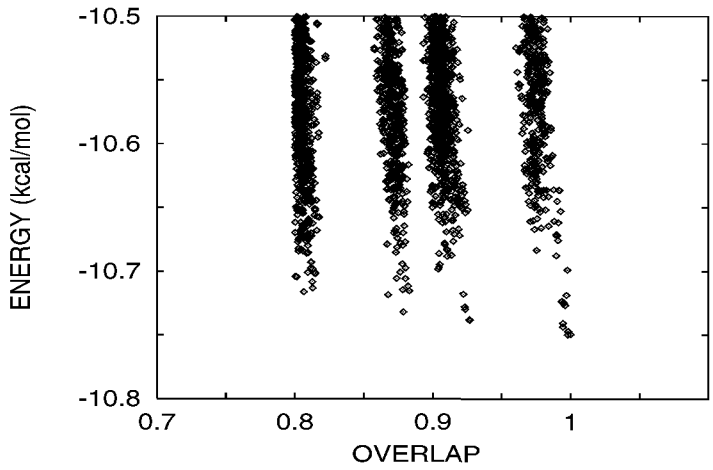


Figure 3: Distribution of microstates with $E \leq -10.5$ kcal/mol with respect to the overlap parameter.

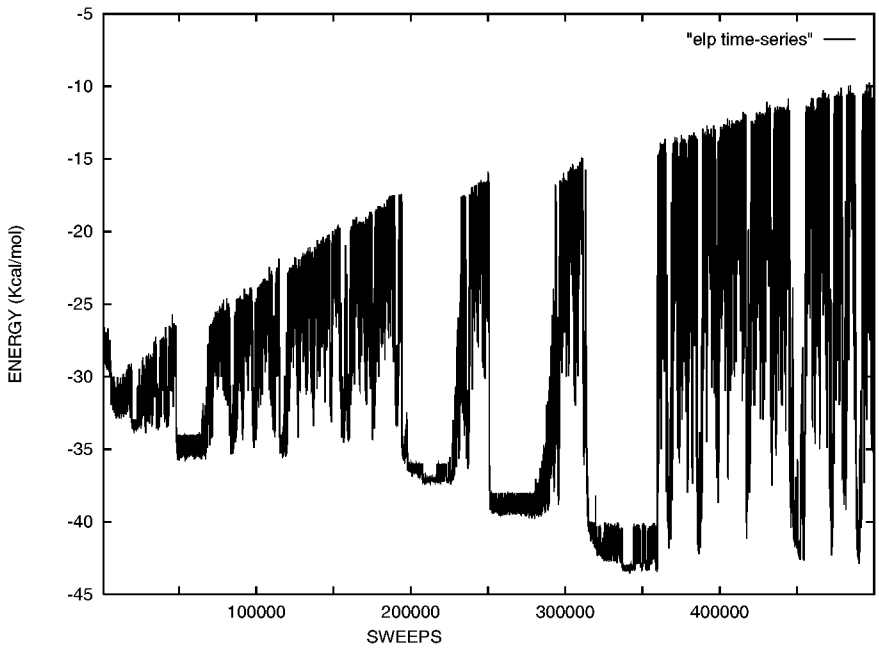


Figure 4: Time series of the energy landscape paving simulation of deltorphin. For comparison, the standart Monte Carlo simulation at $T = 50K$ is performed and a time series fluctuating within a ra ther narrow range of energy -33 kcal/mol $\leq E \leq$ -28 kcal/mol is obtained. MC time ser ies is not plotted, otherwise the figure becomes confusing unless presented in color.

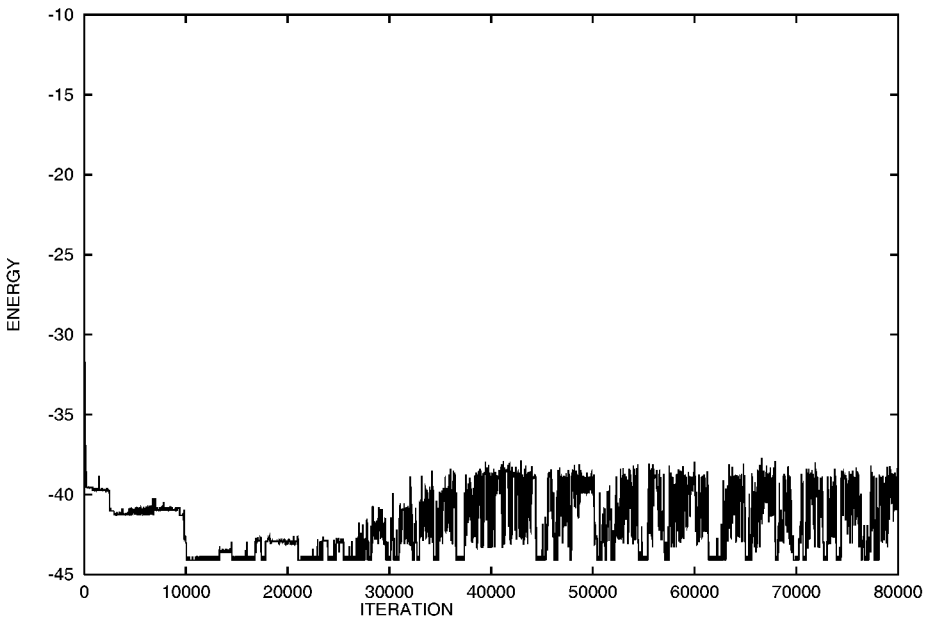


Figure 5: Time series of algorithm used in this work.