

## Conformational properties of random heteropolymers in the folded phase

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1995 J. Phys. A: Math. Gen. 28 1469

(<http://iopscience.iop.org/0305-4470/28/6/005>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 171.66.16.68

The article was downloaded on 02/06/2010 at 02:18

Please note that [terms and conditions apply](#).

# Conformational properties of random heteropolymers in the folded phase

Maria Vittoria Struglia

Dipartimento di Fisica, Università di Roma Tor Vergata, Via della Ricerca Scientifica 1, 00133 Rome, Italy

and INFN, Sezione di Roma Tor Vergata, Rome, Italy

Received 28 July 1994

**Abstract.** We investigated the behaviour at zero temperature of a random self-interacting chain, using a *steepest-descent* algorithm. We mainly studied the energy minima landscape, showing that there are few relevant energy minima at this temperature; we estimate the radius of their basin of attraction.

## 1. Introduction

It is well known that biologically active proteins are characterized by one or very few three-dimensional structures [1–7].

There are interesting connections between the protein folding phenomenon and the physics of disordered systems: it has already been established that the protein folding mechanism is a global optimization problem analogous to the determination of the minimum free-energy configuration of a spin glass [8–16].

So, beyond its importance in biochemical sciences, protein folding provides us with an intriguing statistical mechanics problem, which can be studied in the same framework derived for other disordered systems.

At present, the main efforts of this line of research are oriented in establishing which properties proteins share with random self-interacting chains, in order to distinguish those features that have been selected by natural evolution from the ones exhibited by an entire class of disordered systems. In other words the final aim is to point out which are the universal physical properties governing the behaviour of all these systems, i.e. which properties do not change when we change the chemical details of the polymeric chain slightly.

In this paper we will describe some of the new results we have obtained investigating the behaviour at zero temperature of a random self-interacting chain: our starting point is the results obtained by Iori *et al* adopting a simple model (IMP) of a random heteropolymer [17].

They defined the phase diagram of such a heteropolymer and studied its main properties under the evolution of local Monte Carlo dynamics.

We are now interested in studying the *energy minima landscape* of the model: using a *steepest-descent* algorithm we demonstrate that there are few relevant energy minima in the zero-temperature limit, as was expected if we want to reproduce the characteristics of a native protein. We also estimate the basin of attraction of these minima.

Another important feature emerges from our study: there is evidence of a hierarchical organization, at least with two levels, of the low-lying energy states.

We are also interested in comparing all these results with those obtained for heteropolymers with a reduced number of components.

The paper is organized as follows. In the next section we briefly recall the main results obtained by Iori *et al.*, in the third and fourth sections we will describe the results that have been obtained for a chain made up of 30 points, in the fifth section we compare these results to those obtained for a 15-point chain. In the last section we state our conclusions and outlooks.

## 2. The model

We consider  $N$  sites of a chain in *continuum* three-dimensional space and we indicate their position by the three coordinates  $x_i^\mu$ . We assume that different sites of the chain interact with one another in the following way:

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

where  $r_{i,j}$  is the usual Euclidean distance between sites  $i$  and  $j$ .

The harmonic term between first-neighbours along the chain avoids the system dissociation. The repulsive term becomes relevant at short distances, the attractive term, when it is strong enough, allows the chain to assume a globular shape. The last term in (1) represents the *quenched* potential chosen so as to have zero expectation value and correlation:

$$\langle \eta_{i,j} \eta_{k,l} \rangle = \epsilon \delta_{(i,j),(k,l)}. \quad (2)$$

In a biological picture the sites of the chain represent the points where the protein folds. We expect the disordered part of the potential to simulate the interactions between amino acids themselves, and with solvent molecules.

The Hamiltonian is defined as

$$H \equiv \sum_{i=1}^N \sum_{j>i} E_{i,j}. \quad (3)$$

The first interesting results were obtained using a local Monte Carlo dynamic. In order to obtain information on the system behaviour at equilibrium, the expectation value of the energy, the link, and the gyration radius of the system were measured.

Since we are mainly interested in comparing the different shapes assumed by the polymer during the simulations it is very useful to define a distance in the conformational space.

Let us call the  $\alpha$  and  $\beta$  two configurations that we have generated. We define the distance between two configurations as the minimum over roto-translations of

$$\delta^{(\alpha,\beta)} \equiv \frac{1}{N} \sum_{i=1}^N \sum_{\mu=1}^3 (x_i^{(\alpha)\mu} - x_i^{(\beta)\mu})^2 \quad (4)$$

which we shall call *overlap* between the configurations  $\alpha$  and  $\beta$ . We remind the reader about the possibility of using another definition of distance involving the energies of the two configurations [17].

First of all, varying the parameters in the Hamiltonian, we can recognize three possible different phases for the system: coil, globule but unshaped, frozen. The form of the overlap distribution enables us to distinguish which phase the system is in.

In the first two situations the overlap has a single-peaked distribution, but in the globular phase, the most probable distance between two configurations is much smaller than that of

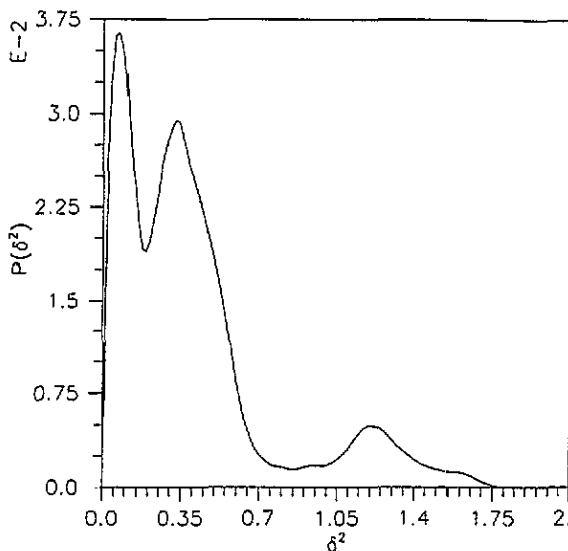


Figure 1.  $P(\delta^2)$  in the frozen phase,  $\epsilon = 6.0$  and  $A = 3.8$ .

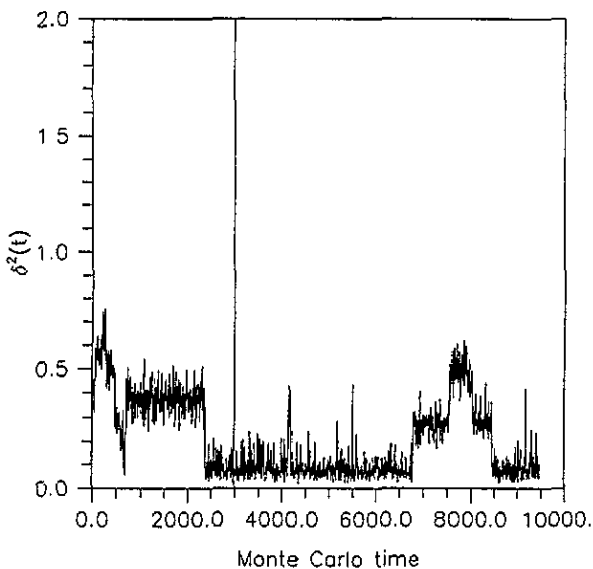


Figure 2. Squared chain distances  $\delta^2$  from some given chains (that are indicated by a vertical line) with  $N = 30$ .

the coil phase. There is evidence that in the coil, as in the globular phase, the system does not have a preferred shape, but in the first case the chain is practically open, in the second case it is folded in a globule, and the overlaps between two different configurations are strongly reduced (see [17]). The transition between these two phases is essentially a transition in the attractive parameter  $A$  at small values of  $\epsilon$ , as shown in [17].

When the random potential strength  $\epsilon$  increases, exceeding a critical value, the distribution of overlap changes and shows a more complex structure. The system studied in [17] was a 30-point chain, and its evolution under Monte Carlo dynamics, in the frozen

phase, was studied with the following values of the parameters:  $R = 2$ ,  $A = 3.8$ ,  $\epsilon = 6.0$ ,  $\beta = 1$ . The results of these simulations are reported in the figures 1 and 2.

As we can see in figure 1 there is a great contribution at very low values of  $\delta^2$ , which corresponds to very similar configurations, namely configurations belonging to the same state. Then there is a second peak centred on a higher value of the distance corresponding to configurations in different states. In figure 2, the distance from a given configuration is plotted versus the Monte Carlo time. We note the existence of quite long-lived states, between which the system jumps. This means that in this phase, the one we called *frozen*, there are a lot of configurations conformationally very similar with one another and also quite stable. As a matter of fact the system spent a lot of time in their neighbours before jumping to another state, and often it came back to a state already visited.

We recall that there is a strong analogy between the phase-diagram of such a random heteropolymer and the phase-diagram of real proteins, where we encounter a *coil* phase, and a *folded* phase, biologically active, characterized by one or few tertiary structures.

### 3. The zero-temperature limit: the energy minima landscape

We shall now show the main results obtained by the application of a *steepest-descent* algorithm on the system just described. We recall that this algorithm provides a method of minimizing functions with more than one variable, searching the minimum in the direction along which the function decreases most rapidly. When the gradient's module becomes smaller than a given tolerance the research is stopped.

Such an algorithm enables us to reduce the thermal noise visible in the Monte Carlo history at  $T = 1$ , and to study the energy minima landscape, that, in analogy with other disordered systems, we expect to be corrugated and characterized by the presence of valleys separated by high-energy barriers.

For these purposes we have selected a wide sample of configurations between those generated by the Monte Carlo with  $\beta = 1$ , and we have applied the steepest-descent algorithm to minimize their energy. In this way we can identify the nearest minima to the initial configurations, namely the energy minima which the system would reach if it were suddenly frozen.

The results are shown in figures 3 and 4: the first one represents the behaviour of the distance of all the minima configurations from the one indicated by the vertical line. As we can see, the picture is very similar to the one obtained at  $\beta = 1$ , but without the thermal noise, and one state is practically reduced to a straight line with rare fluctuations. This essentially means that many configurations, belonging to the same state, are moving towards the same energy minima.

Figure 4 is even more interesting, in which the overlap distribution obtained from the energy minima configurations is shown.

First of all we do not a continuous range of overlaps; it follows that we can distinguish between some particularly relevant values of the overlap, characterized by the higher spikes, and a lot of less interesting little peaks. It means that few energy minima configurations, reached by the system applying the steepest-descent algorithm, are preferred to the others.

This is also confirmed by the fact that the most important contribution is, without any doubt, the one centred on very low values of  $\delta^2$ , coming from those configurations which have fallen into the same minimum. We then encounter a structure extending from  $\delta^2 \sim 0.001$  to  $\delta^2 \sim 0.15$  and corresponding to those configurations that fell into different minima of the same state.

The second major peak and the nearby structures extending from  $\delta^2 \sim 0.25$  to  $\delta^2 \sim 0.5$ ,

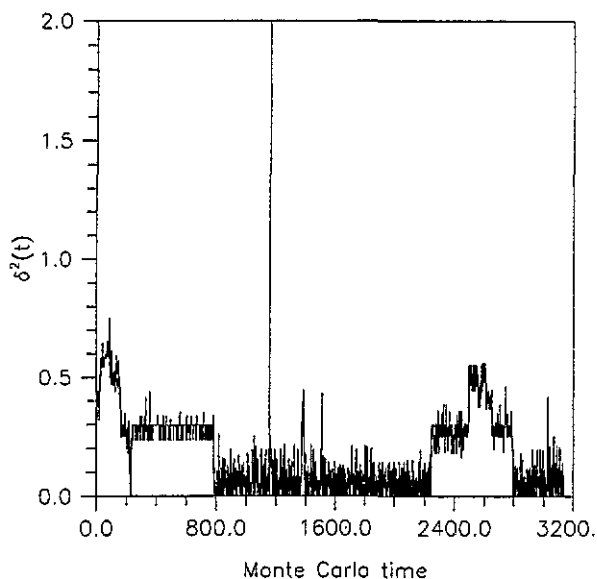


Figure 3. Squared chain distances  $\delta^2$  between cooled configurations with  $N = 30$ .

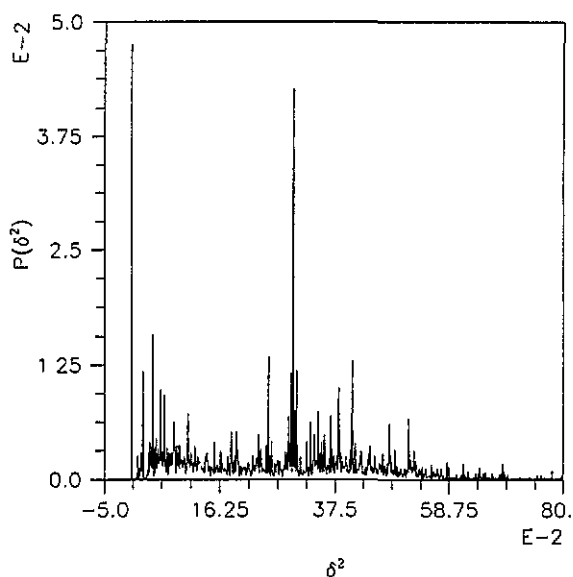


Figure 4.  $P(\delta^2)$  between cooled configurations with  $N = 30$ .

are the contributions coming from configurations which have fallen in the minima belonging to different states.

We can interpret our results in the sense of the existence of at least two levels of *similarity* in our system: configurations belonging to the same state move in no more than five or six different minima, conformationally similar, in the sense of (4), instead configurations belonging to different states have clearly distinct shapes. So we can imagine

the energy landscape of the IMP model as made of some valleys, representing different states, and a *fine structure* inside them, made of the minima belonging to the same state. The dynamics between the large-scale structures is rather slow, on the other hand, inside the micro-structures, we have a faster dynamic and the system oscillates around a 'mean configuration'.

#### 4. Study of stability

Once we have stated that there are not so many relevant energy minima, we are interested in understanding how stable they are under a *perturbation* of their configuration, in order to better characterize the minima energy landscape. Let us explain what we mean: we consider the configurations corresponding to the most relevant energy minima,  $x_M$ , being the ones in which the protein falls most frequently, and we add to them a perturbation  $\delta x$  chosen at random. In this way we obtain a guess configuration:

$$x_G = x_M + \delta x \quad (5)$$

on which we apply the steepest-descent algorithm. For each minimum we try different perturbations with increasing module, in such a way that the squared distance between  $x_G$  and  $x_M$  increases as

$$r^2 = \delta x^2 \quad (6)$$

where  $\delta x^2$  is calculated according to the given definition (4). In this way we can establish with which frequency the system comes back to the initial minimum energy configuration as we propose perturbations always more distant.

In our runs we made fifty trials for every value of  $r^2$ , whose values extended from 0.06 to 0.25 with step 0.01. We refer now to the case of the 30-point chain, particularly to the two longest stable states we can see in the Monte Carlo history.

Let us consider the first stable state, in which the system falls down three times. The most relevant minima encountered in its study have the following energy values:

(i) Stable state

$$\begin{array}{ll} E_1 = -661.6599 & E_2 = -661.1410 \\ E_3 = -660.1942 & E_4 = -659.8622. \end{array}$$

Exploring their surroundings as just described, we found that the system came back to the initial minimum in more or less 55% of the trials, another three or four minima with higher energy were picked up about 1–2%, and a lot of minima with high energy values were picked up just once.

If we draw histograms representing the frequencies with which the system comes back to the initial configuration for the different distances of the proposed perturbations we can guess the radius of the basin of attraction for the three minima, as the point at which the probability to come back to the starting configuration falls to  $\frac{1}{2}$ , obtaining values of  $r^2$  at about 0.15–0.17 (see figure 5). The errors appearing in the histograms are calculated as

$$\epsilon = \frac{1}{M} \left( \frac{K(M-K)}{M} \right)^{1/2}$$

according to the usual binomial distribution, where  $M$  is the number of all of the trials, at a given  $r^2$ , and  $K$  is the number of times that the system came back to the initial minimum.

The results relative to the study of the energy minima belonging to the second stable state are quite similar: we can recognize this state in figure 2(a) as the longest one, in which the system falls down twice. The lowest energy minima we encountered are:

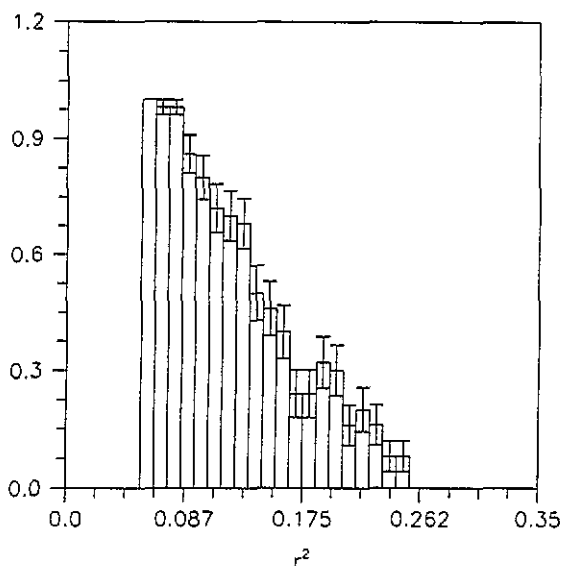


Figure 5. Histogram of the frequency with which the system comes back to  $E = -661.6599$  versus  $r^2$  ( $N = 30$ ).

(ii) Stable state

$$E_1 = -660.1273 \quad E_2 = -659.9702 \quad E_3 = -659.3940$$

and just as in the previous case the polymer came back to the starting configuration in about 55% of the trials.

The values we have obtained for the radius of the basin of attraction of the energy minima are again at about  $r^2 \sim 0.15-0.16$ .

This means that the lowest energy minima we have encountered are similarly relevant in respect of their basin of attraction, in the sense that they are equally stable.

## 5. Simulations with $N = 15$ : results and comparison

It seems particularly relevant to understand the dependence of the behaviour of the system by the number of points constituting the chain, so we have made simulations with  $N = 15$ . In this section we shall show the results obtained.

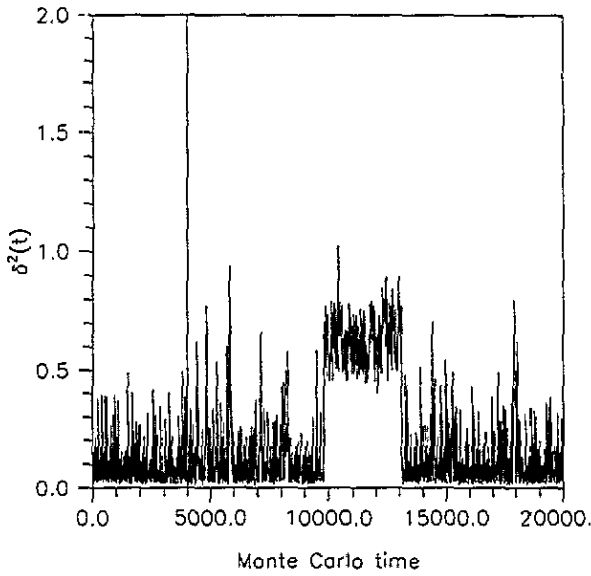
We chose 15 points because it is a sufficient number of components to see the folding, and the CPU time spent in a simulation is quite short.

Just as in the previous case, we first studied the behaviour of the system at  $\beta = 1$  using a Monte Carlo algorithm with the same values for the parameters of the Hamiltonian, and we get similar results. We can observe long-lived states with configurations characterized by a definite fold around which the chain fluctuates: with a reduced number of components we found only two different states between which the system jumps.

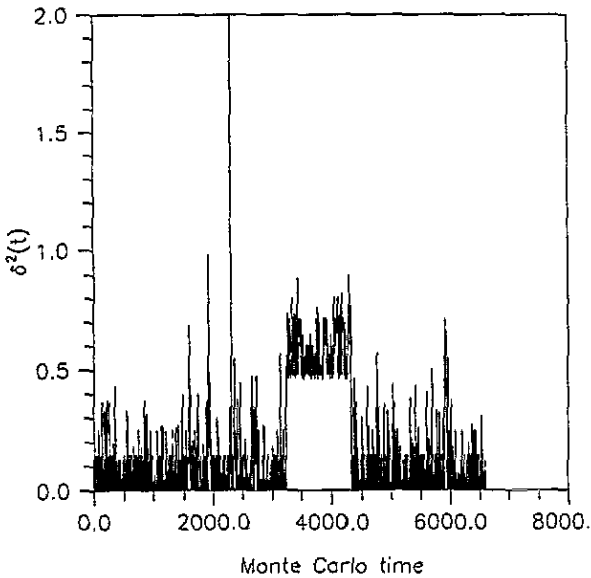
In figure 6 we show the distance of all configurations from a given one, indicated in the figure with a vertical arrow, figure 7 refers to the squared chain distances between cooled configurations.

Using the steepest-descent algorithm we also got the distribution  $P(\delta^2)$  at  $T = 0$ , (figure 8): it is quite similar to that obtained with  $N = 30$ . In figure 8 we can see that the





**Figure 6.** Squared chain distances  $\delta^2$  from some given chains (that are indicated by a vertical line) with  $N = 15$ .



**Figure 7.** Squared chain distances  $\delta^2$  between cooled configurations with  $N = 15$ .

first peak is at  $\delta^2 \sim 10^{-13}$  and it is the contribution coming from those configurations which fell into the same minimum. We can then recognize two structures: the first one extends from  $\delta^2 \sim 0.1-0.3$ , and shows what the typical distances between minima belonging to the same state are. The second one,  $\delta^2 \in (0.5, 0.8)$ , comes from minima belonging to different states.

Comparing these results with the previous ones, obtained in the case  $N = 30$ , we can

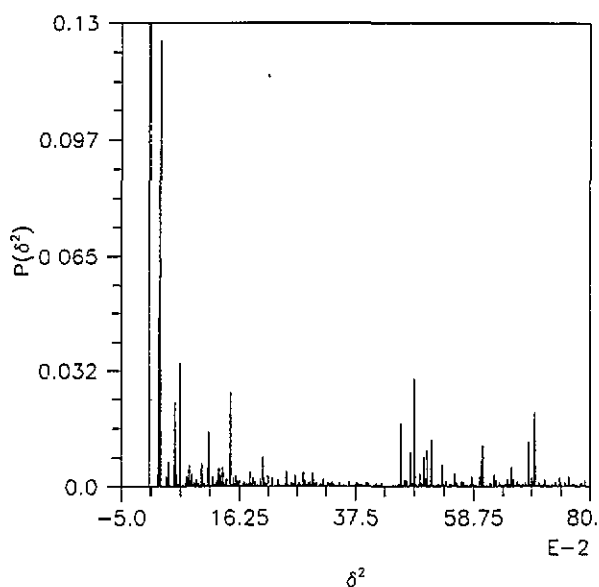


Figure 8.  $P(\delta^2)$  between cooled configurations with  $N = 15$ .

see that in this case the structures relative to the contributions coming from configurations in the same state or in different states are even more clear and definite. In fact, in the 15-point chain case the leading contributions to the distribution come from configurations belonging to the same state, the two structures in the  $P(\delta^2)$  are clearly separate, and the long tail on high values of  $\delta^2$  is practically absent.

We think that all these differences are related to the fact that in our simulation with  $N = 15$  we have encountered just two states.

There is also evidence that in the case  $N = 15$  we have entirely explored the space of the configurations during the simulations: this ensures the validity of our previous considerations about the organization of the minima energy landscape, as discussed at the end of the third section.

We shall now show the results we have obtained during the study of stability. We explored from  $r^2 = 0.02$ – $0.3$ , with steps of  $0.01$ . We did fifty trials for each value of  $r^2$ .

(i) Stable state

$$\begin{array}{ll} E_1 = -252.6436 & E_2 = -252.5755 \\ E_3 = -252.4609 & E_4 = -251.6819 . \end{array}$$

Studying the stability of these minima we obtained an interesting result. The first two minima are stable in the sense that the system came back to the initial configuration in 61% of trials for the lowest one, and in 55% of trials for the second one. But the other two energy minima,  $E = -252.4609$  and  $E = -251.6819$ , are as stable as the others. In both cases the system preferred to fall down in the two lowest minima rather than to come back to the initial minimum, and the radius of the basin of attraction is shorter than that of the other two minima, and it is about at  $r^2 \sim 0.11$  rather than to  $r^2 \sim 0.17$  (see figure 9). So we have a clear differentiation between the minima belonging to the same state, in terms of the relevance of their basin of attraction: the first two configurations will be clearly preferred to the others.

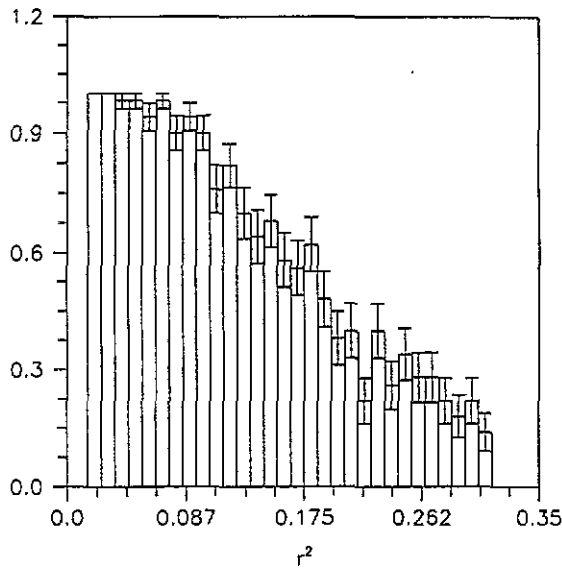


Figure 9. Histogram of the frequency with which the system comes back to  $E = -252.6436$  versus  $r^2$  ( $N = 15$ ).

The results relative to the second state are similar, but we did not find the minima as unstable.

## 6. Conclusions and outlook

Undoubtly, one of the main results to achieve is to understand how many folds the system can assume.

We have shown that, as expected for real proteins, the IMP model has few—one or more—stable states in respect to the usual spin glasses, even with 30 or 15 points.

We think that this important difference is mainly due to the fact that the heteropolymer model we have adopted is less frustrated than a usual spin glass is. The points of the chain can move freely in the three-dimensional continuum space, adjusting their position in order to minimize the possible frustration coming from the random interaction.

The IMP model seems to reproduce some of the most important features of the behaviour of real heteropolymers, such as the phase diagram or the allotropy, giving a clear indication in the sense explained in the introduction: there are some properties not strictly related to the real sequence of amino-acids along a heteropolymeric chain, but they are common to an entire class of physical systems.

We recall that some interesting related results have also been obtained by Shakhnovich and Gutin [13–16], and by Fukugita *et al* [17].

In particular, we recall that the results described in this paper are in good agreement with those described in [17], concerning the study of a two-dimensional IMP model. The authors observe the same dynamical behaviour of the three-dimensional case under local Monte Carlo evolution, and using both deterministic quenching procedures and simulated annealing they find that an increase in the strength of the random interactions  $\epsilon$  leads to the appearance of a single deep global minimum well separated by an energy gap from other local minima.

We also obtain interesting information on the organization of the states, which appear to be divided into two distinct levels of similarity, and on the energy minima landscape.

We are now interested in improving our knowledge about the behaviour of the system if we change the number of components  $N$ , or if we just change the values of the quenched couplings.

## Acknowledgments

I am grateful to my supervisors Enzo Marinari and Giorgio Parisi, who helped me with my work, contributing useful advice and their experience. I also thank Dr Giulia Iori for the helpful and interesting discussions we have had. I thank Professor Giorgio Parisi for careful reading of this paper.

## References

- [1] Ghéllis 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, CO: Westview)
- [5] G6 N 1983 *Ann. Rev. Biophys. Bioeng.* **12** 183
- [6] Creighton T E 1985 *J. Phys. Chem.* **89** 2452
- [7] Iben I *et al* 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.)* **20** 766–70
- [9] Mezard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
- [10] Mezard M and Parisi G 1991 Replica field theory for random manifolds *J. Physique* **1** 809–36
- [11] Bryngelson J D and Wolynes P G 1987 Spin Glasses and the statistical mechanics of protein folding *Proc. Natl Acad. Sci. USA* **84** 7524
- [12] Garel T and Orland H 1988 Mean-field model for protein folding *Europhys. Lett.* **6** 307
- [13] Shakhnovich E I and Gutin A M 1989 The nonergodic spin-glass-like phase of heteropolymer with quenched disordered sequence of links *Europhys. Lett.* **8** 327
- [14] Shakhnovich E I and Gutin A M 1989 Frozen states of a disordered globular heteropolymer *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 Implications of thermodynamics of protein folding for evolution of primary sequences *Nature* **346** 773.
- [17] Fukugita M, Lancaster D and Mitchard M G 1992 Energy level structure of two-dimensional self-interacting random chains *J. Phys. A: Math. Gen.* **25** L121–6
- [18] Iori G, Marinari E and Parisi G 1991 Random self interacting chains: a mechanism for protein folding *J. Phys. A: Math. Gen.* **24** 5349
- [19] Iori G, Marinari E, Parisi G and Struglia M V 1992 *Statistical mechanics of random chain* *Physica* **185A** 98–103