Multi-canonical algorithms for folding processes

B.Berg[a], G. La Penna[b], V. Minicozzi[c], S.Morante[d,e] and G.C. Rossi[d,f]

[a]Department of Physics and Supercomputer, Florida State University, Tallahassee, USA.

[b]CNR-ISMac, Genova, Italy.

[c]Max Planck Institute of Colloids and Interfaces, Golm, Germany.

[d]Department of Physics, University of Rome "Tor Vergata", Roma, Italy. e-mail: morante@roma2.infn.it

[e]INFM, Unitá di Roma 2, Roma, Italy.

[f]INFN, Sezione di Roma 2, Roma, Italy.

Abstract: We present a variant of the Multi-canonical Monte Carlo method in which the proposed configuration for the Metropolis test is generated by performing few steps of Molecular Dynamics. The proposed strategy makes possible to deal with numerical simulation of fully flexible chains of bonded monomers and in perspective to model the presence of the solvent at the fundamental atomic level. By realizing a sort of (unbiased) random walk in energy, the algorithm allows the system to overcome high energy barriers, thus enabling a good exploration of phase space even in presence of a strongly corrugated configuration profile. All the complications of the phase space structure of the system are taken into account by the successive re-weighting procedure. The numerical approach we present here is an important preliminary step in the direction of simulating protein folding in realistic conditions.

Keywords: *Multi-canonical Monte Carlo algorithm; Polymer chains; Folding; Modeling*

1. INTRODUCTION

Rather accurate descriptions of systems of biological interest can be constructed today by developing more or less sophisticated micro-scopic models that, however, in most cases, can only be dealt with numerically. For biological systems it appears of the utmost importance both in applications and from a more theoretical point of view to explore their dynamic and thermo-dynamic behavior in different physico-chemical conditions. This was successfully done in many cases by exploiting some of the most modern numerical techniques, like those that are based on Molecular Dynamics (MD) (Allen and Tildesley, 1990) and Monte Carlo (MC) (Rothe, 1992) methods, or by making use of more speculative approaches inspired by stochastic equations of the Langevin type (Parisi and Wu, 1981).

Biological objects of special importance are the proteins. Proteins are linear polymers having the 20 naturally occurring amino-acids (a.a.'s) as monomers. Chains smaller than a few tens of a.a.'s are called peptides. The biological functionality of a protein crucially depends on its correct folding. Mis-folding is, in fact, known to lead to malfunctioning and in many cases to severe pathologies (e.g., Creutzfeldt-Jacobs disease (Prusiner, 1997) human variant of BSE, Alzheimer disease (Selkoe, 2001), cystic fibrosis (Massiah et al., 1999).

The challenge computational biology is facing today is to predict the folded configuration of a protein or a peptide solely from the knowledge of their a.a. linear composition.

For a number of years the emphasis of numerical investigations was on developing algorithms of the simulated annealing type aimed at finding the global minimum of the potential (or free) energy of the system (see, for instance, Morante and Parisi, 1991). The major difficulty encountered by these approaches is in the existence of a large (actually exponentially large) number of local conformational minima, in which numerical searching algorithms get easily stuck "forever". Similar problems emerge when the configuration space is explored by MD (*micro-canonical ensemble*) or by standard MC (*canonical ensemble*) simulations, with a consequent inadequate exploration of the system configuration space.

To deal with these difficulties two strategies have been developed. The first one is very crude and consists in drastically reducing the number of degrees of freedom (d.o.f.) of the system, while keeping what are believed to be its essential features. This approach was followed with

success in Iori et al., 1991, where it was shown that folding is a generic property of a sufficiently random hetero-polymer.

The second approach relies on the use of the Multi-canonical (MUCA) algorithm (Berg, 2000; Mitsutake et al., 2001) and variants thereof (simulated tempering, replica-exchange, etc.). The method realizes a sort of (unbiased) random walk in energy space, thus allowing the system to overcome in principle any energy barrier. All the complications of the phase space structure are taken into account by the successive re-weighting step. The effectiveness of the method for protein folding has been extensively studied in recent years (Mitsutake et al., 2001) and shown to be fairly good at least for not too long peptides (few tens of a.a. residues). Another appealing feature of the MUCA approach is that from a single (adequately long) run one can recover the whole thermodynamics of the system as a function of the temperature, while within the standard MC scheme one would need a new simulation for each required value of the temperature.

The idea of employing the MUCA strategy to protein folding is due to Hansmann and Okamoto, 1993; Hao and Sheraga, 1994. Since then a lot of work has been done. In particular it has been possible to reproduce the correct folded structure of Poly-alanine$_{10}$ into its α-helix coil (Hansmann and Okamoto, 1999) and of Met-enkephalin$_5$ in its peculiar temperature dependent conformations (Mitsutake et al., 2001; Sugita and Okamoto, 1999). We wish to note that for short peptides, like Poly-alanine, simulated annealing is still a competitive alternative (Morante and Parisi, 1991).

A limitation of the existing MUCA algorithms is that covalent bonds and angles are kept fixed during the updating to bring the number of d.o.f. of the system to a "workable" size. This approximation prevents to model the solvent at the fundamental atomic level. The obvious step to cope with this problem is to introduce full flexibility along the chain. The method we describe in this paper goes in this direction and the preliminary results that we obtain are rather encouraging.

2. THE MULTI-CANONICAL HYBRID MONTE CARLO ALGORITHM

The novelty of the approach we present here lies in the fact that in order to deal with the computational problems posed by the full flexibility of the chain we use MD to generate the configurations that will undergo the Metropolis test (see below). Before explain our approach, we briefly recall what the MC method is about.

2.1. The canonical Monte Carlo method

The canonical MC method is a strategy developed in Statistical Mechanics to compute the partition function of a system at the temperature $T = 1/k_B\beta$ (i.e. endowed with Boltzmann probability distribution) and thermal averages of the type

$$\langle A \rangle = \frac{\prod_{i=1,N} \int d^3 p_i d^3 q_i \, A[\{q,p\}] e^{-\beta H[\{q,p\}]}}{\prod_{i=1,N} \int d^3 p_i d^3 q_i e^{-\beta H[\{q,p\}]}} \quad (1)$$

where N is the number of elementary constituents of the system. Note that, if A only depends on the coordinates, $\{q\}$, the dependence on momenta completely drops out from eq.(1).

The key ingredient of the MC method is the Metropolis acceptance test by means of which a Markov sequence of system configurations, $C_j = C[\{q,p\}_j]$, is collected with probability distribution

$$\begin{aligned} &\wp_{MC}[E_j] \propto n(E_j) e^{-\beta E_j} \\ &E_j = H[\{q,p\}_j] = K[\{p\}_j] + U[\{q\}_j] \end{aligned} \quad (2)$$

where K and U are the kinetic and potential energy respectively, and

$$\begin{aligned} &n(E) = \prod_{i=1,N} \int d^3 p_i d^3 q_i \; \delta(H[\{q,p\}] - E) \\ &\equiv \exp[S(E)/k_B] \end{aligned} \quad (3)$$

is the density of states with *S(E)* the entropy of the system. The Metropolis test is easily realized in the following way. Given the configuration *C*, a new configuration, *C'*, is generated by some (reversible) prescription and accepted with probability

$$P_{MC}[C \to C'] = \min(1, e^{-\beta(H[C'] - H[C])}) \quad (4)$$

Since this procedure fulfils the detailed balance principle, one can prove that the resulting effective probability distribution will be exactly that of eq.(2). We recall that, if the quantity of which we want to evaluate the thermal average does mot depend on momenta, one can use in eq.(4) the potential energy instead of the full Hamiltonian to generate the required Markov sequence.

Given the set of collected configurations $\{C_j, j=1,...,N_{MC}\}$ (with N_{MC} very large), the expectation value of a physical quantity A can be computed according to the simple formula

$$\langle A \rangle = \frac{1}{N_{MC}} \sum_{j=1}^{N_{MC}} A[C_j] + O\left(\frac{1}{\sqrt{N_{MC}}}\right) \quad (5)$$

2.2. The Hybrid Monte Carlo method

The MC method is very efficient for systems with near-neighbor interactions and no (or a limited amount of) frustration. Under these conditions, that are for instance met in the very important case of lattice Quantum Chromo-Dynamics, a thorough (ergodic) exploration of the system configuration space is possible.

Unfortunately in more complicated (complex) systems, like many systems of biological interest, this is almost never the case. The models by which phenomena, like immunological recognition, protein folding or docking, are described in terms of their fundamental constituents (atoms) turn out to be extremely complicated, as they are driven by short (e.g. Lennard-Jones) and long (Coulomb) range potentials with a subtle mixture of attractive and repulsive pieces. As a result the system possesses a very corrugated energy landscape with an extremely large number of nested local minima, in which the MC algorithm gets (almost) inevitably trapped, thus preventing the full exploration of the system configuration space.

Hybrid MC (HMC) method is an evolution of the straightforward MC in which the trial configurations to be subjected to the Metropolis test are constructed using MD. This modification which does not spoil the canonical nature of the algorithm (Rothe, 1992; Scalettar et al., 1986; Gottlieb et al, 1987), allows to perform collective moves, by which all d.o.f. are updated simultaneously. As a consequence, the algorithm can reach more distant configurations leading to a more efficient sampling of the system phase space.

In HMC the proposed new configuration, *C'*, is obtained from the old one, *C*, through the following steps:

1. the initial coordinates for the MD iterations are those of the current configuration, *C*;

2. the initial momenta are taken from a Maxwell distribution at a suitably chosen temperature;

3. n_{MD} (~10) MD steps are performed;

4. the final configuration, *C'*, is submitted to the Metropolis test.

The new configuration, *C'*, can be now accepted (step 5a) or refused (step 5b):

5a. if the configuration, *C'*, is accepted, the corresponding coordinates are stored and a new cycle begins from step 1.

5b. if the configuration, *C'*, is refused, the system is left in the configuration, *C*, the old coordinates are stored and a new cycle begins from step 2.

2.3. The Multi-canonical algorithm

The MUCA-MC method (Berg et al., 1991) was introduced to try to overcome the limitations of the canonical MC discussed in the Introduction. The general idea is to have the system moving freely in energy, thus avoiding remaining stuck in some local energy minimum. This can be done if one could generate configurations with flat distribution, i.e. replace the MC probability distribution (2), ensuing from the Metropolis test (4), with the MUCA probability distribution

$$\wp_{MUCA}(E) \propto \text{constant} \qquad (6)$$

In view of the previous equations, this can be achieved by replacing the Metroplois test (4) based on the Boltzmann factor *exp(-βH)*, with a new Metropolis test based on the (inverse of the) density of states, eq.(3). In fact if, given *C*, we accept the new configuration *C'* with probability

$$P_{MUCA}[C \to C'] = \min\left(1, e^{-(S[C']-S[C])/k_B}\right) \qquad (7)$$

one can prove that the principle of detailed balance is obeyed, so the effective probability distribution will be (see eq.(3))

$$\wp_{MUCA}[C] \propto n[C]\, e^{-S[C]/k_B} = \text{constant} \qquad (8)$$

The problem with this approach is that we do not know *a priori* the density of states of the system. Trying to guess it is hopeless, given the enormous complexity of the system. The only way is to perform a preliminary simulation from which an estimate, $\hat{n}[C]$, of $n[C]$, is extracted. In the literature various iterative strategies have been devised to carry out this step. An exhaustive description can be found in Berg, 2000.

It is important to note at this point that it is not necessary to have an infinitely accurate preliminary determination of $n[C]$ (which by the way if possible would render the whole procedure unnecessary). In fact, given a (reasonable) estimate, $\hat{n}[C]$, of $n[C]$, one proceeds as follows.

1. By replacing (7), configurations are generated with probability

$$\hat{P}_{MUCA}[C \to C'] = \min\left(1, \hat{n}[C]/\hat{n}[C']\right) \qquad (9)$$

2. The resulting effective configuration probability distribution will be

$$\hat{\wp}_{MUCA}[C] \propto n[C]\frac{1}{\hat{n}[C]} \qquad (10)$$

which is only approximately constant.

3. Statistical averages are evaluated using the well known re-weighting formula

$$\langle A \rangle = \frac{\sum_{j=1}^{N_{MUCA}} A[C_j]\hat{n}_{MUCA}[C_j]e^{-\beta H[C_j]}}{\sum_{j=1}^{N_{MUCA}} \hat{n}_{MUCA}[C_j]e^{-\beta H[C_j]}} + O(\frac{1}{\sqrt{N_{MUCA}}})$$

2.4. Introducing full flexibility

As already said in the Introduction, the existing MUCA algorithms keep covalent bonds and angles fixed during the updating and only dihedral (torsional) angles are left free to move. The clear advantage of this strategy is a substantial reduction of the number of d.o.f. with a consequent rather efficient exploration of the configuration space.

It must be immediately noted, however, that torsional moves are greatly disadvantageous in presence of the solvent. In fact, due to "leverage" effects, even a small modification of a rotational angle on one side of a long chain may results in a very large displacement of an atom located far away from the former. Such large displacements will almost inevitably bring some of the atoms of the chain to "collide" with some of the solvent atoms, thus leading to a configuration with a very high total energy which will never be accepted by the Metropolis test. The result is that the only acceptable moves will be those that happen to leave the whole configuration almost unchanged. Thus an adequate sampling of the configuration space will become prohibitively long.

For this reason a simulation in which the solute-solvent interaction is described at atomic level requires the introduction of full flexibility along the chain. The only alternative is to describe solvatation effects by introducing in the Hamiltonian an *ad hoc* contribution proportional to the solvent accessible surface area of the various atomic groups. This is what was done, for instance, in Mitsutake and Okamoto, 2000.

We propose to solve the problems posed by the presence of the solvent by allowing full flexibility of the system, at the same time adopting the HMC strategy described above. Full flexibility means that the chain back-bone is not rigid. Rather covalent bonds and bond angles are driven by harmonic potentials. The total interaction

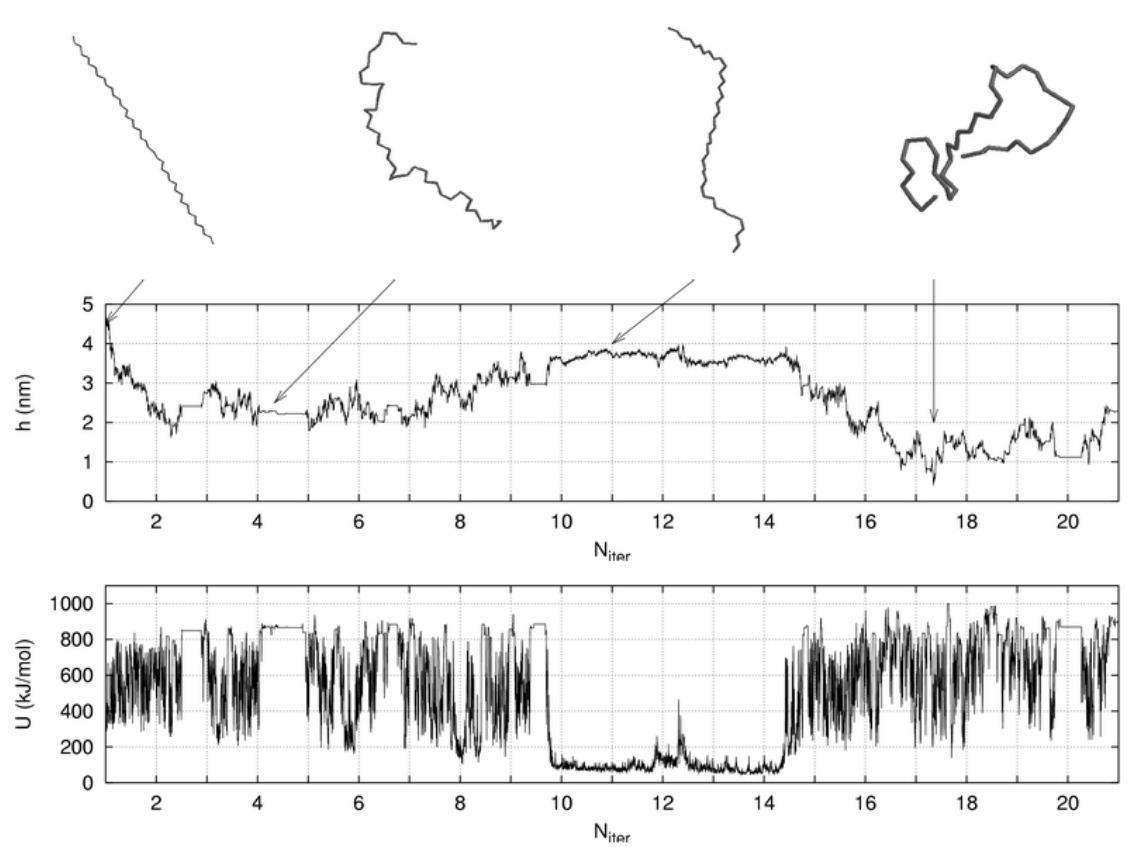

Figure 1. From top to bottom: 1) the tri-dimensional structure of the chain; 2) the history of the head-to-tail length; 3) the history of the total potential energy.

potential, U_{tot}, is then given by (Allen and Tildesley, 1990; La Penna et al., 1997)

$$U_{tot} = U_{stret} + U_{bend} + U_{tors} + U_{nb} \qquad (11)$$

where U_{stret} is the potential associated with bond length stretching, U_{bend} with bond angle bending, U_{tors} with dihedral angle twisting, while the last term is the so-called "non-bonded" potential which includes Coulomb and Lennard-Jones interaction potentials.

3. NUMERICAL SIMULATIONS

3.1. The model system

In order to estimate the efficiency of the proposed strategy we performed MUCA test simulations on the C_{40}=$CH_3(CH_2)_{38}CH_3$ alkane. We simulated the system as a polymeric chain of 40 beads (united-atom approximation). Structural and dynamical properties of model systems of this kind have been extensively studied theoretically (Flory, 1980) and are pretty well known. For this reason and because of the considerable length of the chain, we regard this system as rather interesting in view of testing the efficiency of our approach for the investigation of the significantly more complicated problem of protein folding.

3.2. Results

In Fig.1 we show a synthesis of some preliminary results we have obtained. In the lowest panel we plot the total potential energy (the interactions taken into account are those of eq.(11) with the exclusion of the non-bonded potentials) as function of the number of MUCA-MC moves. The figure clearly shows that the system oscillates among well separated energy states. This feature confirms the expectation that the algorithm is capable of exploring at large the configuration space of the system.

The lowest energy of the system (~100 KJoule) corresponds to a configuration in which the chain is almost completely extended (the so called *all-trans* configuration). This is also the starting configuration of our simulation. As seen in the figure, after few MC steps the chain rapidly evolves to higher energy configurations, then at a certain moment it goes back to a somewhat lower energy state around which it oscillates for a while, and then comes back to configurations of higher energies.

In the middle panel we plot the history of the head-to-tail distance, h, of the chain. h is a kind of measure of the level of chain entangling (Flory, 1980). The all-trans structure corresponds to an average length of about 5nm. The interesting observation is that, while the energy plot reflects something very much like to a two-states dynamics, the corresponding head-to-tail history has a smoother evolution, thus correctly reproducing the fact that many different "folded" structures correspond to very near energies (large entropy). It is also interesting that the extended (unfolded) structure is, on the contrary, almost unique as proved by the almost flat, constant value of h through the low energy region. In the top panel we draw the tri-dimensional structures assumed by the chain in correspondence of four representative values of the energy. On the very left we show the starting extended, all-trans structure. Moving to the right, the second structure corresponds to a configuration with a high energy and a relatively short head-to-tail distance, followed by a situation in which we have an extended structure with a large value of h and a relatively low energy. The shortest value of h we observed corresponds to the tri-dimensional structure drawn at the extreme right of the panel. Somewhat unexpectedly it happens to have a not too small energy.

4. CONCLUSIONS

In this note we have presented a promising variant of the existing MUCA-MC algorithms which appears to be particularly well suited for the study of long flexible chains of monomers. From our preliminary results we have fairly good indications that the modifications we have devised are apt to overcome the difficulties of more standard methods, when the latter are used for simulating systems with a large number of d.o.f. and long-range interactions. A more complete account of our investigation will be presented elsewhere (Berg et al., 2003).

5. ACKNOWLEDGEMENTS

We would like to thank the organizers of MODSIM 2003 for the opportunity offered to us to present this work.

6. REFERENCES

Allen, M.P. and D.J. Tildesley, *Computer Simulation of Liquids,* Clarendon Press, Oxford, 1990.

Berg, B.A., Introduction to multicanonical Monte Carlo simulations, *Fields Institue Communications*, 26, 1-24, 2000 and references therein.

Berg, B.A. and T. Neuhaus, Multicanonical algorithms for first order phase transitions, *Physics Letters B*, 267, 249-253, 1991.

Berg B.A. and T. Neuhaus, Multicanonical ensemble: a new approach to simulate first order phase transitions, *Physical Review Letters*, 68, 9-12, 1992.

Berg, B.A., La Penna, G., V. Minicozzi, S. Morante, G.C. Rossi, in preparation.

Flory, P.J., *Statistical Mechanics of chain molecules*, Hanser Publishers, Oxford University Press, 1980.

Gottlieb, S., W. Liu, D. Toussaint, R.L. Renken and R.L. Sugar, Chiral simmetry breaking in lattice QCD with two and four fermion flavors, *Physical Review D (Particles, fields, gravitation and cosmology)*, 35, 3972-3980, 1987.

Hansmann, U.H. and Y. Okamoto, Prediction of peptide conformation by multicanonical algorithm: new approach to the multi-minima problem, *Journal of Computational Chemistry*, 14, 1333-1338, 1993.

Hansmann, U.H. and Y. Okamoto, Finite-size scaling of helix-coil transitions in the poly-alanine studied by multicanonical simulations, *Journal of Chemical Physics*, 110, 1267-1276, 1999.

Hao, M.-H. and H.A. Sheraga, Monte Carlo simulation of a first-order transition for protein folding, *Journal of Physical Chemistry*, 98, 4940-4948, 1994.

Iori, G., E. Marinari and G. Parisi, Random self-interacting chains: a mechanism for protein folding, *Journal of Physics A*: *Mathematical and General*, 24, 5349-5362, 1991.

La Penna, G., V. Minicozzi, S. Morante, G.C. Rossi and G. Salina, Molecular dynamics with massively parallel APE computers, *Computer Physics Communications*, 106, 53-68, 1997.

Massiah, M.A., Y.H. Ko, P.L. Pedersen, and A.S. Mildvan, Cystic fibrosis transmembrane conductance regulator: solution structures of peptides based on the Phe508 region, the most common site of disease-causing DeltaF508 mutation, *Biochemistry*, 38(23), 7453-7461, 1999.

Mitsutake, A., Y. Sugita and Y. Okamoto, Generalized ensemble algorithms for molecular simulations of biopolymers, *Biopolymers (Peptide Science*), 60, 96-123, 2001.

Mitsutake, A. and Y. Okamoto, Helix-coil transition of amino-acid homo-oligomers in aqueous solution studied by multicanonical simulations, *Journal of chemical physics*, 112, 10638-10647, 2000.

Morante, S. and V. Parisi, Building structural models of peptides : a semi-automatic software, *Computer applications in the biosciences*, 7, 21-26, 1991.

Parisi, G. and Y.-S. Wu, Perturbation theory without gauge fixing, *Scientia Sinica*, 24, 483-496, 1981.

Prusiner, S.B., Prion diseases and the BSE crisis, *Science*, 278, 245-251, 1997.

Rothe, H.J., *Lattice Gauge Theories*, World Scientific Lecture Notes in Physics - Vol. 43, World Scientific, Singapore, 1992.

Scalettar, R.T., D.J. Scalapino and R.L.Sugar, New algorithm for the numerical simulation of fermions, *Physical Review B (Condensed matter and material physics)*, 34, 7911-7917, 1986.

Selkoe, D.J., Alzheimer's disease: genes, proteins, and therapy, *Physiological Reviews*, 81, 741-766, 2001.