**Figure Captions.**

**Figure 1.** $\rho^2$ as a function of $A$ for the homopolymer case ($\epsilon = 0$).

**Figure 2.** As in Fig. 2, but $\lambda^2$.

**Figure 3.** $P(\delta^2)$ for the homopolymer ($\epsilon = 0$) in the open phase ($A = 1.6$).

**Figure 4.** As in Fig. 3, but in the globular phase ($A = 3.8$).

**Figure 5.** $P(\delta^2)$ in the *spin glass* phase, $\epsilon = 6.0$ and $A = 3.8$.

**Figure 6.** $\rho^2$ as a function of Monte Carlo time (in units of $10^4$ sweeps of the chain) in the *spin glass* phase, $\epsilon = 6.0$ and $A = 3.8$.

**Figure 7.** As in Fig. 6, but for $\lambda^2$.

**Figure 8.** Squared chain-distances $\delta^2$ from some given chains (that are indicated by a vertical line).

**Figure 9.** Configurational view of the chains indicated by vertical lines in Fig. 8. In each figure the three projections on the planes $x - y$, $x - z$ and $y - z$.

[14] E. I. Shakhnovich and A. M. Gutin, *Frozen States of a Disordered Globular Heteropolymer,* J. Phys. A: Math. Gen. **22** (1989) 1647.

[15] E. I. Shakhnovich and A. M. Gutin, *Formation of Unique Structure in Polypeptide Chains. Theoretical Investigation with the Aid of Replica Approach,* Pushchino preprint, 1989.

[16] E. I. Shakhnovich and A. M. Gutin, *Implications of Thermodynamics of Protein Folding for Evolution of Primary Sequences,* Nature 346 (1990) 773.

[17] G. Iori, E. Marinari and G. Parisi, in preparation.

# 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** (1983) 183.

[6] T. E. Creighton, J. Phys. Chem. **89** (1985) 2452.

[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** (1989) 1916.

[8] M. Fukugita, H. Kawai, T. Nakazawa and Y. Okamoto, *Monte Carlo Simulation for Folded Structure of Peptides*, presented at the 1990 Lattice Field Theory Conference, Nucl. Phys. B (Proc. Suppl.) to be published.

[9] M. Mezard, G. Parisi and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore 1987).

[10] M. Mezard and G. Parisi, *Replica Field Theory for Random Manifolds*, Ecole Normale preprint LPTENS 90/28 (Paris, december 1990).

[11] J. D. Bryngelson and P. G. Wolynes, *Spin Glasses and the Statistical Mechanics of Protein Folding,* Proc. Natl. Acad. Sci. USA **84** (1987) 7524.

[12] T. Garel and H. Orland, *Mean-Field Model for Protein Folding,* Europhys. Lett. **6** (1988) 307.

[13] E. I. Shakhnovich and A. M. Gutin, *The Nonergodic Spin-Glass-Like Phase of Heteropolymer with Quenched Disordered Sequence of Links,* Europhys. Lett. **8** (1989) 327.

Roberto Benzi started with us this work, and we acknowledge many useful and nice discussions with him during all the course of the work. We warmly thank Marc Mezard for many interesting comments and in particular for suggesting the definition of the distance given in eq. (7). We thank Luca Biferale, David Callaway, Mattia Falcone, Silvia Morante and Valerio Parisi for useful discussions about proteins and other subjects.

protein is in the same state.

The main question is about the minima of the free energy. From Figs. 8 it is clear in first 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 exactly to 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 state: the behavior of the glassy phase one encounters in a spin model (or in a 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.

In Figs. 9 we give the conformational pictures of protein selected at the time-points where we take the distance from in Figs. 8. So in Fig. 9a we have the protein from which we take the distances in Fig. 8a and so on (we give the 3 projections on the $x-y$, $x-z$ and $y-z$ planes: the figures are after minimization of $\delta$ over roto-translations, i.e. the projections are, at least in principle, as similar as they can be). It is remarkable that the two states (that we consider *stable* states, since the chain finds them again after billions of Monte Carlo steps) are conformationally completely different. It is very impressive how Fig. 9a is similar to Fig. 9d, and Fig. 9c to Fig. 9f: the intermediate configurational states are completely different, but the chain comes back after many millions of Monte Carlo iterations, to the some configuration.

We have given evidence for the existence of a glassy phase in the dynamics of heteropolymers: we have shown that such a phase has typical features that are different from the ones of a disordered spin model, that are due to the chainy features of the model, and that such features are exactly what one needs to apply such a model to the description of the dynamics of protein folding. In a next paper we will give some more informations about the structure of the free energy minima: we will discuss how the states do cluster, and the possibility of applying an ultrametric description to the states.

**Acknowledgments**

we can see very long-living structures. The macroscopic jump in the radius survives for order of 20 million Monte Carlo sweeps.

In the series of Figs. 8 we give the square distance $\delta^2$ of the protein chains we have encountered in the course of the dynamics from the specific chain we indicate by drawing a vertical line on the selected time. We compute the chain-distance $\delta^2$ from the considered chain to all the chains preceeding it in the (Monte Carlo) time and to the chains following it. Obviously enough the distance is zero in the point specified form the arrow, i.e. the chain-distance of a protein with itself is zero: small $\delta$ means the two configurations are in a similar state, large $\delta$ they are in a different state. We warn the reader that these figures have to be understood in detail, since they do constitute the main point of this work. All the features we will remark in Fig. 8 would persist when looking at the same figures done for the $\Delta$ chain-distances (based, as we have seen, on site energy differences).

Fig. 8a gives the square distance $\delta^2$ from the chain obtained after 15 millions of iterations. The chain is in this moment in a *stable* state: we see from this figure that the chain will return (twice) to the same state after more than 50 millions of Monte Carlo iterations. We can see the start very far from thermal equilibrium (at the beginning the protein is in a transient state, at large distance from all the equilibrium configurations), and after a while (as we said 15 millions of Monte Carlo sweeps) the chain we have decided to take the distance $\delta$ from.

Before 20 millions of iterations (Fig. 8b) the chain goes in a long living state (it last $O(20\ 10^6)$ iterations) where it will not return during all the run. In Fig. 8c the chain is in its second stable state, where it has spent more than 45 millions of Monte Carlo iterations. It should be noticed that the chain is visiting this state for the second time, and that it will come back to the same state once more.

In Fig. 8d the chain is back to the first state. In Fig. 8e it is in a transient state: from Fig. 6, where we have given the gyration radius $\rho^2$ as a function of the Monte Carlo time, we see that such a state is macroscopically different from the other ones, and it is characterized form a different value of $\rho$ In Fig. 8f the chain is back (for the third time) to the second state.

A good way to proceed is to compare the chain-distance $\delta$ and the link length $\lambda$: in a globule unshaped state the typical value of $\delta$ is larger than the distance of two chain sites. On the contrary in a well folded, well shaped phase $\delta$ is very small (on the scale fixed by $\lambda$) for all the time in which the

dynamics, two protein chains are at distance $\delta$ (we pick up one configuration over $10^4$ and we compute the distances of all possible couples). It turns out to be the one one expects in a normal (replica symmetric) phase: we plot it in the coil phase in Fig. 3 for $A = 1.6$ and in the globule phase, for $A = 3.8$, in Fig. 4. We cannot distinguish any kind of structure, in the sense that the $P$ are in this case single peaked usual distributions. In the open phase the probability distribution has a tail for large values which is likely connected to fluctuations in the radius $\rho$, which we expect to be much larger in the coil phase than in the globular one. Such a tail is consistently absent in the globular phase.

We have done simulations for a few different realizations of the $\eta_{i,j}$. The results for small values of $\epsilon$ 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 (we use a jack-knife technique in order to control the convergence of our error estimators).

Increasing the strength of the disorder (at fixed $\beta$ and $R$) we find, 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 very very large (we are not able to determine it), and the jump from the two phases (coil and unshaped globule) with *reasonable* correlation times to the new phase is very abrupt. The $P(\delta^2)$ in the new phase is non-trivial, and we can observe the system to survive in a given state for very long times. We give a typical example (after a very long run of $\simeq 2\ 10^8$ complete chain updating sweeps) of $P(\delta^2)$ in Fig. 5. $N = 30$, $\epsilon = 6.0$ and $A = 3.8$ in these and next figures. The distribution $P(\delta^2)$ has a first peak at a very small value of $\delta$, typical of two chain-configurations that are in the same state, and are very similar. The other part of the distribution correspond to configurations which are macroscopically different: $\delta$ is non-negligible compared to $\lambda$.

Let us discuss in some detail the dynamics in the glassy phase. In Fig. 6 we give $\rho^2$ as a function of the Monte Carlo time, and in Fig. 7 the link squared length $\lambda^2$. Already at such a very rough level (we will see that using our chain-distance criteria we can gather by far more detailed informations)

and the *link length*

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

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

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

# 4 Numerical Simulation

Let us start by summarizing our 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 unshaped 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.

We start our simulation without the random part of the potential ($\epsilon = 0$), with $N = 30$. We have set $\beta = 1$ and $R = 2$, in such a way to get values of $\frac{\rho}{N}$ and $\lambda$ of $O(1)$ for $A = 0$. We compute the relevant quantities for different values of $A$. In Fig. 1 we give $\rho^2$ as a function of $A$, and in Fig. 2 we give $\lambda^2$. The change of regime from a coil phase at small $A$ to a globular phase for large $A$ is clear, around $A = 2$.

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

The probability of a given chain square distance, $P(\delta^2)$, is defined as the normalized number of times that, during the course of the Monte Carlo

the first distance could be completely misleading, in which one could find, by overlapping the centers of the two instances, a completely spurious position. Here definition (7) can help, in which 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 indicator: we consider this as being a very good consistency check, that show 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, $\epsilon$.

The different parameters we have described are deeply interconnected. In our numerical simulations we have mostly fixed $\beta = 1$, and studied the phase diagram in $A$ for different values of the noise $\epsilon$: the repulsive coefficient $R$ has been fixed in such a way to match the scale fixed by the temperature. We have tried some runs with $\beta = 2$, and they have confirmed the idea that roughly a rescaling in $\beta$ corresponds to a rescaling in the other parameters.

The most part of our runs (a part from exploratory ones, in which we have varied $R$) have been done with $R = 2$, and $N = 30$ sites on the chain. We have done some runs with $N = 60$ and some with $N = 10$ and $N = 20$ in order to get informations about the scaling laws of the system.

A part 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 $\bar{\cdot}$ : the most part of the times 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*

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

In order to understand the structure of the equilibrium states of the model (stable and metastable states), we want to use the concept of overlap. From the physical interpretation of the replica approach we are lead to be interested in the *differences* between the different configurations we encounter in the course of the Monte Carlo dynamics we use to sample the equilibrium probability. Let us call $\alpha$ and $\beta$ two configurations that we have generated. In defining their distance we have to remind that there is a rotational and translational motion that is not relevant for defining a distance: we are interested in a parameter that measures shape differences. We want to know if we find a structure in the chain shape: we want to be able, for example, to distinguish between an unshaped closes 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. Practically we bring back protein $\beta$ over the protein $\alpha$ (overlapping the two barycenter), and then we find the optimal rotation of $\beta$ which minimizes $\delta^{(\alpha,\beta)}$.

Such a definition of overlap is by no means unique. We also use a completely different distance, that 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 $(\alpha)$.

The definition (6) is very natural, in which it gives 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 problems with definition (6) comes if one part of the two chains is very similar and another part is completely different (that usually happen during the folding procedure, when the folding is not yet completed). In this case

ferent 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 neighbour interaction on the chain, keeps the chain together. The repulsive $R$ term forbids the collapse of the chain, and the attractive $A$ contribution allows to fold the chain. The choice of a Lennard-Jones form is a convenient, well understood one; other choices or the exponent are obviously possible and we tend to believe that the qualitative behaviour of the model should not change as far as the potentials go to zero at infinity sufficiently fast. We could also have chosen an exponentially dumped interaction, but we have not done this choice for practical numerical reasons.

In absence of the random quenched term we are dealing with an *homopolymer*, and we expect an usual *coil-globule* transition. The globule state of an homopolymer has no definite shape. A quenched disorder could allow (and we will show it does) to form 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 (the one we will call *folded* in the following) would be suitable in order to describe protein folding.

# 3 Dynamics and Overlaps

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 with the correct probability. We always keep the acceptance ratio (i.e. the percentage of accepted updates) to 50%. According to the popular belief such a choice nearly optimizes the efficiency of the simulation.

stable states. We are developing a more detailed analysis which will be presented in a forthcoming paper [17].

# 2 The Model

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 $x_i^\mu$, where in the following latin indices $i$, $j$,... label the $n$-th site of the chain, and greek indices $\mu$, $\nu$,... label the 3 spatial directions (only the ones from $\mu$ on in the alphabet, since we will use $\alpha$, $\beta$,... 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 neighbours on the chain. The deterministic part of the potential has the usual Lennard-Jones form. The main difference from an usual homopolymer 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, that 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 dif-

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 sequence 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 behaviour of random heteropolymers.

The times are ripe for starting such an enterprise. Many crucial progresses have recently been done in the studies of complex systems [9]. Starting from the specific example of amorphous materials, very soon generalized to very different situations, a whole new formalism, the mechanism of replica symmetry breaking, has lead to many new results. In the last 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, that characterize for example phases that cannot be found in usual, non-random spin models, random membranes share some of such 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 S-K infinite range spin model (and also of a Random Energy Model) with what one knows about protein folding. The many, completely disconnected minima structure would not match with protein that always appear to be in one of few allowed states.

In this paper we will see that important features that have been noticed in the approach of ref. [10] can be explicitly found during the numerical simulations of a $N = 30$ heteropolymeric chain. We find, along with the usual coil-globule phase transition, a new *folded* phase, which seems suitable to describe protein folding as a generic phenomenon. We will see that 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 will relate the existence of such a phase to the presence in the system of a strong, quenched disorder.

We refer to the work of [11, 12] for connections between disordered systems and protein folding. In refs. [13, 14, 15, 16] a mean field treatment for heteropolymeric chains has been elaborated.

We will present in this paper our first results, describing the phase diagram of the model and giving the first conclusions about the structure of

# 1 Introduction

Proteins are a fascinating subject (we refer for example to references [1, 2, 3, 4, 5, 6, 7] for an approach to many of the sides of the problem). Proteins are very elegant and multifunctional 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 animal 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) are contained in the *linear* sequence of the messenger RNA: there is no space for explicit coding of the $3d$ structure, that must be determined from the interaction laws of the constituent amino-acids. The given sequence of amino-acids, that does eventually constitute the working protein, is coded in the RNA: the different amino-acids have different interactions, and interact in a different way with the solvent.

Folding is surely a complex and quite a mysterious procedure. We just remind that the time scales involved in the problem are very different: folding time variate a lot, and the time scale involved is much longer of the one needed of a steepest descent to a simple minimum and too short (obviously) for an exhaustive search of configurational space. One or a very few allowed folded state characterizes 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 an universal behavior is calling for the attention of physicists. We want to understand which is the mechanism that allows such a crucial mechanism to work. We want eventually to be able to build the native configuration using the physical 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 many times independent from 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, transitions between different regimes, only depend on such general features: only very specific features (like the value of the critical temperature) depend on the

# RANDOM SELF-INTERACTING CHAINS: A MECHANISM FOR PROTEIN FOLDING.

Giulia IORI, Enzo MARINARI and Giorgio PARISI,

Dipartimento di Fisica,
Università di Roma *Tor Vergata*,
Via E. Carnevale, 00173 Roma, Italy
and
Infn, Sezione di Roma *Tor Vergata*

January 6, 1995

**Abstract**

We investigate by Monte Carlo simulations 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 ones, of a new phase, the *folded phase*, characterised by long relaxation times and by the existence of few stable states.

**ROM2F-91-4**