Random self-interacting chains: a mechanism for protein folding

1991 J. Phys. A: Math. Gen. 24 5349

(http://iopscience.iop.org/0305-4470/24/22/019)

J Phys A Math Gen 24 (1991) 5349-5362 Printed in the UK

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 Istituto Nazionale di Fisica Nucleare, Sezione di Roma 'Tor Vergata', Roma, Italy

Received 14 February 1991, in final form 23 July 1991

Abstract. We investigate by Monte Carlo simulations the thermodynamic behaviour 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

1. Introduction

Proteins are a fascinating subject (we refer for example to [1-7] for various approaches to the many 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) 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

Folding is certainly a complex and quite mysterious procedure. The timescales involved in the problem are very different folding times vary a lot, but the timescale 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 behaviour is attracting the attention of physicists We want to understand what the mechanism is 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 [8]): in other words we want to 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 for 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 in contrast which of their properties are selected by natural evolution: such a study has to be started by investigating in detail the behaviour of random heteropolymers.

The time is ripe for starting such an enterprise. 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 led to many new results. In recent months many results have been obtained for the behaviour 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 [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 this paper we will see that important features that have been noted in the approach of [10] can be explicitly found during the numerical simulations of an $N=30$ 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 timescales, 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 [13-16] a mean field treatment for heteropolymeric chains has been elaborated.

In this paper we will present our first results, describing the phase diagram of the model and giving the first conclusions about the structure of 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* 3D space is characterized by the three values of the coordinates

x_i^μ, where in the following Latin indices $i, j, \ldots$ label the ith site of the chain, and Greek indices $\mu, \nu, \ldots$ label the three spatial directions (only those from μ on in the alphabet, since we will use $\alpha, \beta, \ldots$ 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

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

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 a usual homopolymer is given by the quenched $1/r^6$ 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

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

that is [illegible] if $i=j$ and $k=l$ or if $i=l$ and $j=k$. This effective random intera[illegible] in the biological picture, many different factors: the complex interactio[illegible] 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 $\beta = 1/T$. In the following we will try to reach a good understanding of the role of the disorder (given by the quenched random potential) and of the Lennard-Jones interaction on the chain.

The deterministic part of the potential had 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 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 behaviour of the model should not change as long as the potentials go sufficiently fast to zero at infinity. 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.

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 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 almost optimizes the efficiency of the simulation.

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. 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 a closed shapeless globule and a frozen well shaped structure. In order to do that we define

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

after taking the minimum over roto-translations. In practice we bring protein β over protein α(overlapping the two barycentres), 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

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

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 centres 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 $\beta = T^{-1}$ and the strength of the quenched disorder, ε.

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 ε: the repulsive coefficient R has been fixed in such a way as to match the scale fixed by the temperature. We have tried some runs with $\beta = 2$, and they have confirmed the idea that a rescaling in β roughly corresponds to a rescaling in the other parameters.

Most of our runs (apart 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 information about the scaling laws of the system.

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

$$E = \langle H \rangle \tag{8}$$

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*

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

and the *link length*

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

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 in 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.

4. Numerical simulation and results

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 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.

We start our simulation without the random part of the potential ($\varepsilon = 0$), with $N = 30$. We have set $\beta = 1$ and $R = 2$, in such a way to get values of ρ/N and λ of O(1) for $A = 0$. We compute the relevant quantities for different values of A. In figure 1 we give ρ^2 as a function of A, and in figure 2 we give λ^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 of the gyration radius behaves as N, while in the globule phase it behaves as $N^{1/3}$. Such a criterion allows a good empirical definition of the transition point. In contrast 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 (we pick out one configuration in 10^4 and we compute the distances of all possible couples). It turns out to be the probability distribution one expects in a normal (replica symmetric) phase· we plot it in the coil phase in figure 3 for $A = 1.6$ and in the globule phase, for $A = 3.8$, in figure 4 We cannot distinguish any kind of structure, in the sense that the P are in this case the usual single-peaked distributions. In the open phase the probability distribution has a tail for large values which is probably connected to fluctuations in the radius ρ, 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 ε are quantitatively very similar to those with $\varepsilon = 0$. 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 β and R) we find, close to a given value of $\varepsilon = \varepsilon_c$, a transition to a new phase, completely different in character from those we have discussed before. Such a phase has all the typical features of a frozen

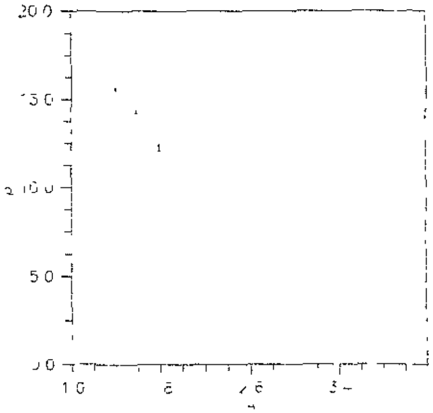

Figure 1. ρ^2 as a function of A for the homopolymer case ($\varepsilon = 0$)

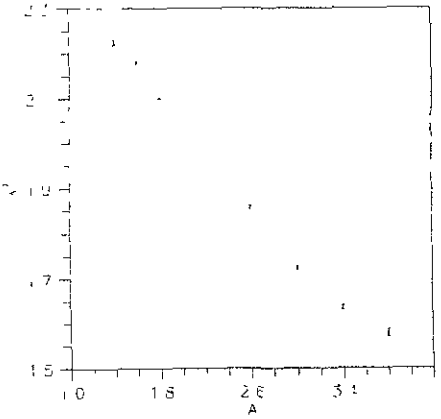

Figure 2. λ^2 as a function of A for the homopolymer case ($\varepsilon = 0$).

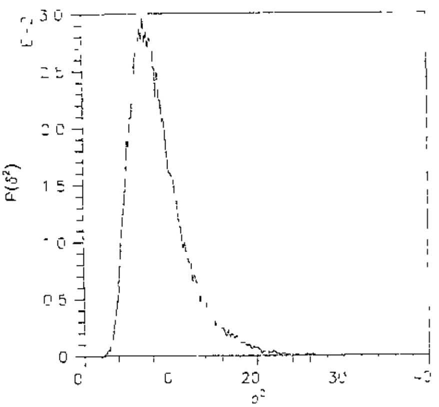

Figure 3. $P(\delta^2)$ for the homopolymer ($\varepsilon=0$) in the open phase ($A=1\ 6$)

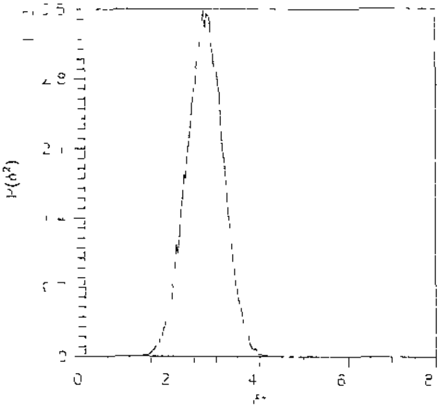

Figure 4. As in figure 3, but in the globular phase ($A=3\ 8$)

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 is non-trivial, and we can observe the system surviving in a given state for very long times. We give a typical example (after a very long run of $\simeq 2\times10^8$ complete chain updating sweeps) of $P(\delta^2)$ in figure 5. In this and in the following figures $N=30$, $\varepsilon=6.0$ and $A=3.8$. The distribution $P(\delta^2)$ has a first peak at a very small value of δ, typical of two chain-configurations that are in the same state, and are very similar. The other part of the distribution corresponds to configurations which are macroscopically different: δ is non-negligible compared to λ.

Let us discuss in some detail the dynamics in the glassy phase. In figure 6 we give ρ^2 as a function of the Monte Carlo time, and in figure 7 the link squared length λ^2

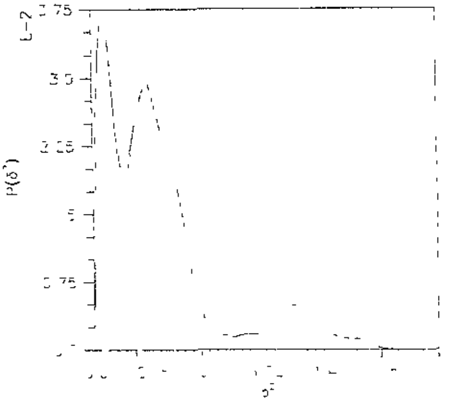

Figure 5. $P(\delta^2)$ in the *spin glass* phase, $\varepsilon=6\ 0$ and $A=3\ 8$.

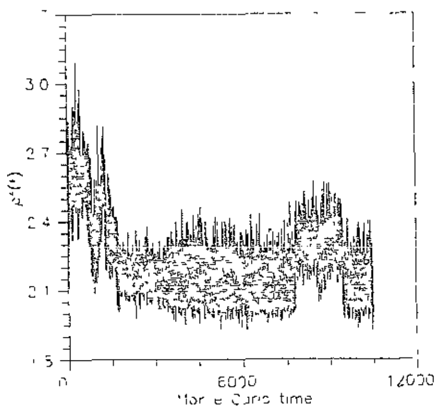

Figure 6. ρ^2 as a function of Monte Carlo time (in units of 10^4 sweeps of the chain) in the *spin glass* phase $\epsilon=6\,0$ and $A=3\,8$

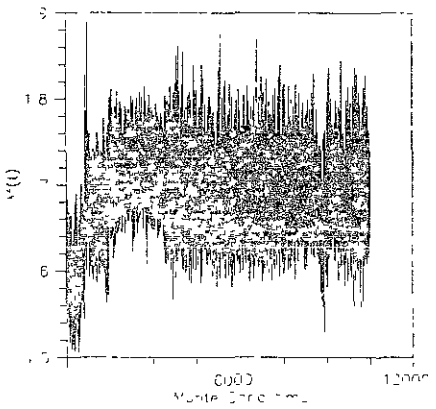

Figure 7. As in figure 6, but for λ^2

Already at such a very rough level (we will see that using out chain-distance criteria we can gather far more detailed information) we can see very long-lived structures. The macroscopic jump in the radius survives for the order of 20 million Monte Carlo sweeps.

In the series of figures 8 we give the squared distance δ^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 δ^2 from the considered chain to all the chains preceding it in the (Monte Carlo) time and to the chains following it. Obviously the distance is zero at the point specified by the vertical line, i.e. the chain-distance of a protein with itself is zero: small δ means the two configurations are in a similar state, large δ they are in a different state. We warn the reader that these figures have to be understood in detail, since they constitute the main point of this work. All the features we note in figure 8 would persist when looking at the same figures done for the Δ chain-distances (based, as we have seen, on site energy differences).

Figure 8(a) gives the squared distance δ^2 from the chain obtained after 15 million iterations. The chain is at 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 million Monte Carlo iterations We can see that at the start we are very far from thermal equilibrium (at the beginning the protein is in a transient state, at a large distance from all the equilibrium configurations), and after a while (as we said 15 million Monte Carlo sweeps) we see the chain from which we have decided to measure the distance δ

Before 20 million iterations (figure 8(b)) the chain goes into a long-lived state (which lasts $O(20\times10^6)$ iterations) where it will not return during the entire run. In figure 8(c) the chain is in its second stable state, where it spends more than 45 million 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 figure 8(d) the chain is back to the first state. In figure 8(e) it is in a transient state: from figure 6, where we have given the gyration radius ρ^2 as a function of the Monte Carlo time, we see that such a state is macroscopically different from the others,

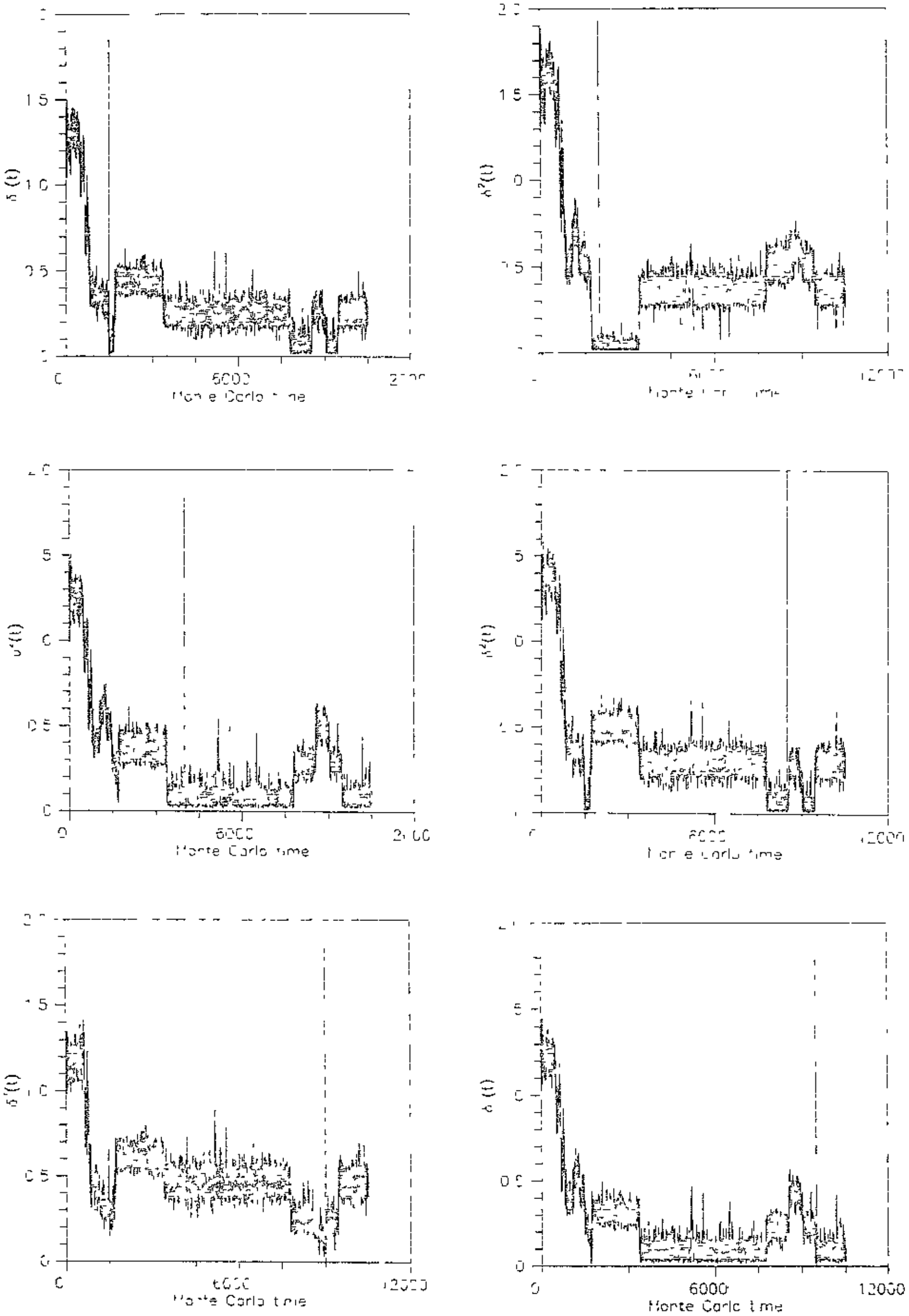

Figure 8. Squared chain-distances δ^2 from some given chains (that are indicated by a vertical line)

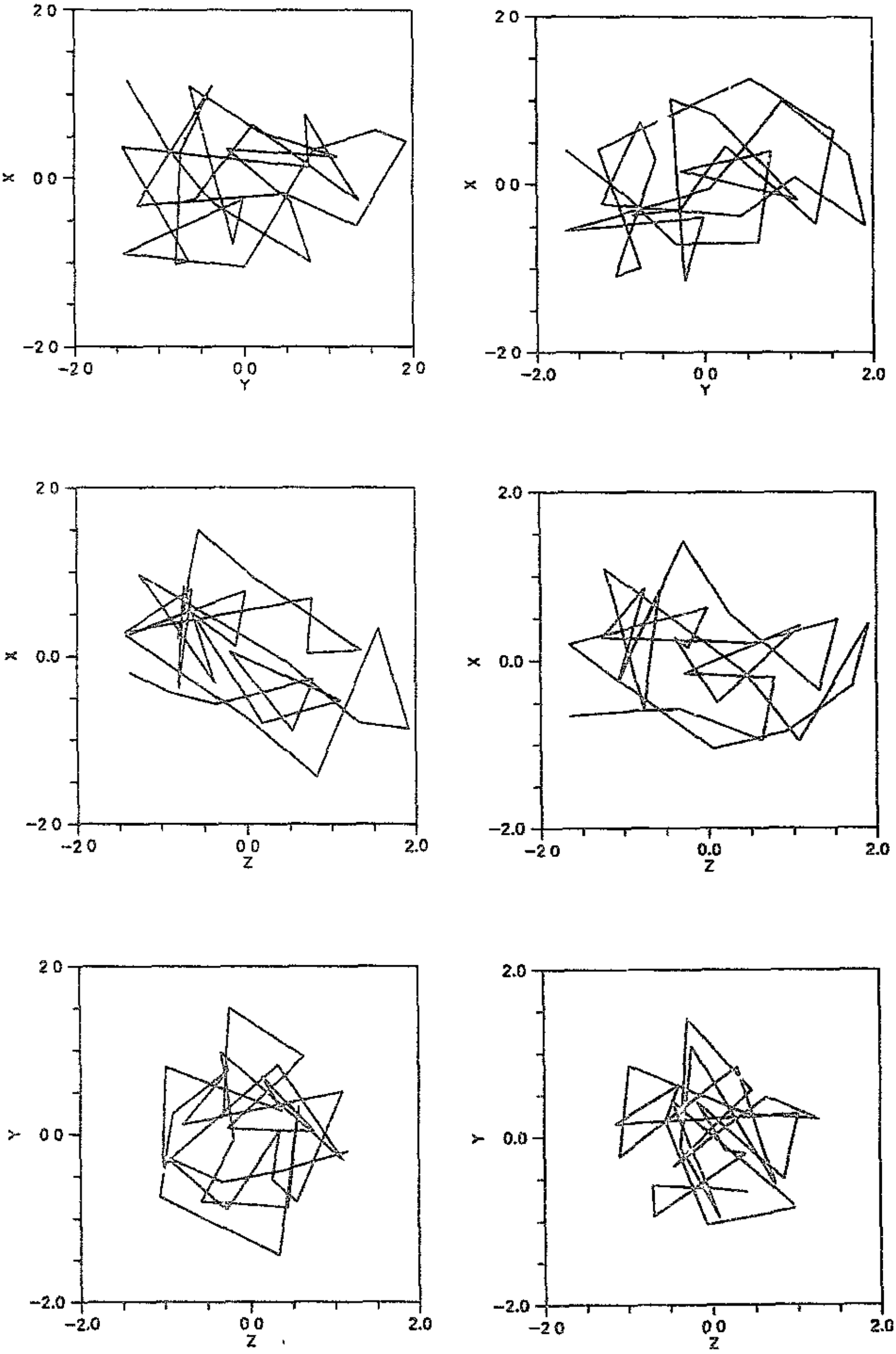

Figure 9. Configurational view of the chains indicated by vertical lines in figure 8 Each figure shows the three projections on the x–y, x–z and y–z planes.

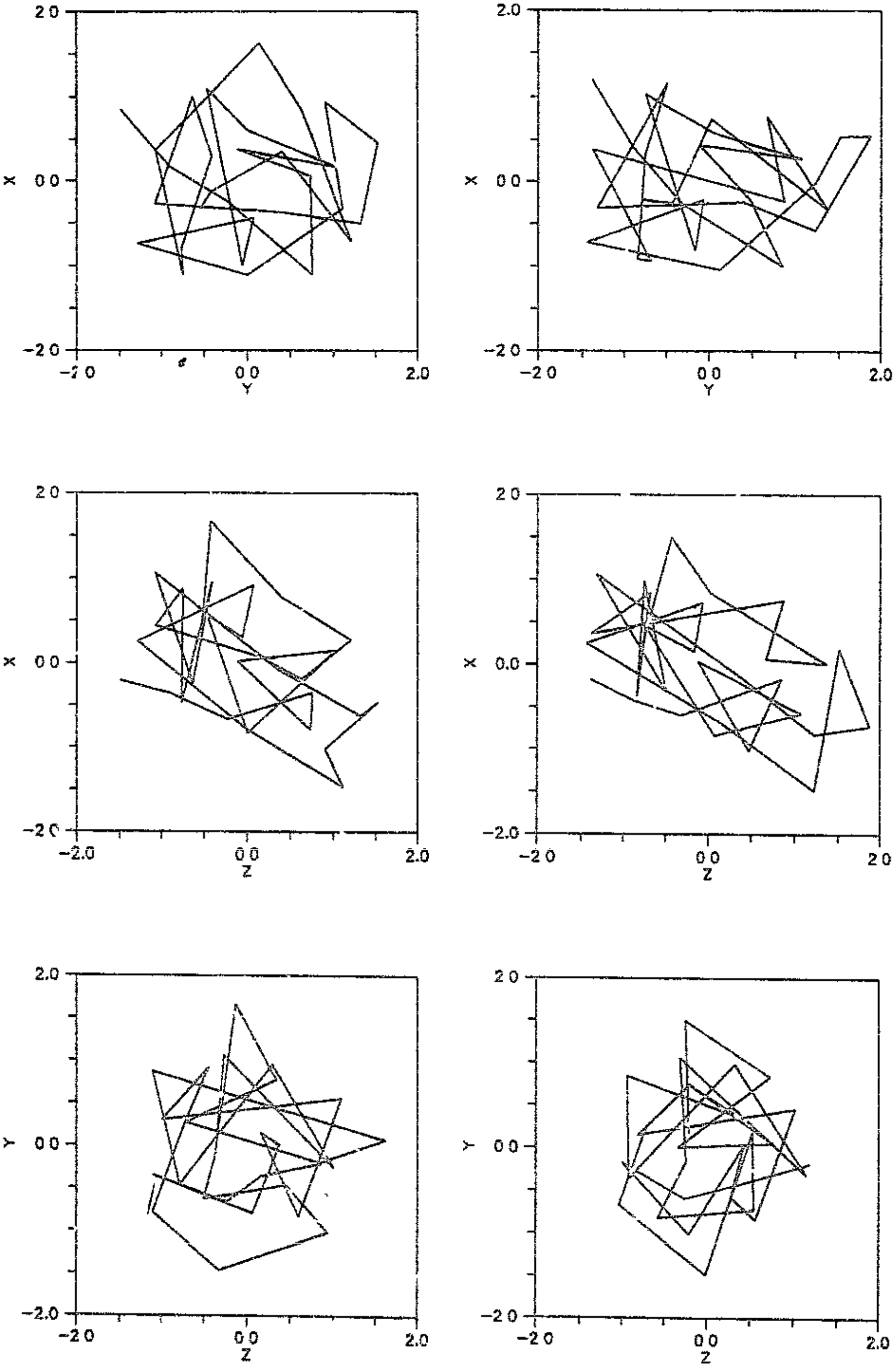

Figure 9. (contınued)

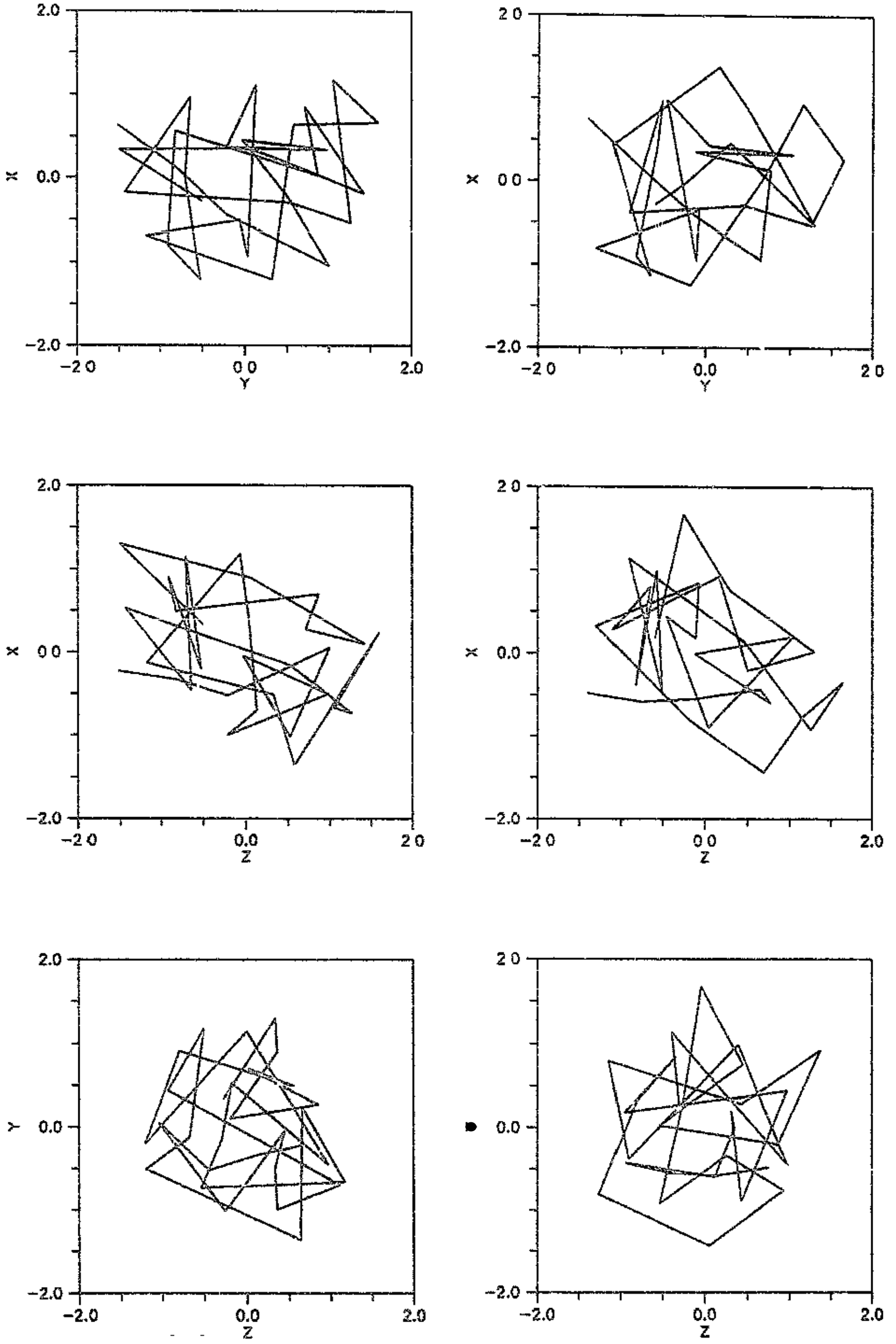

Figure 9. (continued)

and it is characterized by a different value of ρ In figure 8(f) the chain is back (for the third time) to the second state.

A good way to proceed is to compare the chain-distance δ and the link length λ: in a globule shapeless state the typical value of δ is larger than the distance between two chain sites. In contrast, in a well folded, well shaped phase δ is very small (on the scale fixed by λ) for all the time in which the protein is in the same state.

The main question is about the minima of the free energy. From figures 8 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 [10] points in this direction).

Proteins fold in one or very few stable states the behaviour 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.

In figures 9 we give the conformational pictures of proteins selected at the time-points from where we take the distance in figures 8. So in figure 9(a) we have the protein from which we take the distances in figure 8(a) and so on (we give the three projections on the x–y, x–z and y–z planes: the figures are after minimization of δ 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 (which 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 figure 9(a) is similar to figure 9(d), and figure 9(c) to figure 9(f): the intermediate configurational states are completely different, but the chain comes back, after many million Monte Carlo iterations, to the same 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 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. In a forthcoming paper we will give some more information about the structure of the free energy minima: we will discuss how the states cluster, and the possibility of applying an ultrametric description to the states.

Acknowledgments

Roberto Benzi started on this work with us, and we acknowledge many useful and pleasant discussions with him during the course of the work. We warmly thank Marc Mézard for many interesting comments and in particular for suggesting the definition of the distance given in (7). We thank Luca Biferale, David Callaway, Mattia Falcone, Silvia Morante and Valerio Parisi for useful discussions about proteins and other subjects. We warmly thank Paul Coddington for his essential critical reading of the manuscript.

References

[1] Ghélis C and Yon J 1982 *Protein Folding* (New York Academic)

[2] Creighton T E 1984 *Proteins Their Structure and Molecular Properties* (San Francisco Freeman)

[3] Kotani M (ed) 1984 *Advances in Biophysics* (Amsterdam Elsevier)

[4] Wetlaufer D (ed) 1984 *The Protein Folding Problem* (Boulder Westview)

[5] Gô N 1983 *Annu Rev Biophys Bioeng* **12** 183

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

[7] Iben I, Braunstein D, Doster W, Frauenfelder H, Hong M K, Johnson J B, Luck S, Ormos P, Schulte A, Steinbach P J, Xie A H and Young R D 1989 *Phys Rev Lett.* **62** 1916

[8] Fukugita M, Kawai H, Nakazawa T and Okamoto Y, 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] Mezard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore. World Scientific)

[10] Mezard M and Parisi G, Replica field theory for random manifolds *Preprint* Ecole Normale, LPTENS 90/28 (Paris, December 1990)

[11] Bryngelson J D and Wolynes P G 1987 *Proc Natl Acad Sci USA* **84** 7524

[12] Garel T and Orland H 1988 *Europhys Lett* **6** 307

[13] Shakhnovich E I and Gutin A M 1989 *Europhys Lett* **8** 327

[14] Shakhnovich E I and Gutin A M 1989 *J Phys A Math Gen* **22** 1647

[15] Shakhnovich E I and Gutin A M 1989 Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of replica approach *Preprint* Pushchino

[16] Shakhnovich E I and Gutin A M 1990 *Nature* **346** 773

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