Delineation of the native basin in continuum models of proteins

Mai Suan Li and Marek Cieplak
Institute of Physics, Polish Academy of Sciences, Al. Lotnikow 32/46, 02-668 Warsaw, Poland

Received 12 April 1999, in final form 26 May 1999

Abstract. We propose two approaches for determining the native basins in off-lattice models of proteins. The first of them is based on exploring the saddle points on selected trajectories emerging from the native state. In the second approach, the basin size can be determined by monitoring random distortions in the shape of the protein around the native state. Both techniques yield similar results. As a byproduct, a simple method to determine the folding temperature is obtained.

Chains of beads on cubic or square lattices, with some effective interactions between the beads, often serve as simple models of proteins (see for instance [1]). A more realistic modelling, however, requires considering off-lattice systems. Simple off-lattice heteropolymers have been discussed recently by Iori *et al* [2], Irback *et al* [3], Klimov and Thirumalai [4] and by the present authors [5]. The purpose in using such models is to understand the basic mechanism of folding to the native state. In lattice models, the native state is usually non-degenerate and it coincides with the ground state of the system. Delineating boundaries of the native basin in off-lattice systems, however, is difficult, especially when the number of degrees of freedom is large, yet it is essential for studies of almost all equilibrium and dynamical properties of proteins. For instance, stability of a protein is determined by estimating the equilibrium probability to stay in the native basin: the temperature at which this probability is $\frac{1}{2}$ defines the folding temperature, T_f. The native basin consists of the native state and its immediate neighbourhood, as shown schematically in figure 1, and it should not be confused with the whole folding funnel. The latter involves a much larger set of conformations which are linked kinetically to the native state.

In most studies, such as in [3, 4], the size of a basin is declared by adopting a reasonable but ad hoc cutoff bound. Systematic approaches, however, are needed and will be presented here. The task of delineating of the native basin is facilitated by introducing the concept of a distance between two conformations a and b, δ_{ab}. The distance should be defined in a way that excludes effects of an overall translation or rotation. There are two definitions of δ_{ab}, for a sequence of N monomers, that we shall use. The first one is [2, 6]:

$$\delta_{ab}^2 = \min \frac{1}{N} \sum_{i=1}^{N} |\vec{r}_i^{\,a} - \vec{r}_i^{\,b}|^2 \tag{1}$$

where $\vec{r}_i^{\,a}$ denotes the position of monomer i in conformation a and the minimization is performed over translations, rotations and reflections. In practice, we put chain a over chain b by overlapping the two centres of mass, and then we find the optimal rotation of b which minimizes δ_{ab}.

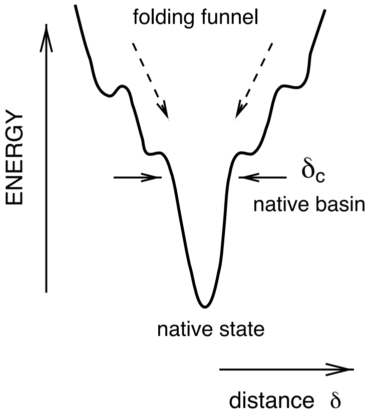

Figure 1. A schematic plot of the energy in the vicinity of the native state. δ_c denotes a characteristic size of the basin.

The second is

$$\delta'^2_{ab} = \frac{1}{N^2 - 3N + 2} \sum_{i \neq j, j \pm 1} (|\vec{r}_i^a - \vec{r}_j^a| - |\vec{r}_i^b - \vec{r}_j^b|)^2. \tag{2}$$

The first definition is closer to an intuitive understanding of the distance between shapes whereas the second is easier to compute, especially in three-dimensional situations. Nevertheless, both are expected to be physically equivalent.

In this paper, we develop two techniques for determining the basin of the native state. In the first approach, we generate an image of the phase space by sampling it by low-energy trajectories that start from the native state and continue by displacing the monomers in a way that increases the distance away from the native state while preserving the connectedness of the chain. For each trajectory, a dependence of the potential energy on the distance away from the native state is obtained and locations of the saddle points are determined. The average distance to a first encountered saddle point, $\langle \delta_s \rangle$, may be considered as characterizing the size of the basin.

In the second technique, which we find to be more statistically reliable and more automatic in its implementation, we adopt a variant of de Gennes's idea of the 'ant in a labyrinth' [7]. Specifically, we characterize the geometry of the native basin by monitoring random shape distortions of the heteropolymer in the basin. The distortions are induced by diffusive-like displacements of individual beads. The method of the shape distortion is implemented by first placing the polymer in an initial conformation, $a(0)$, which usually coincides with the native state. Subsequently, one performs random displacements of individual beads in the chain, through a Monte Carlo routine, and the conformation at time t acquires a shape $a(t)$. The process is characterized by determining the evolution of a mean square distance, $\langle \delta^2 \rangle_t = \langle \delta^2_{a(0)a(t)} \rangle$ between $a(t)$ and $a(0)$. The focus is on short time behaviour and the average is over many trajectories that start from the same $a(0)$. The characteristic size of the basin, δ_c is obtained by studying features in $\langle \delta^2 \rangle_t$ as described later. In order to scale the walls of the basin one needs to make the polymer 'crawl' up these walls without involving any kinetic energy. Thus the process does not correspond to any 'real' evolution in time, as defined e.g. through the molecular dynamics. As opposed to the first technique, the fluctuations in the shape of the polymer are understood to be due to coupling to a heat bath.

Models

In order to illustrate the two techniques, we consider a two-dimensional version of the model introduced by Iori *et al* [2]. The Hamiltonian is given by

$$H = \sum_{i \neq j} \left\{ k(d_{i,j} - d_0)^2 \delta_{i,j+1} + 4 \left[\frac{C}{d_{i,j}^{12}} - \frac{A_{ij}}{d_{i,j}^{6}} \right] \right\} \tag{3}$$

where i and j range from 1 to the number of beads, N, which in our model is equal to 16. The distance between the beads, $d_{i,j}$ is defined as $|\vec{r}_i - \vec{r}_j|$, where $\vec{r}_i$ denotes the position of bead i, and is measured in units of the standard Lennard-Jones length parameter σ. The harmonic term in the Hamiltonian couples the adjacent beads along the chain. The remaining terms represent the Lennard-Jones potential. In [2] A_{ij} is chosen as $A_{ij} = A_0 + \sqrt{\epsilon}\eta_{ij}$, where A_0 is constant, η_{ij} are Gaussian variables with zero mean and unit variance; ϵ controls the strength of the quenched disorder. The case of $\eta_{ij} = 0$ and $A_0 = C$ would correspond to a homopolymer. Our choice for the values of A_{ij} is that all of the A_{ij} are positive, which corresponds to attraction. We measure the energy in units of C and consider k to be equal to 25 in units of C/σ^2. Smaller values of k may violate the self-avoidence of the chain [5].

We focus on two 16-monomer sequences in two dimensions which were characterized in detail previously [5]. One of them, G, is a good folder and the other, R', is a bad folder. We used two criteria for the quality of folding: (1) based on the evaluation of the specific heat and structural susceptibility [8] and (2) based on the location of the folding temperature, T_f, relative to the temperature at which the onset of the glassy effects takes place [9]. G has been designed as an off-lattice go-like sequence [10] and the A_{ij} are taken to be 1 for native contacts and 0 for non-native ones. The target conformation is on lattice and is shown in figure 1(*a*). This target is the same as for the lattice sequence R of [11, 12]. The ground state of G has approximately the shape of the target—note that the interaction between two beads forming a contact also slightly affects other beads. The bad folder R' is constructed following the rank-ordering procedure which is an off-lattice analogue of what was done in [11, 12]. We choose the equlibrium interbead distance, d_0 of $2^{1/6}$ and 1.16 (the latter value is determined by the average of A_{ij} over all couplings, see [5]) for G and R', respectively. The target conformation is chosen initially to be the same as in figure 1(*a*). The most strongly attractive Lennard-Jones interactions are assigned to the nine native contacts. They are enhanced by making the corresponding A_{ij} bigger than 1. The remaining couplings have A_{ij} which are smaller than 1. Rank ordering of the contacts generates good folders among lattice models. In the off-lattice models, however, the non-native Lennard-Jones interactions, due to their long range nature, overconstrain the system and frustrate it leading to a deterioration of folding properties. The values of A_{ij} for R' are listed in [5]. The resulting native state of R' is shown in figure 1(*b*).

Figures 2(*c*) and (*d*) show the distributions of distances δ for local energy minima of sequences G and R', respectively. The native states and the local energy minima have been obtained by multiple quenches from random conformations. Note that the sequence G has a much smaller number of local energy minima compared with sequence R' and the energy gap to the first excited local minimum is significantly larger. In each case, there is one minimum which has the closest geometrical distance, δ_{min} to the native state. δ_{min} is thus an upper bound for the size of the native basin (0.5 for G and 0.3 for R').

Trajectories in the energy landscape

The first technique is implemented by creating the trajectories from the native state at $T = 0$ in a stochastic way. The bead positions are displaced randomly and the moves are accepted if

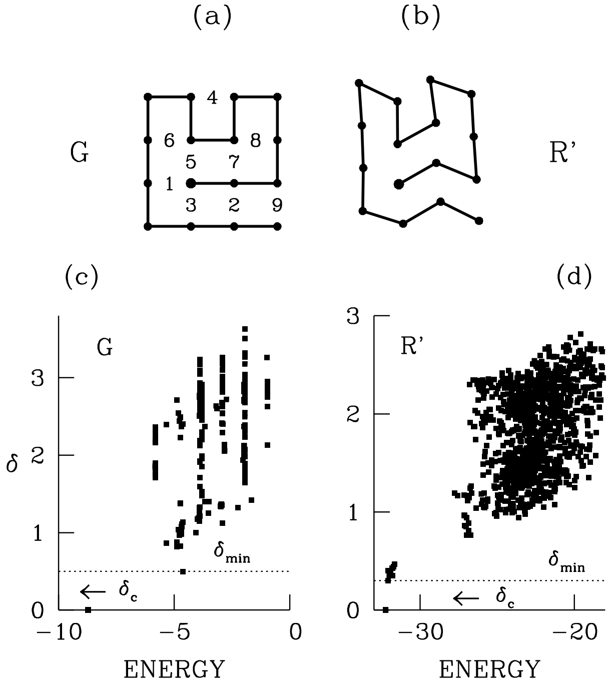

Figure 2. (*a*) The target native conformation for sequence G. Its energy is equal to -8.716 in units of C. The assignment of the couplings A_{ij} used to construct sequence R'. The numbers indicate the relative strengths of the contacts. (*b*) The native conformation of sequence R' is also shown. (*c*) The distances δ from local minima to the native state, for sequence G, shown as a function of energy of the minima. The dotted line corresponds to the minimal distance δ_{min}. The horizontal arrow indicates the size of the basin boundary, as obtained from studies of the shape distortions. (*d*) The same as (*c*) but for sequence R'.

they increase the geometrical distance to the native state and keep the distance between nearest beads in the interval $1 < d_{i,i+1} < 1.1d_0$. In addition, one has to keep the interaction energy for any pair of monomers in sequence R' to be negative. For G, however, the non-native pairs interact only repulsively. In order to minimize the repulsion as much as possible we explore only those trajectories which keep the distances between the monomers of non-native contacts sufficiently large. We choose the minimal distance between the beads of non-native contacts, d_m, to be $d_m = 1.5$ (the choice of 1.6 for d_m yields similar results) and reject trajectories which lead to a monotonic increase in energy even after entering into the region of overall positive energies. Smaller values of d_m usually lead to trajectories with unreasonably large positive energies. Substantially larger values of d_m generate short trajectories which terminate on a conformation which does not allow for a further increase in the distance.

Figure 3 shows typical trajectories for G and R'. For each trajectory, we define the postion of the saddle point δ_s. This point appears as a local maximum on the energy versus δ curve. The average value, $\langle\delta_s\rangle$, defines the basin size. Averaging over 50 trajectories we obtain $\langle\delta_s\rangle = 0.17 \pm 0.02$ and 0.12 ± 0.01 for sequences G and R', respectively.

It should be noted that the technique described here is conceptually simple and easy to implement for a small number of trajectories but its resulting statistics on the basin size are inherently poor since there is no convenient control over the choice of important trajectories.

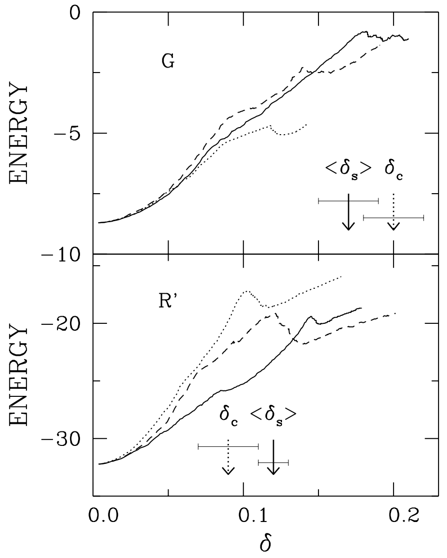

Figure 3. Typical trajectories from the native state generated by the first method. The solid and dotted arrows denote the basin sizes $\langle\delta_s\rangle$ and δ_c obtained by the first and second approaches, respectively.

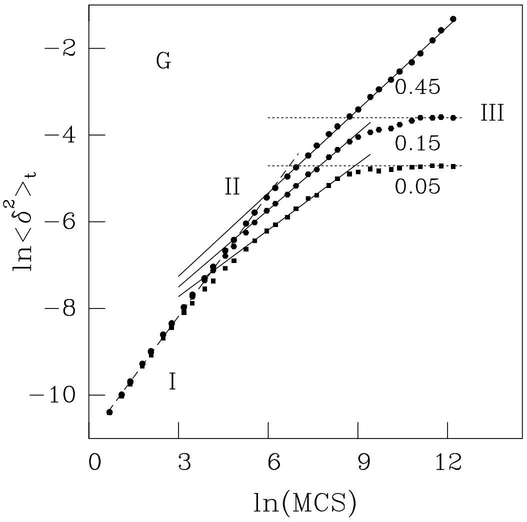

Figure 4. The dependence of the square of the distance to the native state on the Monte Carlo time for G for several values of the temperature which are shown next to the curves.The dashed, solid and dotted lines correspond to regimes I, II, and III respectively. The results are averaged over 400 trajectories. The error bars are smaller than the symbol sizes. The optimal rotation, required in equation (1), is picked from 1000 discrete values into which the full 360° angle is divided.

Following patterns in the force field deterministically is a possible improvement but another stochastic approach, described below, combines simplicity with reliability of the results.

Fluctuations in the shape of the polymer

In order to compute $\langle\delta^2\rangle_t$ we update the monomer positions randomly within circles of radius of 0.01 (the choice of 0.02 yields similar results). We assume that the system is in contact with a heat bath corresponding to temperature T which provides a controlling device. Figure 4 shows the dependence of $\langle\delta^2\rangle_t$ on t for sequence G. We observe that there are, in general, three regimes of behaviour of $\langle\delta^2\rangle_t$. For sequence G, all of the three regimes appear below $T_c = 0.19$ and are shown, in figure 4, for $T = 0.05$ and 0.15.

In the first regime, I, corresponding to very short timescales during which only several per

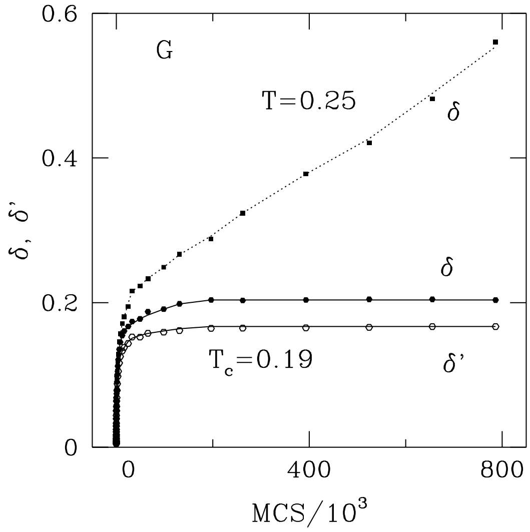

Figure 5. The time dependence of $\delta = \langle\delta^2\rangle_t^{1/2}$ (black hexagons) and $\delta' = \langle\delta'^2\rangle_t^{1/2}$ (open hexagons) for sequence G at the critical temperature $T_c = 0.19$. For $T = 0.25$, δ (black squares) does not saturate below the the distance equal to the minimal distance, $\delta_{min} = 0.5$, between the native state and the local minimum which is the nearest geometrically (see figure 2).

cent of a linear size of the basin are covered, one has a power law

$$\langle\delta^2\rangle_t \sim t^{\nu_0}. \tag{4}$$

Pliszka and Marinari [13] have demonstrated that ν_0 is sensitive to the details of the Hamiltonian. We find that ν_0 is equal to 0.96 ± 0.02 and 0.94 ± 0.02 for sequences G and R' respectively. Both values are close to 1 and stay the same for all T (even for T up to 50) which suggests a simple diffusive behaviour.

In the second regime, II, the plot of $\langle\delta^2\rangle_t$ versus t acquires a T-dependence. Sommelius [6] has focused on this regime and has postulated that the behaviour here appears to follow a power law, at least approximately. The corresponding exponent ν is then found to depend on T in a way that relates to characteristic temperatures of the system. We have found, however, that this power law is not robust—the effective exponent depends on the size of the steps in which one implements distortions. More importantly, the spatial extent of this regime is necessarily limited, and it would probably remain so even for very long heteropolymers.

In the third regime, III, observed only below a critical temperature T_c, $\langle\delta^2\rangle_t$ saturates at a constant value as explained in detail in [5]. Above T_c, the shape distortion ceases to be confined to the native basin and the type-II behaviour continues to take place, as illustrated by the data points corresponding to $T = 0.45$ in figure 4. The limiting saturation value of $\sqrt{\langle\delta^2\rangle}$ at T_c defines a characteristic basin size, δ_c, that can be used, e.g., when deciding if a folding took place if the system started to evolve from and unfolded state. Naturally, T_c is a measure of the folding temperature T_f which characterizes thermodynamic stability of the system.

Figure 5 switches from the logarithmic scale of figure 4 to the linear scale in which the transition between the staturation and lack of saturation in $\langle\delta^2\rangle_t$ shows in a more convincing way. Figure 5 compares the time dependence of $\langle\delta^2\rangle_t$ to that of $\langle\delta'^2\rangle_t$ for G at T_c and demonstrates that the actual choice of the definition of the distance has small effect on the results. For sequence G we have $\delta_c = 0.2 \pm 0.02$ (the second definition of the distance yields $\delta'_c = 0.17 \pm 0.02$) For sequence R' we find that $T_c \approx 0.09$ and $\delta_c = 0.09 \pm 0.02$. Within the error bars, δ_c is close to $\langle\delta_s\rangle$ defined by the first approach which is also indicated in figure 3. For both sequences, the values of δ_c and $\langle\delta_s\rangle$ are smaller than δ_{min}. Thus *the saturation value of the distance to the native state at $T = T_c$ may indeed serve as a measure of size of the native basin δ_c*, i.e. $\delta_c = [\langle\delta^2\rangle_{sat}(T = T_c)]^{1/2}$ determines the true boundary of the native basin.

The fact that δ_c for G is found to be larger than for R' indicates greater stability of G relative to R'. This also correlates well with the higher value of T_c which in turn suggests that T_c is a measure of T_f—the folding temperature. This interpretation is confirmed by comparing the values of T_c to T_f obtained by calculating the probability to be within the cutoff distance, δ_c, away from the native state [5] and by studying positions of the peaks in the structural susceptibility [5]. These studies yield T_f of 0.24 ± 0.03 and $\approx$0.12 for G and R', respectively. Both of these values are close to T_c obtained from the shape distortion. It should be noted that the calculation of T_c by monitoring stochastic shape distortions is significantly less involving computationally. We have also used this technique to determine the sizes of basins of low-lying local energy minima. The corresponding values of δ_c are found to be smaller than for the native state.

We now consider, following Struglia [14], the basin of attraction for random initial conformations which are subsequently quenched in the steepest descent fashion. The basin of attraction is defined in terms of a distance at which the probability to fall onto the native state is bigger than or equal to p_c. The corresponding basin size will be denoted by δ_f. In [14], p_c was taken to be equal to $\frac{1}{2}$. In our studies, we took $p_c = 1$ and determined the basin of attraction for sequences G and R' through a standard quenching procedure. We considered 200 trajectories and obtained $\delta_f \approx 0.55$ and ≈ 0.35 for sequences G and R', respectively. These values exceed not only our estimate of δ_c but also that of the minimal distance between the native state and the nearest minimum δ_{min}. The values of δ_f would become even larger for a p_c that was less than 1. This emphasizes the point that the procedure used by Struglia probably delineates the folding funnel and not the native basin itself.

In summary, we have explored the native basins of two off-lattice sequences by monitoring the shape distortion and by exploring the saddle points of the trajectories from the native state. We have devised with computationally simple methods to delineate the boundaries of the basins and to estimate the folding temperature. The bad and good folders are found to have native basins which are comparable in size even though the structure of their folding funnels must be very different.

Acknowledgments

We thank T X Hoang for discussions. This work was supported by Komitet Badan Naukowych (Poland; grant number 2P03B-025-13).

References

[1] Dill K A, Bromberg S, Yue K, Fiebig K M, Yee D P, Thomas P D and Chan H S 1995 *Protein Sci.* **4** 561

[2] Iori G, Marinari E and Parisi G 1991 *J. Phys. A: Math. Gen.* **24** 5349

[3] Irback A, Peterson C and Pottast F 1996 *Proc. Natl Acad. Sci., USA* **93** 9533
Irback A, Peterson C and Pottast F 1997 *Phys. Rev.* E **55** 860
Irback A, Peterson C, Pottast F and Sommelius O 1997 *J. Chem. Phys.* **107** 273

[4] Klimov D K and Thirumalai D 1997 *Phys. Rev. Lett.* **79** 317
Veitshans T, Klimov D K and Thirumalai D 1997 *Fold. Des.* **2** 1

[5] Li M S and Cieplak M 1999 *Phys. Rev.* E **59** 970

[6] Sommelius O 1997 *Preprint* cond-mat/9703227

[7] de Gennes P G 1976 *La Recherche* **7** 919

[8] Camacho C J and Thirumalai D 1993 *Proc. Natl Acad. Sci.* **90** 6369
Klimov D K and Thirumalai D 1996 *Phys. Rev. Lett.* **76** 4070

[9] Socci N D and Onuchic J N 1994 *J. Chem. Phys.* **101** 1519

[10] Go N and Abe H 1981 *Biopolymers* **20** 1013

[11] Cieplak M, Vishveshwara S and Banavar J R 1996 *Phys. Rev. Lett.* **77** 3681

[12] Cieplak M and Banavar J R 1997 *Fold. Des.* **2** 235
[13] Pliszka P and Marinari E 1993 *Europhys. Lett.* **22** 3
[14] Struglia M V 1995 *J. Phys. A: Math. Gen.* **28** 1469