International Journal of Modern Physics C, Vol. 0, No. 0 (1993) 000–000

HETEROPOLYMER FOLDING ON A APE-100 SUPERCOMPUTER*

GIULIA IORI
Dipartimento di Fisica and INFN, Università di Roma Tor Vergata,
Viale della Ricerca Scientifica 2, 00133 Roma (Italy)

ENZO MARINARI†
NPAC and Physics Department, Syracuse University,
Syracuse, NY 13244 (USA)

and

GIORGIO PARISI
Dipartimento di Fisica and INFN, Università di Roma La Sapienza,
P. Aldo Moro 2, 00185 Roma (Italy)

We discuss a model of protein folding, the IMP model of a random heteropolymer. We discuss the thermodynamic behavior of a linear heteropolymer in which the interaction between different monomers contains a quenched random component. We show the existence, along with the usual coil and globule phases, of a new *folded phase*, characterized by long relaxation times and by the existence of few stable states. We discuss further relevant features of the model. We describe massively parallel simulations of this model on the APE-100 supercomputer.

1. Introduction

Proteins are a fascinating subject (we refer for example to references[1−7] for various approaches to the many sides of the problem). Proteins are very elegant and multi-functional entities, large and complex on the scale of their fundamental constituents, but very simple if regarded on the scale of the structures they eventually constitute (for example cow bodies).

Protein folding is one of the essential and most interesting features. Biologically active proteins are in a folded state: a globular state with a precise shape, characteristic of the given protein. The information about folding (i.e. the $3d$ stable structure of a working protein) is contained in the *linear* sequence of the messenger RNA: there is no space for explicit coding of the $3d$ structure, which must be determined from the interaction laws of the constituent amino-acids. The given sequence of amino-acids, which eventually constitutes the working protein, is coded in the RNA: the different amino-acids have different interactions, and interact in a different way with the solvent.

*Talk given by Enzo Marinari
†also at Dipartimento di Fisica and INFN, Università di Roma *Tor Vergata*, Viale della Ricerca Scientifica 2, 00133 Roma (Italy)

Folding is certainly a complex and quite mysterious procedure. The time scales involved in the problem are very different: folding times vary a lot, but the time scale is much longer than that needed for a steepest descent to a simple minimum, and too short (obviously) for an exhaustive search of the configuration space. One or a very few allowed folded states characterize a given, biologically active, protein.

The fact that one can hope to understand some features of such a problem on the basis of first principles and of a universal behavior is attracting the attention of physicists. We want to understand what is the mechanism that allows such a crucial process to work. We want eventually to be able to build the native configuration using a physically relevant approach (for an essay in this direction see ref.[8]): in other words we want try to understand which are the relevant mechanisms (that have to be very stable and simple) used by nature in the process of folding.

Physicists are used to approaches based on the idea of *universality*: relevant mechanisms are often independent of the details of the interaction laws, and just depend on very general features of the problem (for example the symmetries of the problem). Critical phenomena, which occur in the transition between different regimes, depend only on such general features: only very specific features (like the value of the critical temperature) depend on the details of the interaction. *Protein folding* is an exquisite candidate to such an approach. It is clear to us that real proteins are the products of natural evolution and they are *not* random sequences of random interacting amino-acids. It is however extremely interesting to understand which properties proteins share with generic random heteropolymers, and on the contrary which of their properties are selected by natural evolution: such a study has to be started by investigating in details the behavior of random heteropolymers.

Much crucial progress has recently been made in studies of complex systems[9]. Starting from the specific example of amorphous materials, and soon generalized to very different situations, a whole new formalism, the mechanism of replica symmetry breaking, has lead to many new results. In recent months many results have been obtained for the behavior of membranes in random potentials[10].

Such an approach seems crucial in order to try to apply ideas concerning disordered systems to the description of protein folding. Indeed if random spin systems have their own typical features (which characterize for example phases that cannot be found in usual, non-random spin models), random membranes share some of these new features, but are in some sense different, and this difference can be quite crucial. For example it would be difficult to match the structure of states of a Sherrington-Kirkpatrick infinite range spin model (and also of a Random Energy Model, see for example the Introduction and the reprinted papers contained in ref.[9]) with what one knows about protein folding. The many completely disconnected minima in these models do not match with proteins that always appear to be in one of few allowed states.

In our original work[11] we have seen that important features that have been noticed in the approach of ref.[10] can be explicitly found during the numerical simulations of an heteropolymeric chain. We found, along with the usual coil-globule phase transition, a new *folded* phase, which seems suitable to describe protein folding as a generic phenomenon. Its features match very well many of the intriguing features of the protein folding dynamics: we have breaking of ergodicity and very long time scales, and few stable states in which the chain folds. We can relate the existence of such a phase to the presence in the system of a strong, quenched disorder.

We refer to the work of[12,13] for connections between disordered systems and protein folding.

We will here introduce the model and its (bio)-physical motivations. We will describe the phase diagram. We will give some hints about a series of further results we have obtained[15]. At last we will describe our recent simulations[14] on the APE-100 supercomputer, and describe our first results.

2. The IMP Model

2.1. *The Hamiltonian*

Let us start by defining the Hamiltonian of our model. We consider N sites of a chain (they will be identified, in the protein analogy, with *sequences* of amino-acids). Their position in *continuum* 3 dimensional space is characterized by the 3 values of the coordinates x_i^μ, where in the following Latin indices i, j,... label the n-th site of the chain, and Greek indices μ, ν,... label the 3 spatial directions (only the ones from μ on in the alphabet, since we will use α, β,... to label the copies of the chain we encounter in the course of the Monte Carlo dynamics).

We define the distance between two sites of the chain by

$$r_{i,j} \equiv \sqrt{\sum_{\mu=1}^{3}\left(x_i^\mu - x_j^\mu\right)^2} \ , \tag{1}$$

and the energy between two sites of the chain is

$$E_{i,j} \equiv \delta_{i,j+1}\ r_{i,j}^2 + \frac{R}{r_{i,j}^{12}} - \frac{A}{r_{i,j}^6} + \frac{\eta_{i,j}}{r_{i,j}^6} \ . \tag{2}$$

The harmonic term couples first neighbors on the chain. The deterministic part of the potential has the usual Lennard-Jones form. The main difference from a usual homopolymers is given by the quenched $\frac{1}{r^6}$ contribution. The quenched part of the potential has a zero expectation value (we have explicitly written an attractive deterministic contribution, which we will call the A term, in the definition of the couple energy, (2))

$$\langle \eta_{i,j} \rangle = 0 \ , \tag{3}$$

it is symmetric ($\eta_{i,j} = \eta_{j,i}$)
and has a correlation of the form

$$\langle \eta_{i,j} \; \eta_{k,l} \rangle = \epsilon \; \delta_{(i,j),(k,l)} \; , \tag{4}$$

that is non-zero only if $i = j$ and $k = l$ or if $i = l$ and $j = k$. This effective random interaction represents, in the biological picture, many different factors: the complex interactions between different groups of different amino-acids, the effect of the solvent (typically water molecules), etc.

The Hamiltonian is defined as

$$H \equiv \sum_{i=1}^{N} \sum_{j>i} E_{i,j} \; , \tag{5}$$

and the model is brought to thermal equilibrium under the Boltzmann distribution $e^{-\beta H}$, where $\beta \simeq \frac{1}{T}$. In the following we will try to reach a good understanding of the rôle of the disorder (given by the quenched random potential) and of the Lennard-Jones interaction on the chain.

The deterministic part of the potential has a simple form. The harmonic term, with a first neighbor interaction on the chain, keeps the chain together. The repulsive R term forbids the collapse of the chain, and the attractive A contribution allows the chain to fold. The choice of a Lennard-Jones form is a convenient, well understood one: other choices of the exponent are obviously possible and we tend to believe that the qualitative behavior of the model should not change as long as the potentials go to zero at infinity sufficiently fast. We could also have chosen an exponentially damped interaction, but we have not done so for practical numerical reasons.

In the absence of the random quenched term we are dealing with a *homopolymer*, and we expect the usual *coil-globule* transition. The globule state of a homopolymer has no definite shape. A quenched disorder could allow (and we will show it does) the formation of a globular phase with a definite, frozen shape: we would be dealing with a closed globule, in which the positions of the elementary parts of the chain are definite and fixed. This kind of phase (which we will call *folded* in the following) would be suitable in order to describe protein folding.

2.2. *The Chain Distance, the Dynamics and the Observable Quantities*

We use a local Monte Carlo dynamics: we propose a local updating move for a given link of the chain, and we accept or reject the proposed update with the correct probability. We always keep the acceptance ratio (i.e. the percentage of accepted updates) to 50%. According to popular belief such a choice nearly optimizes the efficiency of the simulation.

In order to understand the structure of the equilibrium states of the model (stable and meta-stable states), we want to use the concept of overlap. The physical interpretation of the replica approach suggests that we should study the *differences* between the configurations we encounter in the course of the Monte Carlo dynamics we use to sample the equilibrium probability distribution. Let us call α and β

two configurations that we have generated. In defining their distance we have to remember that there is a rotational and translational motion that is not relevant for defining a distance: we are interested in a parameter that measures differences in shape. We want to know if we find a structure in a chain shape, and we want to be able, for example, to distinguish between an closed shapeless globule and a frozen well-shaped structure. In order to do that we define

$$\delta^2_{(\alpha,\beta)} \equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{\mu=1}^{3}\left(x_i^{(\alpha)\mu} - x_i^{(\beta)\mu}\right)^2 , \tag{6}$$

after taking the minimum over roto-translations. In practice we bring protein β over protein α (overlapping the two gravity centers), and then we find the optimal rotation of β which minimizes $\delta^{(\alpha,\beta)}$.

Such a definition of overlap is by no means unique. We also use a completely different distance, which does not need the minimization procedure. In this case we use the energy of the site couples in order to define

$$\Delta^2_{(\alpha,\beta)} \equiv \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j>i}\left(E_{i,j}^{(\alpha)} - E_{i,j}^{(\beta)}\right)^2 , \tag{7}$$

where $E_{i,j}^{(\alpha)}$ is the site energy (2) of the configuration (α).

The definition (6) is very natural, in that it identifies the physical similarities of two configurations of the same chain (when we say the same chain we mean that we are in the same realization of the random quenched potential: a given *protein* is characterized by the sequence of the amino-acids, and sequences of such elementary constituents do interact in a definite way). Once we have eliminated the rotational and the translational degrees of freedom we are left with an indicator which is zero if the two proteins are identical.

The problem with definition (6) comes if one part of the two chains is very similar and another part is completely different (which usually happens during the folding procedure, when the folding is not yet completed). In this case the first distance could be completely misleading, and one could find, by overlapping the centers of the two configurations, a completely spurious position. Here definition (7) can help, since it locally recognizes part of the two chains which are in a similar energetic state. In our simulations we always find the same answer when looking at the two distance indicators: we consider this as being a very good consistency check, which shows that indicators (6) and (7) are really measuring the intrinsic similarity of two different chains.

The parameters that characterize our model are the number of elementary sequences (sites of the chain) N, the attractive coefficient A, the repulsive coefficient R, the inverse temperature $\beta \simeq T^{-1}$ and the strength of the quenched disorder, ϵ. The different parameters we have described are deeply interconnected.

Apart from the overlap distances we have measured some local observable quantities. We have measured the expectation value of the energy of the system,

$$E \equiv \langle H \rangle \ , \tag{8}$$

where by $\langle \cdot \rangle$ we mean the thermal average over configurations in a given realization of the random potential (we indicate the average over different instances of the random potential by $\bar{\cdot}$: most of the time we will discuss results obtained in a given realization of the potential, because this is the real problem we are eventually interested in). We have monitored the *gyration radius*

$$\rho \equiv \frac{1}{N} \langle \sum_{i=1}^{N} \sqrt{\sum_{\mu=1}^{3} \Big(x_i^\mu - \langle x^\mu \rangle \Big)^2} \rangle \ , \tag{9}$$

and the *link length*

$$\lambda \equiv \frac{1}{N-1} \langle \sum_{i=1}^{N-1} \sqrt{\sum_{\mu=1}^{3} \Big(x_i^\mu - x_{i+1}^\mu \Big)^2} \rangle \ . \tag{10}$$

The coil-globule phase transition is characterized by a sudden jump in ρ when varying A at fixed R (and low ϵ).

The model we are discussing here turns out to have a very rich structure: it is quite easy to implement, and the two definitions of chain-distance we have given allow us to extract a lot of additional relevant information. Another possible approach consists of defining the protein on the lattice (in this case the main advantage is the large computational speed one can reach, and the relative ease of an operational definition of a chain-distance), but the continuum approach turns out, as shown by the results we discuss in this paper, to be very effective.

3. The Phase Diagram of IMP

Let us summarize these first results. In absence of the noise (homopolymer) we observe, when increasing the coefficient of the attractive contribution A, a (well known) phase transition from an open coil state to a globule shapeless phase. For low quenched noise the situation does not change. In the strong noise regime we get an abrupt transition to a completely different phase.

In the coil phase the square of the gyration radius behaves as N, while in the globule phase it behaves as $N^{\frac{1}{3}}$. Such a criterion allows a good empirical definition of the transition point. On the contrary we will see that the *frozen phase* is characterized by a non-trivial structure in the probability distribution of the distances.

The probability of a given chain squared distance, $P(\delta^2)$, is defined as the normalized number of times that, during the course of the Monte Carlo dynamics, two protein chains are at distance δ.

The results for small values of ϵ are quantitatively very similar to the ones with $\epsilon = 0$. Fluctuations from one instance of the potential to a different one are very small, and the statistical errors on the measured quantities can be reliably measured.

Increasing the strength of the disorder (at fixed β and R) we have found, close to a given value of $\epsilon = \epsilon_c$, a transition to a new phase, completely different in character from the ones we have discussed before. Such a phase has all the typical features of a frozen phase in a spin glass, plus some bonuses we will discuss in the following, that make it very suitable to describe the state of a folded, biologically active protein.

The correlation time in the glassy phase is extremely large (we are not able to determine it), and the jump from the two phases (coil and shapeless globule) with *reasonable* correlation times to the new phase is very abrupt. The $P(\delta^2)$ in the new phase looks non-trivial, and we have observed the system to survive in a given state for very long times.

The main question is about the minima of the free energy. From these small scale preliminary runs it is clear that there is, as expected, a very complex structure. There are few stable states (we see at least two different states in which the chain comes back after many iterations). The most important point is perhaps that there are *few* stable states: the fact that after many Monte Carlo iterations, and after visiting a completely different state (see next paragraph), we come back to exactly the same state, is very remarkable, and is a feature that is quite different from the pattern of stable state in disordered spin models (the recent work of ref.[10] points in this direction).

Proteins fold in one or very few stable states: the behavior of the glassy phase one encounters in a spin model (or in the Random Energy Model) would not be consistent with such a phenomenon. In order to explain protein folding by the effect of disorder one has to find few stable structures: this is what we have shown to happen for heteropolymers in random, strong enough disorder.

From these simulations evidence for the existence of a glassy phase in the dynamics of heteropolymers. We have shown that such a phase has typical features which are different from those of a disordered spin model, and are due to the chain-like features of the model, and that such features are exactly what one needs in order to apply such a model to the description of the dynamics of protein folding.

4. Additional Results

In ref.[14] we have obtained more results about the IMP model. Such results have again been based on small scale simulations, but they contribute to define a first picture. The presence of a stretched exponential time behavior has been noted in the second of references[14]. Here the system undergoes an abrupt (not too large) perturbation, and we analyze the following relaxation to equilibrium. This is good evidence for the existence of *quasi-states*. We get the stretched exponential behavior by visiting very similar quasi-states. The stretched exponential is not generated by visiting different minima with underlying different time scales.

In the third of refs.[14] we have discussed in better detail the dynamics of the polymer shape. The $T = 0$ structure of the model has been discussed in the fifth of references[14]. A steepest descent procedure has been used to study the relation among the finite T landscape and the one at $T = 0$. One finds that the structure of the minima does not change much from $T = 1$ to $T = 0$, for example. The free energy landscape is quite stable.

5. On the Ape

APE-100[16] is a SIMD parallel super-computer. In the *tube* version, the one we have used here, it has a theoretical speed of 6.5 Gflops (and, as we will discuss, reaches an unusually high sustained speed). In its largest version it reaches a theoretical speed of 100 Gflops.

Why do we need a supercomputer? We want for example to try to understand if properties like the existence of one or few ground states is a generic property of random heteropolymers or just the pattern followed by some realizations of the random system. We want to compute, in essence, a reliable estimate of

$$\overline{P(\delta)} \; , \tag{11}$$

where here by $P(\delta)$ we indicate the probability distribution of the distances, and with the long bar we indicate the average over the random coupling realizations.

We also have to understand if our results do characterize thermal equilibrium or if they are just typical of a transient behavior (due to long correlation times). For doing that we use a quite standard criterion. We simulate two independent copies of the same system, and we compute two different estimators of the $P(\delta)$. In the first one we consider an *history* distance (i.e. the distance of representatives of the same history at different times), while in the second we compute the distances of the two copies at the *same time*. Only when (and if) these two estimators converge to the same result we can be sure we are looking at an equilibrium phenomenon.

The other open question for which we will need a very large statistical sample concerns determining the scaling law of the mass gap (energy difference of the ground state and the first excited state) with the chain size.

In our parallelization scheme we have one independent realization of random couplings on each processor, and, as we have said, we simulate the evolution of two copies of the system in each coupling set (i.e. in each processor).

In our preliminary runs (one week of the 6.5 Gflops Ape) we have studied the system at different T values (with 16×10^6 full sweeps on 128 30 site systems for each temperature). At $T = 2.5$ and 2.0 the system is at equilibrium, but the true $\overline{P(\delta)}$ is trivial. At $T = 1.0$ equilibrium seems an hopeless task. At $T = 1.5$ longer runs in a second phase lasted two weeks show a non-trivial equilibrium $P(\delta)$. We estimate here a correlation time of order 30 million sweeps.

The computer program we have written is, as we have already said, based on a very naive parallelism, and it is highly efficient on Ape (it reaches order of 70% efficiency). All the program is written in a very high level language.

We are now trying runs based on the *tempering*[17] updating scheme, that we hope will allow us to establish the existence of highly non-trivial $P(\delta)$ deep in the folded phase.

References

1. C. Ghélis and J. Yon, *Protein Folding* (Academic, New York, 1982).
2. T. E. Creighton, *Proteins: Their Structure and Molecular Properties* (Freeman, San Francisco, 1984).
3. M. Kotani, editor, *Advances in Biophysics* (Elsevier, Amsterdam, 1984).
4. D. Wetlaufer, editor, *The Protein Folding Problem* (Westview, Boulder, 1984)..
5. N. Gô, *Annu. Rev. Biophys. Bioeng.* **12**, 183 (1983).
6. T. E. Creighton, *J. Phys. Chem.* **89**, 2452 (1985).
7. I. Iben, D. Braunstein , W. Doster, H. Frauenfelder, M. K. Hong, J. B. Johnson, S. Luck, P. Ormos , A. Schulte, P. J. Steinbach, A. H. Xie and R. D. Young, *Phys. Rev. Lett.* **62**, 1916 (1989).
8. M. Fukugita, H. Kawai, T. Nakazawa and Y. Okamoto, *Nucl. Phys.* **B** (Proc. Suppl.) **20**, 766 (1991).
9. M. Mezard, G. Parisi and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore 1987); G. Parisi, *Field Theory, Disorder and Simulations* (World Scientific, Singapore 1992).
10. M. Mezard and G. Parisi, *Replica Field Theory for Random Manifolds, J. Phys.* **I1**, 809 (1991).
11. G. Iori, E. Marinari and G. Parisi, *J. Phys. A: Math. Gen.* **24**, 5349 (1991).
12. J. D. Bryngelson and P. G. Wolynes, *Spin Glasses and the Statistical Mechanics of Protein Folding, Proc. Natl. Acad. Sci. USA* **84**, 7524 (1987).
13. T. Garel and H. Orland, *Mean-Field Model for Protein Folding, Europhys. Lett.* **6**, 307 (1988).
14. G. Iori, E. Marinari, G. Parisi and M. V. Struglia, *Statistical Mechanics of Heteropolymer Folding, Physica* **A 185**, 98 (1992); G. Iori, E. Marinari and G. Parisi *Non-Exponential Time Scales in Disordered Systems: an Application to Protein Dynamics* Roma Tor Vergata preprint ROM2F/92/35; P. Pliszka and E. Marinari, *On Heteroplymer Shape Dynamics, Europhys. Lett.*; G. Iori, PhD Thesis, unpublished; M. V. Struglia, *Conformational Properties of Random Heteropolymers in the Folded Phase*, Roma Tor Vergata preprint ROM2F/92/42.
15. G. Iori, E. Marinari and G. Parisi, *Large Scale Numerical Simulation on a Parallel Supercomputer*, to be published.
16. See C. Battista et al., *The Ape-100 Computer: the Architecture,* reprinted in the second of references[9] and in press on *International Journal of High Speed Computing*, and references therein. For a general discussion see E. Marinari, *A Review Talk about Computers and Theoretical Physics, Nucl. Phys. B (Proc. Suppl.)* **30** (1993)123;
17. E. Marinari and G. Parisi, *Europhys. Lett.* **19**, 451 (1992);