Roma I, 1111/95
chem-phys/yymmdd

A Heteropolymer Model Study for the Mechanism of Protein Folding

Masataka Fukugita[(a)], David Lancaster[(b)] and Mark G. Mitchard[(c)]

(a): Yukawa Institute for Theoretical Physics, Kyoto University
Kyoto 606, Japan

(b): Dipartimento di Fisica
Università di Roma I, *La Sapienza*
Piazza A. Moro 2, 00185 Roma, Italy

(c): Department of Physics, Kyoto University
Kyoto 606, Japan

8/95

[(a)] Also at: Institute for Advanced Study, Princeton, NJ08540, USA.

[(b)] (djl@liocorno.roma1.infn.it)

[(c)] Present address: 3-5-16-203 Nishiasabu, Minato-ku, Tokyo, Japan.

Synopsis: A heteropolymer model of randomly self-interacting chains in two dimensions is studied with numerical simulations in order to elucidate the folding mechanism of protein. We find that the model occasionally shows folding propensity depending on the sequence of random numbers given to the chain. We study the thermodynamic and kinematic roles in the folding mechanism by grouping the local energy minima found in the simulations into clusters according to the similarity of their conformations. It is suggested that the local minima to which some heteropolymers show a folding tendency are always the lowest energy states of the energy spectrum within a cluster, though which cluster is selected depends on the sequence. We also find that folding propensities are correlated with the global structure of the average energy surface, implying a dominant kinetic role in the folding mechanism, although thermal factors cannot be ignored as the mechanism of choosing the ground state within a cluster of states connected by small deformations. We suggest that a hierarchical cluster structure plays an important role in selecting a unique folded state out of the huge number of local minima of heteropolymers.

1. Introduction

The native conformation of protein is determined by the information contained in the linear sequence of amino acids, yet the mechanism of protein folding remains an important puzzle [1]. With a few exceptions, protein forms a folded structure in vitro, and therefore the folding should be determined by chemical and physical principles [2]. There have been long running discussions over the relative importance of thermodynamic and kinetic factors to protein folding: on the one hand, it seems natural to suppose that the native state should be thermodynamically stable with the lowest free energy. On the other hand, as Levinthal illustrated [3], there are a huge number of local energy minima available to the molecular system of protein and some specific mechanism of kinetic character is needed to find a unique state. Based on the kinetic point of view, a few mechanistic models have been proposed for the formation of the folded state of globular proteins [4].

With the hope of elucidating the mechanism of folding, we have been studying a simple model of heteropolymers with the aid of numerical simulations. Several models have been proposed as laboratories for investigating folding. A number of interesting results suggestive of certain aspects of a folding mechanism have been obtained using lattice models [5]. However, we prefer to work with models defined in the continuum, since much of the information regarding the energy barriers that select which molecular deformations are likely may be lost on the lattice. The cost of working in the continuum is that one requires extremely long simulations and that one must deal with the problems of continuum configurations. The specific model we employ is that of self interacting random chains originally proposed by Iori, Marinari and Parisi [6]. This is a model of a linear chain of N sites where each pair of sites is subject to an interaction with strength proportional to a quenched random variable η_{ij}, in addition to an ordinary r^{-6} attractive force and a short-range r^{-12} repulsive hard core. All side chain information is suppressed and there is no inbuilt propensity for local structure formation. It has been shown that this model displays glassy characteristics within some temperature range [6][7][8]. The present authors have shown that not all heteropolymers are alike and that the thermodynamic properties depend on the sequence of random variables η_{ij} appearing in the potential: there are "folding sequences", in which the molecule tends to fall into the ground state as the temperature decreases, and "non-folding sequences", for which there is no favoured state even at low temperature [9]. It was shown that all sequences undergo a continuous phase transition separating the random phase and the globule (or glassy) phase. More interesting is that the folding sequence has another transition at a lower temperature separating the frozen folded phase from the globule phase, and that this transition is apparently discontinuous [9].

In this paper we examine more sequences, our goal being to find physical features that characterise the folding. We use long Monte Carlo runs, taking over a hundred million iterations, and quench the configurations extracted from the runs to zero temperature in order to find the local minimum closest to the conformation realised by the Monte Carlo at that instant. Thousands of local minima are found in this way. We use the word "states" to refer to these local minima. We first hypothesise an operational criterion for folding based on these data using the propensity for the heteropolymer to be in a favoured state during the Monte Carlo runs. We then study the energy level structure to see whether the presence of the gap between the ground state and the excited states correlates with the folding propensity [10]. The motion of the heteropolymers in conformational space is investigated from the same viewpoint. We notice for some sequences a strong kinetic effect that tends to constrict the size of the configuration space visited as the temperature is reduced and that the sequences for which this effect occurs tend to have greater folding propensity according to the operational criterion.

We attempt to coarse-grain some of our measurements by grouping together nearby local minima into clusters with the aid of a friend-of-friends algorithm in the conformational space. We also attempt to characterise properties of the energy surface of the model. The energy surface is of course highly irregular, both in the sense that the hard core part of the potential leads to very high energies when the configuration in any minimum suffers a random perturbation, and also in the sense that the set of minima and the easiest pathways over the barriers separating them, form a very complex and rugged system [11]. We try to distinguish global features of the overall landscape from local ones around the putative folded state. Local features, for example, the energy gap between the ground and excited states belonging to the same cluster, are relevant to the thermodynamic stability of the folded state. Distinct pathways or large scale basins of attraction visible on the global scale would constitute kinetic features. Using the technique of grouping together geometrically similar configurations into clusters, local properties can be investigated relatively directly by looking at the energy levels of local minima that correspond to configurations within the same cluster. The configuration space, of which the energy is a function, is very high dimensional and difficult to visualise, so to gain a feeling for the form of the energy landscape we consider simple one dimensional plots of energy against rms distance from some chosen state. To make this global information useful some of the bumpiness of the surface must be removed by appropriate averaging. We see that the energy surface thus constructed shows a correlation with the folding propensity.

In section 2 we define the model of heteropolymers and describe the method of simulation. We introduce in section 3 an operational criterion for folding and non-folding and discuss other possible indicators of folding: the energy spectrum, the rms deviation of the motion of the polymer in configuration space and its correlation function. Section 4 is the central part of the analysis: we classify the local minima into clusters by grouping similar configurations and study the local energy levels and the global energy surface to elucidate the relative role of thermodynamics and kinetics in folding. Physical conclusions concerning the folding mechanism are given in section 5.

2. The model and method of simulation

The system is described by the partition function:

$$Z = \sum_{\text{all configurations}} e^{-\beta H}, \tag{2.1}$$

where the energy of the self-interacting chain of length N is written [6]:

$$H = \sum_{1\leq i<j\leq N} \left(\delta_{i+1,j}\, r_{ij}^2 - \frac{A}{r_{ij}^6} + \frac{R}{r_{ij}^{12}} + \frac{\epsilon \eta_{ij}}{r_{ij}^6} \right). \tag{2.2}$$

Here r_{ij} is the distance between sites i and j of the chain; the four terms represent, in order, a bond holding adjacent sites together thus forming the chain, a van der Waals attraction, a hard-core repulsion, and the term that models the widely varying van der Waals potentials between different pairs of atoms. This last term contains η_{ij} which is an uncorrelated set of values taken from a uniform distribution of mean zero and variance unity.

We work with 20-site chains in two dimensions. In two dimensions the phase space is greatly restricted so that chain folding is more easily frustrated than in the three dimensional case. We hope that simulations with rather short chains will exhibit some of the behaviour of much longer chains in three dimensions [7]. Eight different representations are taken for η_{ij}, and these systems are labelled A–H. These sequences were chosen completely randomly. From previous studies we fix the parameters of the Hamiltonian (2.2) to be $A = 2, R = 2$ and $\epsilon = \sqrt{10}$. These values ensure that a glassy phase does develop in the model and that we are not merely dealing with a perturbed heteropolymer.

For simulations we use standard Monte Carlo techniques but we apply a deterministic quenching procedure to bring configurations to the nearest local energy minimum and this allows us to accurately classify states. This technique is known as catchment analysis. We work with two temperatures, $\beta = 0.8$ and $\beta = 1.0$. Both these temperatures lie between the random-globule transition at about $\beta \sim 0.6$ and the globule-folded transition which occurs for the typical folding sequence at $\beta \sim 1.1$. A study of the autocorrelation function and associated time history at $\beta = 0.8$ allows us to conclude that this temperature is high enough to admit fast moves from configuration to configuration and to thoroughly explore the kinetically accessible region of configuration space. On the other hand, $\beta = 1.0$ is just above the phase transition for the typical folding sequence (B) and we shall see that the motion is already quite restricted for the sequences with folding propensity.

We have started simulations with the polymer in a straight line configuration. We have made 250 million Metropolis sweeps at $\beta = 1.0$ and 100 million at $\beta = 0.8$ for each sequence and extract configurations every 50000 sweeps at $\beta = 1.0$ (every 20000 at $\beta = 0.8$) for quenching. This provides us with 5000 samples at each temperature. For the two sequences that represent typical folding and typical non-folding properties, we performed additional runs in which we varied the temperature over fairly small steps to study thermal properties. The study of these specific examples has shown that the non-folding characteristic persists to a much lower temperature for one of the two sequences and that the folding characteristics are more conspicuously observed at lower temperatures for the other sequence as reported in [9].

3. Results from Monte-Carlo runs

3.1. Folding and Non-Folding

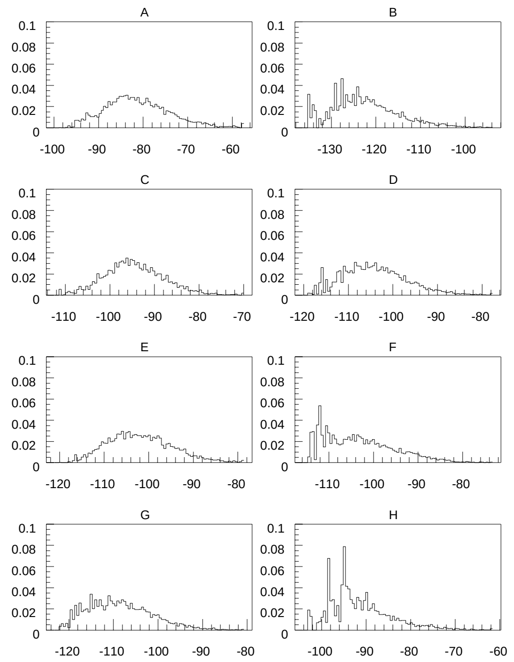

Fig. 1: Probability with which a given energy state is found in runs at $\beta = 0.8$.

We show in Fig. 1 the probability of finding a given energy state in Monte-Carlo runs for $\beta = 0.8$, Figs. 1(A)–(H) refer to the sequences A–H respectively. In each figure the population of the quenched states arising from the catchment analysis of the 5000 Monte Carlo samples is binned. The shape of the distribution differs wildly from sequence to sequence. The distribution in Fig. 1(A) which turns out to represent a typical non-folding sequence, shows a gaussian-like shape in which no particular states are favoured during the run. In Fig. 1(B) in contrast, a number of marked peaks are visible. This figure shows that a substantial amount of time is spent in the lowest energy state (LES). We shall see below that sequence B represents the most typical example of a folding sequence amongst our eight representations.

Results for six further sequences are presented in Figs. 1(C)–(H). Sequences C and E show features similar to A, whereas D resembles B, though the favoured state is not the lowest energy state found in our simulations. Other sequences lie between the two cases. We see that G has features more similar to the non-folder A, and that F and H show folding characteristics to some extent. We arrange the sequences so that non-folding polymers appear on the left-hand side and polymers with some folding propensity appear on the right-hand side of the panels, with the most typical examples at the top.

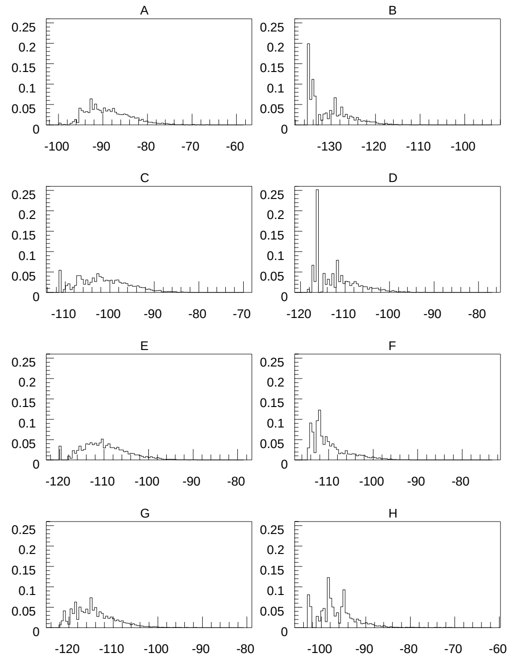

Fig. 2: Probability with which a given energy state is found in runs at $\beta = 1.0$.

Fig. 2 shows similar distributions for the lower temperature run at $\beta = 1.0$. While the features seen in Fig. 1(A) change little, the peaks, especially the one corresponding to the lowest energy state in Fig. 1(B) become more pronounced and in fact the system spends about 20% of its time in the LES. In reference [9], by performing additional runs at lower temperatures, we found that the system becomes frozen in this state (LES) at temperatures below the folding transition. For this reason we take sequence B to be folding.

The shape of the distribution changes little also for C and E although some small peaks are now visible at the lowest energies. In D the probability of finding the system in one low lying state is greatly enhanced as the temperature is reduced. The peak, however, does not correspond to the LES. This LES was rarely visited in the $\beta = 1.0$ simulation.

In the other figures, 1(F) and 1(H), the peaks become significantly higher as the temperature is lowered, but we do not observe a unique selection of any state.

	Sequence							
	A	B	C	D	E	F	G	H
$\beta = 0.8$	0.03	0.05	0.03	0.03	0.03	0.05	0.03	0.08
$\beta = 1.0$	0.06	0.20	0.06	0.25	0.05	0.12	0.07	0.12

Table 1: maximum peak height at at $\beta = 0.8$. (top row) and at $\beta = 1.0$ (bottom row). Normalised to be probabilities.

The peak height information is summarised in Table 1. The maximum peak height is shown; both the absolute value and the ratio between the values at each temperature are relevant to our considerations in distinguishing folding and non-folding. The distinction between our typical non-folding A, and folding B sequences is clear: for the folding sequence, we not only see that the peak height is large at low temperature, but also that the increase of height towards low temperature is dramatic (see B and D).

In conclusion, our model heteropolymers display a variety of different shapes of frequency distribution: some have a character typical of frustrated systems with no favoured state and little change in the distribution as the temperature is lowered. However we sometimes see behaviour reminiscent of the folding polymer, in which the system tends to stay in the ground state with probability that grows rapidly as the system is cooled. The majority of sequences lie between these two extreme cases. Often, a number of peaks emerge at low temperature, showing that these sequences have some weak tendency to fold: it seems rare that no peak at all appears in the distribution.

3.2. Energy level structure

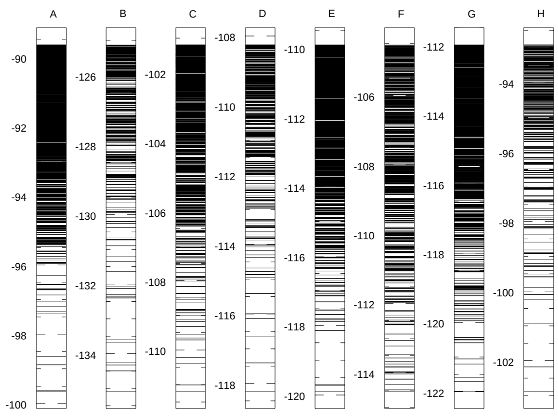

Fig. 3: Spectrum of energy levels.

We study the energy levels of these heteropolymers in Figs. 3(A)–(H), where each panel corresponds to the sequence shown in Figs. 1 and 2. There are several isolated states above the LES, but the number of states rapidly increases with the energy. A difference between the folding and non-folding sequences is not obvious. Note that the non-folding sequence A has a well separated LES, even more distinct than the typical folding sequence B. This demonstrates that the appearance of a distinct ground state, separated from the densely populated excited states, does not give a criterion for folding propensity. We see however that the density of energy levels for the folding sequences B and D are rather low at low energy compared with that for A, C and E.

3.3. Geometrical distance

The rms distance travelled in the simulation with respect to some fixed configuration has been used as an indicator of phase transitions [6], and it might also give us information to distinguish different polymers. The correlations between configurations at different times tells us about the time-scales involved in the dynamics and about the typical distances travelled over. The rms distance between states labeled α and β is defined by:

$$g_{\alpha\beta} = \sum_{\text{sites } i} |\mathbf{r}_i^\alpha - R\mathbf{r}_i^\beta|^2, \tag{3.1}$$

where $\mathbf{r}_i^\alpha$ are the site coordinates (at fixed centre of mass) for the conformation α. The quantity is minimized with respect to the rotation and reflection matrix R for each conformation. We display the rms distance with respect to the most probable state (MPS), the state which appears most frequently in the simulation and corresponds to the highest peak in Fig. 2. The autocorrelation function is defined by the rms distance between states at fixed time difference appropriately averaged over the run:

$$C(t) = \sum_{t_0} g_{\alpha(t_0+t),\beta(t_0)}. \tag{3.2}$$

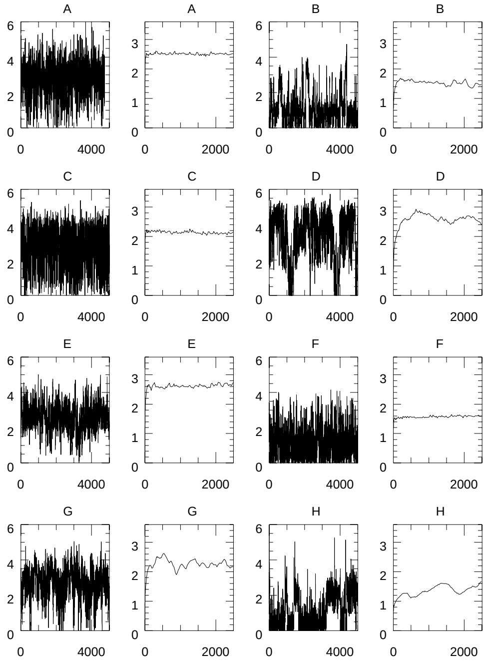

Fig. 4: Time history of rms distance with respect to the most probable state (left hand column) and Autocorrelation function (right hand column) at $\beta = 0.8$. The horizontal time axis is in units of 20000 sweeps.

Results are shown in Figs. 4 and 5 for Monte Carlo runs at $\beta = 0.8$ (Fig. 4) and $\beta = 1.0$ (Fig. 5). The time history $g_{\alpha(\mathrm{MPS}),\beta(t)}$ is shown in the left hand columns and $C(t)$ on the right hand side. Rapid fluctuations are always seen in the time history. At $\beta = 0.8$, although some longer time-scale behaviour can often be seen in the time histories (notably for the sequences D and H), the autocorrelation function reaches a reasonably stable plateau and we take this as evidence that at $\beta = 0.8$ the kinematically accessible region of configuration space is thoroughly explored. This is corroborated by the observation that states and their opposite parity partners are both seen during the course of the $\beta = 0.8$ runs, whereas this is not always the case for the $\beta = 1.0$ runs since potential barriers are sometimes too high. At the lower temperature, $\beta = 1.0$, long time-scale behaviour is more apparent, indicating that the polymer stays around a certain state for a while and then jumps to another state. In the case of the sequences D and H, one large jump is observed (in addition to other flip-flop motion for D). We are not certain, however, whether the states that appear in the earlier stage of the run could be revisited if we made a longer run, or whether this is merely a transient effect in the initial

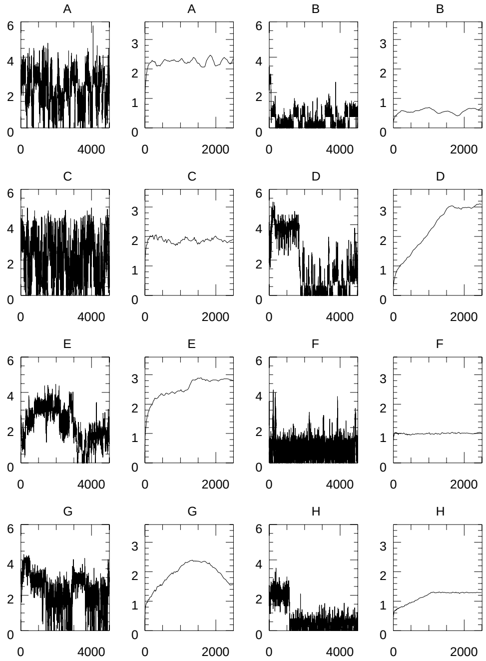

Fig. 5: Time history of rms distance with respect to the most probable state (left hand column) and Autocorrelation function (right hand column) at $\beta = 1.0$. The horizontal time axis is in units of 50000 sweeps.

thermalisation process.

Size of Configuration Space		
Sequence	$\beta = 0.8$	$\beta = 1.0$
A	2.50±0.03	2.24±0.13
B	1.50±0.08	0.55±0.09
C	2.11±0.03	1.88±0.05
D	2.57±0.09	$\sim$
E	2.64±0.05	$\sim$
F	1.59±0.02	1.02±0.01
G	2.24±0.07	$\sim$
H	1.40±0.12	$\sim$

Table 2: Asymptotic value of the mean square deviation autocorrelation function.

The effects of the long time-scale behaviour also show up in the autocorrelation function shown in Figs. 4 and 5 above, giving it a non-trivial behaviour for the sequences with a long time-scale. However, there are several sequences for which our runs are long enough to average over the flip-flop time-scales and for these we see a clear asymptotic plateau in the autocorrelation function. For the cases in which the autocorrelation displays stable behavior we have given its asymptotic value in Table 2. This value varies considerably from sequence to sequence: for the typical non-folder A this asymptotic value is large at both temperatures, whereas for the typical folder B it is not so large at $\beta = 1.0$ and becomes substantially smaller at the lower temperature. These tendencies are visible for other sequences, in so far as the flip-flop time-scale is small enough that the autocorrelation function shows a plateau. In particular, the ratio of the asymptotic value at high and low temperature shows a clear distinction between the sequences A, C and B, F. This indicates that for the sequences with folding propensity, there is some kinetic mechanism at work that rapidly constricts the volume of the portion of configuration space visited as the temperature is reduced.

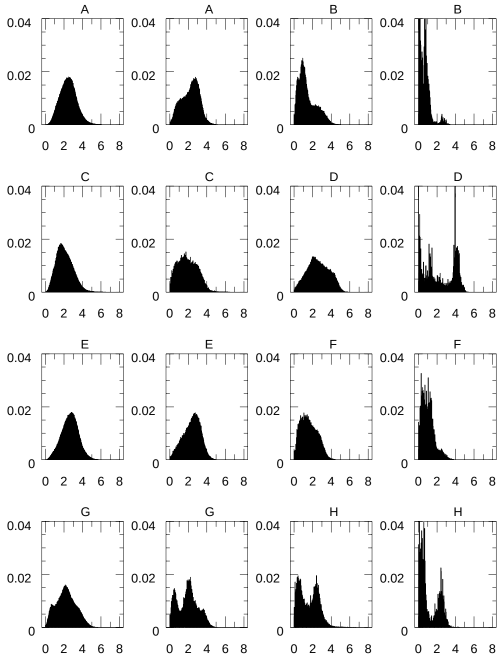

Fig. 6: Histogram for the distribution of $g_{\alpha\beta}$ between each pair of states at both high temperature (left) and low temperature (right).

The histogram for the distribution of $g_{\alpha\beta}$ between all pairs of states taken from the 5000 samples of each run is shown in Fig. 6($\beta = 0.8$ on the left and $\beta = 1.0$ on the right). At high temperature there is little structure in the distributions except for the cases of B and H. At low temperature however, only the definitely non-folding sequences A, C and

E retain their simple shape, the other sequences displaying multiple peaks. The simple structure observed for A,C and E suggests that the configuration space is fairly uniformly explored, with no region receiving special attention. The strong peaks at the origin seen for B,D and to some extent for F and H, indicate that there is a large contribution from identical or almost identical states. Other peaks such as in D, indicate the presence of more than one favoured region in the configuration space.

4. Classification of conformations

4.1. Local energy spectrum

It usually happens that many local energy minima are connected by small deformations of the polymer, and that these local minima form clusters. If such clusters of minima are separated from other clusters by larger deformations it suggests that the structure of the configuration space is hierarchical. If this is correct then the potential surface can be considered at two levels: energy surfaces within a cluster; and those at large scales, coarse grained over clusters. Local features within a cluster are relevant largely to thermodynamic stability. For the cluster containing the folded state, the thermodynamic stability of the folded state may depend on the form of this restricted potential surface, and it is also relevant to understanding the possible fluctuations of the heteropolymer around its folded state [11][12]. On the other hand, the coarse grained, large scale form of the potential is perhaps relevant to large scale motion of the polymer and in particular to the kinetic aspect of the folding mechanism.

In order to distinguish kinetic and thermodynamic aspects of folding we attempt to perform a classification of the states that correspond to local minima into clusters. By introducing a cut-off on the rms distance between two states, we define a cluster as the set of states such that no one state in the cluster is farther than the cut-off from *some* other state in the cluster: the cluster containing a particular state consists of itself, all states within the rms deviation cut-off of that state, all states within the cutoff of those configurations and so on. This is known as the “friend of friends” scheme. Any member of a cluster is thus separated from any member of any other cluster by at least the cutoff.

We should also introduce another cut-off in the energy space, so that the two states in one class are connected via not too highly excited states in the “friend of friends” algorithm. We take this cut-off to be ten energy units above the LES.

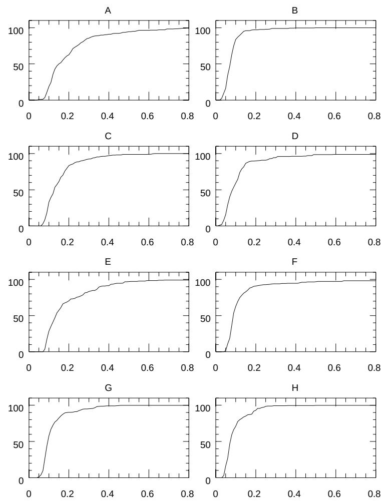

Fig. 7: Percentage of the number of states which belong to clusters whose population is larger than 1% of the total number of states. Shown as a function of the rms deviation cutoff.

In order to gain an idea as to a reasonable value of the cut-off for the rms deviation, we have plotted in Fig. 7 the fraction of the number of states which belong to clusters whose population is larger than 1% of the total number of states. It is interesting to observe that the increase of the fraction for B and D (also for F and H) is faster than for A, C, and E which are non-folding sequences. This indicates that configuration space of the folding sequence is more strongly clustered than that of the non-folding sequences. We shall work with a cut-off distance of $r_{\rm cut} = 0.15$ which corresponds to 50-90% of the fraction in Fig. 7.

	Sequence							
	A	B	C	D	E	F	G	H
States	1046	320	793	367	1129	640	975	306
Clusters	311	16	190	54	263	80	103	38

Table 3. Number of states (1st row) and clusters (2nd row) appearing in the spectra.

In Table 3 we show the number of states found in the course of the runs up to 10 energy units above the LES, and correspondingly, the number of clusters obtained with the friend-of-friends scheme taking the above cut-off. The table demonstrates that the folding sequences, especially B and D, are indeed more strongly clustered.

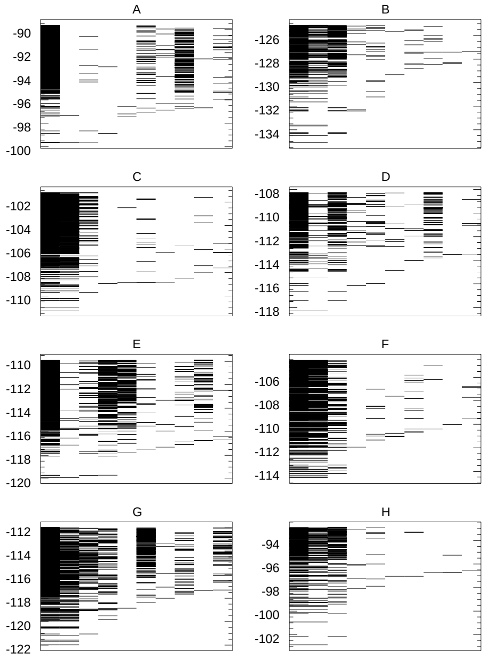

Fig. 8: Spectra of energy levels within each cluster. For each sequence the left hand column shows the complete spectrum as in Fig. 3. The other columns correspond to the first nine clusters arranged according to the energy of the lowest energy state in the cluster.

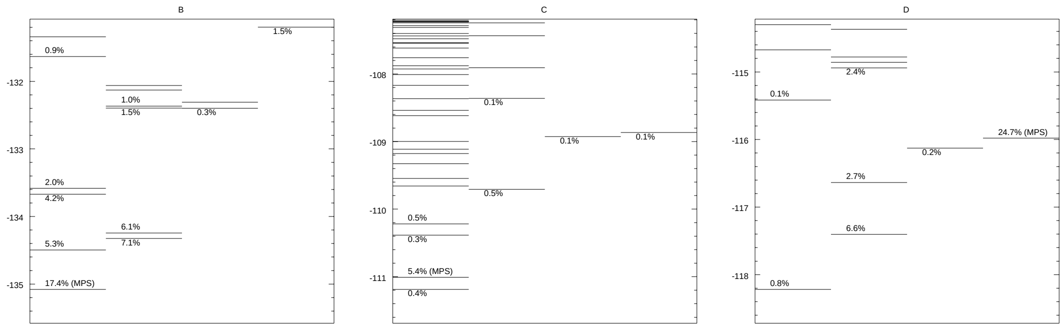

Fig. 9: Close up of the low lying levels of the energy spectrum within clusters (Fig. 8) shown for sequences B, C and D.

The energy level structure within each cluster is shown in Fig. 8: the tower of energy levels corresponds to that in each cluster, i.e., energy levels of "local excitations". The towers are ordered according to the energy of the lowest energy state of each tower.

We find that most of the notable peaks appearing in the frequency distribution of Figure 2 correspond to the lowest excited states (or low lying excited states) of the clusters.

Three examples (B,D and C) are depicted in Fig. 9, which is similar to Fig. 8, but only the important parts are reproduced and the probability of visiting these specific states are indicated. In our prototypical folding sequence B, the relative probability distribution is just as we expected: the largest probability peaks are always found in the lowest energy state of each cluster, with the highest probability corresponding to the true LES. In D, the conspicuous peak seen in Fig. 2(D) corresponds to the lowest energy level of the fourth cluster, while the lowest energy level of the first cluster is visited only a small fraction (0.8%) of the time. In C, the most probable state appears as the second lowest state, but we see that the LES and this state are separated by 0.55 rms units of distance; a slightly smaller choice of the cut-off separates these two states into different clusters.

For other sequences we find that the most probable state occasionally appears at some excited level (e.g. sequences F and H). We find, however, that the cluster containing the most probable state is a large cluster in the sense that a slight reduction of $r_{\rm cut}$ produces many sub-clusters. From our observations we would conjecture that the notable peaks that appear in the frequency distribution always correspond to the lowest energy level of some cluster with an appropriate cut-off distance.

Unfortunately we cannot clearly distinguish the features of folding and non-folding sequences from these energy level diagrams. At least it is clear that a large energy gap between the lowest and the first excited state within a cluster does not necessarily mean that system has a greater folding propensity.

4.2. Global structure of the energy surface

The energy surface resides in a multi-dimensional space rendering it difficult to visualise. Here we attempt to display the energy surface in one dimensional space, integrating out most of the information. The simplest approach is to plot the energy against the rms distance from a fixed reference state: we average over all states ($E_{\rm LES} < E < \infty$) that are a given radial distance up to some binning width from the reference state. We also apply smoothing, to avoid too rugged a landscape: this smearing consists of taking all states within the cluster containing the reference state and averaging over them. We choose the reference state to be the LES and take the cut-off parameter that defines the cluster to be $r_{\rm cut} = 0.1$. This diagram corresponds to the reduced energy surface approximately corresponding to the average of the energy.

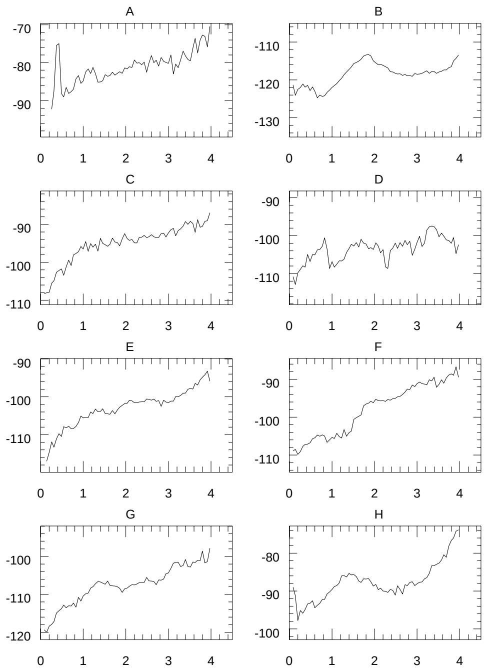

Fig. 10: Reduced energy surface at $\beta = 0.8$, as a function of the rms distance from the LES.

In Figs. 10 and 11 we show such plots for the $\beta = 0.8$ and $\beta = 1.0$ data. The curve of energy shows a global increase as a function of the rms distance from the LES, showing that the cluster containing the LES mostly coincides with the global free energy minimum. The general observation to be made from studying these plots is that the sequences *B, D* and *F*, which have folding tendencies, develop large scale structure, while the non-folding sequences *A, C* and *E* have little such structure. The surface for the sequences *A, C* and *E* display similar behaviour at both temperatures: they are rugged on the short scale, but are smooth on the large scale. We note that the large narrow peak at small distance in sequence *A* comes from the fact that the smearing is not sufficient — the class that contains the LES is so small that the smearing has little effect.

A conspicuous large scale structure is visible in B for both $\beta = 1.0$ and $\beta = 0.8$: global potential basins are separated by large barriers. Similar structure is also visible for *D* at $\beta = 1.0$. It is worthwhile to note that such clear structure is not seen at $\beta = 0.8$, in fact the characteristic peak is not so conspicuous at $\beta = 0.8$ (Fig. 1(D)) and it grows rapidly towards $\beta = 1.0$, (Fig. 2(D)). Another interesting case is *H*: the notable structure seen at $\beta = 0.8$ disappears at $\beta = 1.0$. Correspondingly, we note that conspicuous peaks observed at $\beta = 0.8$ (Fig. 1(H)) do not grow significantly as the temperature is lowered to $\beta = 1.0$

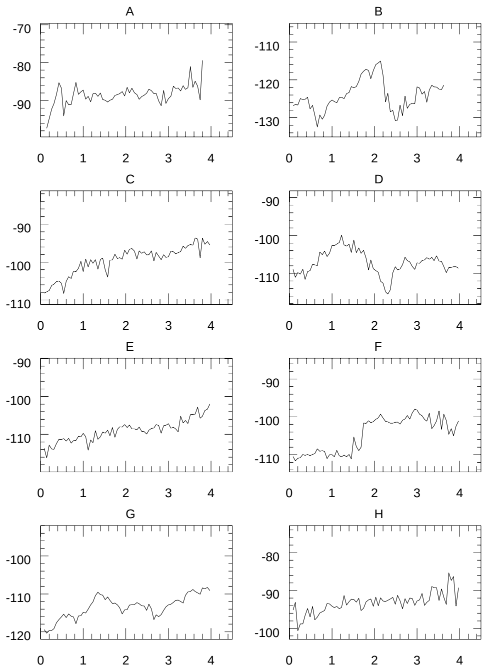

Fig. 11: Reduced energy surface at $\beta = 1.0$, as a function of the rms distance from the LES.

(Fig. 2(H)). We also remark a peculiar behaviour of the free energy surface in F. We see a wide and flat basin at $\beta = 1.0$, suggesting that unique selection of the LES is hampered by thermodynamical fluctuations even at low temperature.

We conclude that the large scale structure of the energy surface is correlated with the appearance of peaks in the frequency distribution. We expect to have folding sequences when the large scale structure of the potential surface develops towards lower temperatures.

5. Conclusion and discussion

We have seen that heteropolymers are occasionally endowed with some folding propensity. In spite of a very large number of local minima expected in such heteropolymers, some of the configurations are preferred over others when the temperature is lowered. We may call the most conspicuous case a folding heteropolymer. We find that these favoured states generally correspond to the lowest energy level among a number of local minima which are connected by small deformations of the polymer chain. It sometimes happens that the most probable state coincides with the true ground state of the system, but there is also the case where the most probable state is not the ground state of the system, but simply a metastable state kinematically well separated from the ground state. We cannot tell whether both cases correspond to native protein or whether only the former corresponds to protein.

This observation suggests that the clustering structure of the local minima plays an important role in the folding: it facilitates the selection of unique or semi-unique states out of the huge number of local minima. Probably, clustering does not take place at one level, but large clusters contain a number of smaller sub clusters, which may also be subdivided as in the Parisi tree structure found in spin glass models [11][13].

Our analysis clearly indicates that the energy level structure, for instance whether the excited states are well separated from the LES, has little to do with folding propensities, in contrast to what has been suggested in [10]. The difference may arise from the difference between the models, those on the lattice and those in the continuum. It seems that more important is the size of the configuration space visited by polymers. Our analysis suggests that for folding sequences it is severely constricted as the temperature is reduced.

We suggest that the folding propensity is correlated with the large scale structure of free energy surfaces and more importantly whether this structure develops towards lower temperatures. This seems to corroborate the hypothesis that the folding is dominantly controlled by the kinematic mechanism, although thermal stability plays a role locally in the configuration space.

It is clear that our heteropolymer chains are still too short to really mimic the reality. We have also worked in only two dimensions. We may ascribe the frequent appearance of peaks in the frequency distribution to a small size effect. We may speculate that the majority of heteropolymers behave more like sequence A as the length N increases and the appearance of folding sequences becomes less and less likely. A simulation for large N polymers, however, requires much longer Monte Carlo runs, in addition to a larger set of representations to obtain a folding sequence.

Acknowledgments

DJL and MGM thank the Yukawa Institute for Theoretical Physics for hospitality.

References

[1] Karplus, M. and Shakhnovich, E. (1992) ‘Protein Folding: Theoretical studies of Thermodynamics and Dynamics’ in Protein Folding, Ed. Creighton, T. ;
Wolynes, P.G. ‘Spin Glass Ideas and the Protein Folding Problem’ in Spin Glasses and Biology, ed. Stein, D., World Scientific.

[2] Anfinsen, C.B. (1973) Science **181**, 223.

[3] Levinthal, C. (1968) J. Chim. Phys. **65**, 44.

[4] Karplus, M. and Weaver, D.L. (1976) Biopolymers **18**, 1421 ;
Karplus, M. and Weaver, D,L. (1976) Nature **266**, 404.

[5] Go, N. and Taketomi, H. (1978) Proc. Nat. Acad. Sci. USA **75**, 559 ;
Miller, R., Danko, C.A., Fasolka, M.J., Balazs, A.C., Chan, H.S. and Dill, K. (1992) J. Chem. Phys. **96**, 768 ;
Garel, J.R., Garel, T. and Orland, H. (1989) J. Phys (France) **50**, 3067 ;
Shakhnovich, E., Farztdinov, G., Gutin ,A.M. and Karplus, M. (1991) Phys. Rev. Lett. **67**, 1665.

[6] Iori, G., Marinari, E. and Parisi, G. (1991) J. Phys. **A24**, 5349.

[7] Fukugita, M., Lancaster, D. and Mitchard, M.G. (1992) J. Phys. **A25**, L121.

[8] Pliszka, P. and Marinari, E. (1993) Europhys Lett. **22**, 167 ;
Iori, G., Marinari, E. and Parisi, G., work in progress.

[9] Fukugita, M., Lancaster, D. and Mitchard, M.G. (1993) Proc.Natl.Acad.Sci.USA. **90**, 6365.

[10] Sali, A., Shakhnovich, E. and Karplus, M. (1994) J. Mol. Biol. **235**, 1614.

[11] Austin, R.H., Beeson, K., Eisenstein, L., Frauenfelder, H., Gunsalus, I.C. and Marshall, V.P. (1973) Phys.Rev.Lett. **32**, 403; ;
Iben, I.E.T., Braustein, 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.

[12] Go, N (1983) Ann. Rev. Biophys. Bioeng. **12**, 183.

[13] Mézard, M., Parisi, G. and Virasoro, M.A. (1987) “Spin Glass Theory and Beyond”, World Scientific.