Structural susceptibilities in toy models of proteins

2000 J. Phys. A: Math. Gen. 33 7699

(http://iopscience.iop.org/0305-4470/33/43/303)

J. Phys. A: Math. Gen. **33** (2000) 7699–7707. Printed in the UK PII: S0305-4470(00)15839-1

Structural susceptibilities in toy models of proteins

Mai Suan Li

Institute of Physics, Polish Academy of Sciences, Al. Lotnikow 32/46, 02-668 Warsaw, Poland
and
Institut für Theoretische Physik, Universität zu Köln, Zülpicher Straße 77, D-50937 Köln, Germany

Received 25 July 2000, in final form 14 September 2000

Abstract. New definitions of the structural susceptibilities based on the fluctuations of distances to the native state of toy protein models are proposed. The calculation of such susceptibilities does not require the basin of the native state and the folding temperature can be defined from the peak if the native conformation is compact. The number of peaks in the derivatives of distances to the native state with respect to temperature, when plotted versus temperature, may serve as a criterion for foldability. The thermodynamic quantities are obtained by Monte Carlo and molecular dynamic simulations.

The understanding of many aspects of protein folding has been recently advanced through studies of toy lattice models [1, 2]. A more realistic modelling, however, requires the consideration of off-lattice systems. In lattice models, the native state is usually non-degenerate and it coincides with the ground state of the systems. In the case of off-lattice models the native state has a zero measure, and delineating boundaries of the native basin in off-lattice systems is vital for studies of almost all equilibrium and dynamical properties. For instance, stability of a protein is determined by estimating the equilibrium probability of staying in the native basin: the temperature at which this probability is $\frac{1}{2}$ defines the folding temperature, T_f.

In most studies, such as in [3–5], the size of a basin, δ_c, is declared by adopting a reasonable but *ad hoc* cutoff bound. In [6], for instance, the folding kinetics are studied by monitoring the number of native contacts. The definition of the native contacts remains, however, ambiguous because it depends on the choice of the cutoff distance. We have developed two systematic approaches to delineate the native basin [7]. One of them is based on exploring the saddle points on selected trajectories emerging from the native state. In the second approach, the basin is determined by monitoring random distortions in the shape of the protein around the native state. It should be noted that the implementation of these methods becomes difficult in the case of long chains. The question we ask in this paper is what one can learn about the folding thermodynamics and the foldability of the off-lattice sequences without knowledge of δ_c.

We start our discussion by introducing the following distances to the native state:

$$\delta_d = \sqrt{\frac{2}{N^2-3N+2}\sum_{i\neq j,j\pm1}(d_{ij}-d_{ij}^{NAT})^2}$$
$$\delta_{ba} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N-2}(\theta_i-\theta_i^{NAT})^2} \qquad \delta_{da} = \sqrt{\frac{1}{N-3}\sum_{i=1}^{N-3}(\phi_i-\phi_i^{NAT})^2}. \tag{1}$$

0305-4470/00/437699+09$30.00

Here $d_{ij} = |\vec{r}_i - \vec{r}_j|$ are the monomer-to-monomer distances in the given structure, N is the number of beads. The subscripts d, ba and da refer to the distances, the bond angles and the dihedral angles, respectively. The superscript NAT corresponds to the native state. The bond angle, θ_i, is defined as the angle between two successive vectors $\vec{v}_i$ and $\vec{v}_{i+1}$, where $\vec{v}_i = \vec{r}_{i+1} - \vec{r}_i$. The dihedral angle, ϕ_i, is the angle between two vector products $\vec{v}_{i-1} \times \vec{v}_i$ and $\vec{v}_i \times \vec{v}_{i+1}$. The angular distances to the native state have not been studied in previous papers.

We define the structural susceptibilities corresponding to the distances (1) as follows:

$$\begin{aligned} \chi_{\mathrm{d}} &= \langle \delta_{\mathrm{d}}^2 \rangle - \langle \delta_{\mathrm{d}} \rangle^2 \\ \chi_{\mathrm{ba}} &= \langle \delta_{\mathrm{ba}}^2 \rangle - \langle \delta_{\mathrm{ba}} \rangle^2 \\ \chi_{\mathrm{da}} &= \langle \delta_{\mathrm{da}}^2 \rangle - \langle \delta_{\mathrm{da}} \rangle^2 \end{aligned} \tag{2}$$

where the angular brackets indicate a thermodynamic average. As one can see below, these three susceptibilities behave qualitatively in the same way. The sharpness of their peaks may, however, be different (see, for instance, figure 4) and it is useful to calculate all of them.

In the case of an off-lattice model the departure of the sequence geometry from its native conformation is usually described through the structural overlap function [8] as

$$\delta_{\mathrm{o}} = 1 - \frac{2}{N^2 - 3N + 2} \sum_{i \neq j, j \pm 1} \Theta(\delta_{\mathrm{c}} - |d_{ij} - d_{ij}^{\mathrm{NAT}}|) \tag{3}$$

where $\Theta(x)$ is the Heaviside function. The overlap structural susceptibility, χ_{o}, is then defined as the thermal fluctuation of χ_{s}:

$$\chi_{\mathrm{o}} = \langle \delta_{\mathrm{o}}^2 \rangle - \langle \delta_{\mathrm{o}} \rangle^2. \tag{4}$$

The maximum in χ_{o}, when plotted against T, may be interpreted as a signature of the folding temperature T_{f} [8,9]. The advantage of the new definitions of the structural susceptibilities (2) compared with χ_{o} is that the native basin δ_{c} is not involved in their computation.

We have also studied the following derivatives of distances with respect to T:

$$\begin{aligned} D_{\mathrm{d}} &= \frac{\mathrm{d}\langle \delta_{\mathrm{d}} \rangle}{\mathrm{d}T} \qquad & D_{\mathrm{ba}} &= \frac{\mathrm{d}\langle \delta_{\mathrm{ba}} \rangle}{\mathrm{d}T} \\ D_{\mathrm{da}} &= \frac{\mathrm{d}\langle \delta_{\mathrm{da}} \rangle}{\mathrm{d}T} \qquad & D_{\mathrm{o}} &= \frac{\mathrm{d}\langle \delta_{\mathrm{o}} \rangle}{\mathrm{d}T} \\ D_{\mathrm{g}} &= \frac{\mathrm{d}\langle R_{\mathrm{g}} \rangle}{\mathrm{d}T} \end{aligned} \tag{5}$$

where R_{g} is the gyration radius. Naively one can expect that the peaks of the derivatives D, when plotted against T, would coincide with those of the corresponding susceptibilities χ. It is, however, true only when the native conformations are compact.

Using the Monte Carlo and the molecular dynamic simulations we have demonstrated that T_{f} locates at the peaks of χ_{d} (D_{d}), χ_{ba} (D_{ba}) or χ_{da} (D_{da}) provided the native conformations are compact. Thus, the determination of T_{f} does not require the native basin δ_{c}. This is the main advantage of the new quantities given by equations (2) and (5).

The situation becomes more complicated when the native conformations are not compact. In this case the native basin is necessary for the accurate estimate of T_{f}. The information about the foldability may, however, be obtained without δ_{c} monitoring the temperature dependence of D_{d}, D_{ba} and D_{da}. Namely, a good folder would have only one peak in D_{d}, D_{ba} or D_{da}, when plotted against temperature, whereas a bad folder would have two peaks. This may serve as a criterion to distinguish the good folders from the bad ones.

We focus on four sequences whose native conformations are shown in figure 1. The 27-monomer lattice chain, L_{27}, is a Go sequence [10]. Its Hamiltonian is as follows:

$$H = \sum_{i<j} \alpha_{ij} \Delta_{ij} \tag{6}$$

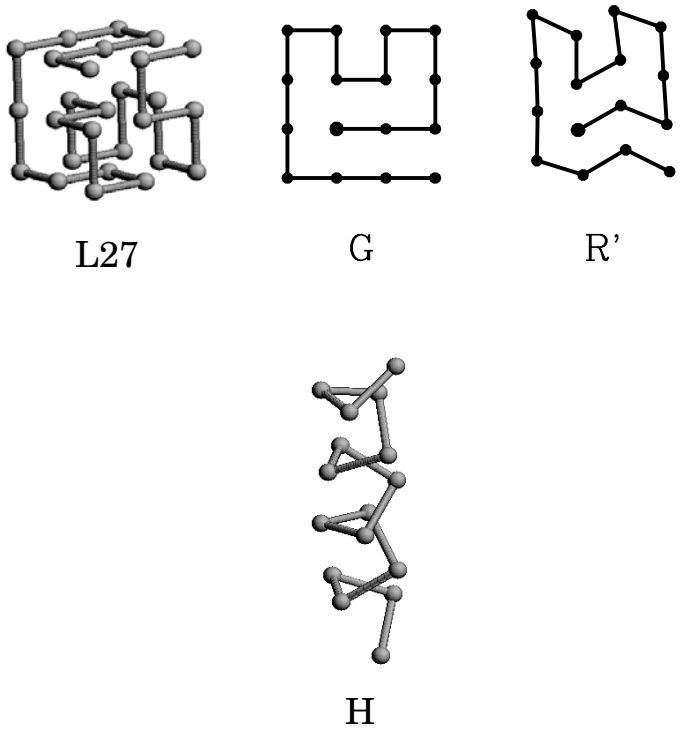

Figure 1. Native conformations of four sequences studied in this paper.

where $\Delta_{ij} = 1$ if monomers i and j are in contact and 0 otherwise. The quantity $\alpha_{ij} = -1$ if monomers i and j are in contact in the native conformation and 0 otherwise. We use L_{27} to check the behaviour of the new quantities δ_{ba} (D_{ba}) and δ_{ha} (D_{ha}) for the lattice models.

The sequences denoted by G and R' are two-dimensional versions of the model introduced by Iori *et al* [11]. The Hamiltonian is given by

$$H = \sum_{i \neq j} \left\{ k(d_{i,j} - d_0')^2 \delta_{i,j+1} + 4\epsilon \left[\frac{C}{d_{i,j}^{12}} - \frac{A_{ij}}{d_{i,j}^{6}} \right] \right\} \tag{7}$$

where i and j range from 1 to $N = 16$. d_{ij} is measured in units of σ, the typical value of which is $\sigma = 5$ Å. We take d_0' to be equal to $2^{1/6}\sigma$ and 1.16σ for G and R', respectively [12]. The harmonic term in the Hamiltonian, with the spring constant k, couples the beads that are adjacent along the chain. The remaining terms represent the Lennard-Jones potential. Random values of A_{ij} describe the quenched disorder. In equation (7) ϵ is the typical Lennard-Jones energy parameter. We adopt the units in which $C = 1$ and consider k to be equal to 25ϵ. Smaller values of k may violate the self-avoidance of the chain. The coupling constants A_{ij} for system R' are listed in [12]. These are shifted Gaussian-distributed numbers with the strongest attracting couplings assigned to the native contacts. For system G, A_{ij} is taken to be 1 or 0 for the native and non-native contacts respectively. System R' has been shown to be structurally overconstrained and hard to fold.

The helical system H has a native state that mimics typical α-helix secondary structures. In this case the distances between beads are assumed to have the length $d_0 = 3.8$ Å. As one

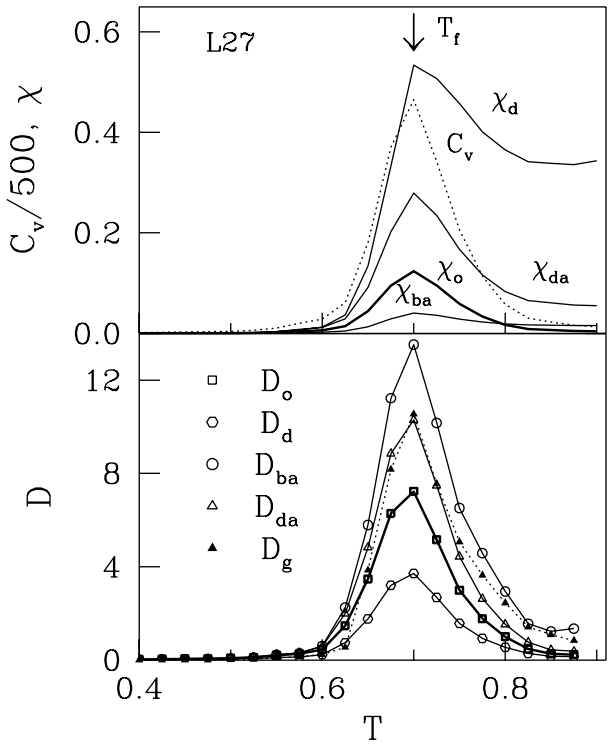

Figure 2. The temperature dependence of C_v, χ and D for sequence L_{27}. The arrow corresponds to T_f. T_f is defined as a temperature at which the probability of being in the native state is $1/2$. C_v and D_g are denoted by dotted curves whereas χ_o and D_o are denoted by thick curves. The results are averaged over 50 starting conformations. The error bars are smaller than the symbol sizes.

proceeds along the helix axis from one bead to another, the bead's azimuthal angle is rotated by 100° and the azimuthal length is displaced by 1.5 Å. The Hamiltonian used to describe the helix is a Go-like modification of equation (7) and it reads [13]

$$H = V^{\mathrm{BB}} + V^{\mathrm{NAT}} + V^{\mathrm{NON}}. \tag{8}$$

The first term is a backbone potential which includes the harmonic and anharmonic interactions

$$V^{\mathrm{BB}} = \sum_{i=1}^{N-1} [k_1(d_{i,i+1} - d_0)^2 + k_2(d_{i,i+1} - d_0)^4]. \tag{9}$$

We take $d_0 = 3.8$ Å, $k_1 = \epsilon$ and $k_2 = 100\epsilon$. The interaction between residues which form native contacts in the target conformation is chosen to be of the Lennard-Jones form

$$V^{\mathrm{NAT}} = \sum_{i<j}^{\mathrm{NAT}} 4\epsilon \left[\left(\frac{\sigma_{ij}}{d_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}} \right)^{6} \right]. \tag{10}$$

We choose σ_{ij} so that each contact in the native structure is stabilized at the minimum of the potential, i.e. $\sigma_{ij} = 2^{-1/6} d_{ij}^N$, where d_{ij}^N is the length of the corresponding native contact. Residues that do not form the native contacts interact via a repulsive soft core potential V^{NON}, where

$$V^{\mathrm{NON}} = \sum_{i<j}^{\mathrm{NON}} V_{ij}^{\mathrm{NON}} \tag{11}$$

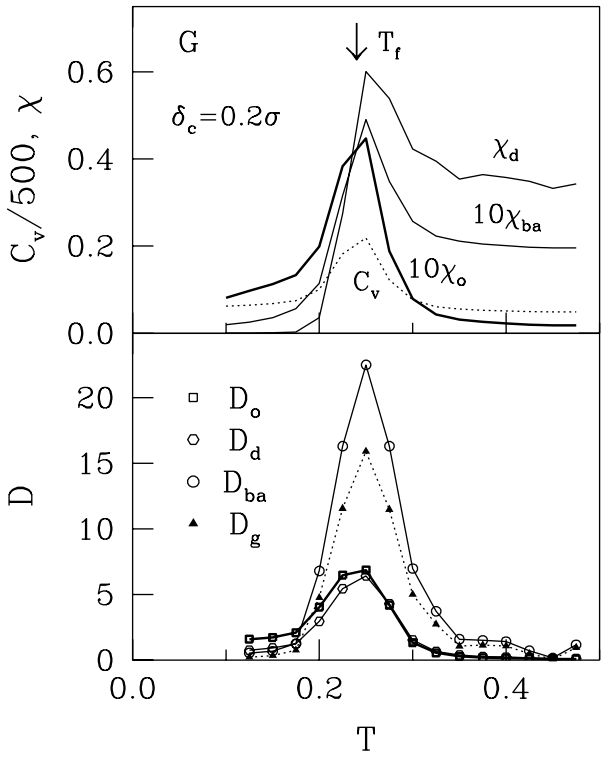

Figure 3. The temperature dependence of C_v, χ and D for sequence G. The native basin defined by the shape distortion approach [7] is equal to $\delta_c = 0.2\sigma$. The results are averaged over 100 molecular dynamic trajectories.

$$V_{ij}^{\mathrm{NON}} = \begin{cases} 4\epsilon\left[\left(\dfrac{\sigma_0}{d_{ij}}\right)^{12} - \left(\dfrac{\sigma_0}{d_{ij}}\right)^{6}\right] + \epsilon & d_{ij} < d_{\mathrm{cut}} \\ 0 & d_{ij} > d_{\mathrm{cut}}. \end{cases} \tag{12}$$

Here $\sigma_0 = 2^{-1/6} d_{\mathrm{cut}}$, $d_{\mathrm{cut}} = 5.5$ Å. The difference between a Go and Go-like sequences is in the choice of the non-native contact interaction energy which is taken to be zero for the Go sequence and non-zero for the latter one.

The thermodynamics of L_{27} are studied by a Monte Carlo procedure that satisfies the detailed balance condition [14, 15]. The dynamics allow for single and two-monomer (crankshaft) moves. For each conformation of the chain, one has A possible moves and the maximum value of A, $A_{\max}$, is equal to $A_{\max} = N + 2$. In our 27-monomer case $A_{\max} = 29$. For a conformation with A possible moves, the probability to attempt any move is taken to be $A/A_{\max}$ and the probability not to do any attempt is $1 - A/A_{\max}$ [15]. In addition, the probability to do a single move is reduced by a factor of 0.2 and to do the double move, by 0.8 [15]. The attempts are rejected or accepted as in the standard Metropolis method. The equilibration is checked by monitoring the stability of data against at least three-times longer runs. We have used typically 10^6 Monte Carlo steps (the first 5×10^5 steps are not taken into account when averaging).

Figure 2 shows the temperature dependence of C_v, χ and D for L_{27}, where χ (D) is a common notation for χ_{d} (D_{d}), χ_{ba} (D_{ba}), χ_{da} (D_{da}) and χ_{o} (D_{o}). In this on-lattice case the

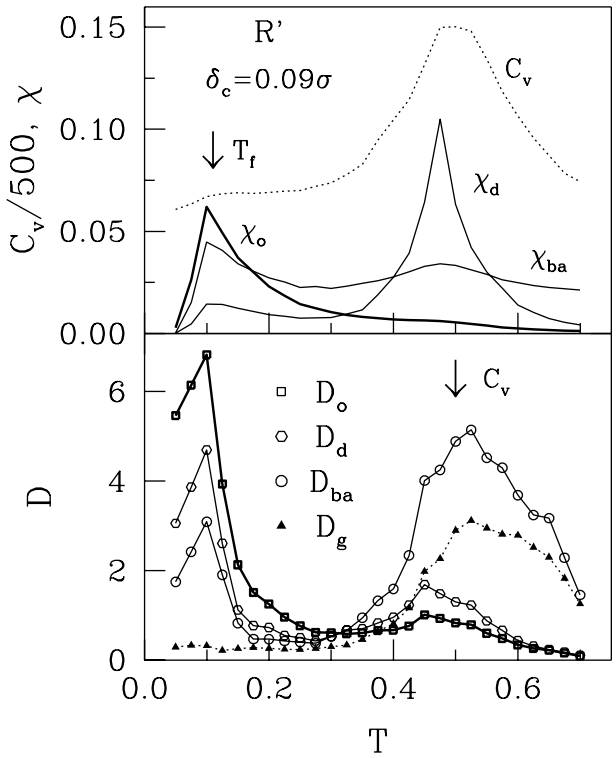

Figure 4. The temperature dependence of C_v, χ and D for sequence R'. The native basin is equal to $\delta_c = 0.09\sigma$. The results are averaged over 170 molecular dynamic trajectories.

overlap structural susceptibility χ_o is also given by equation (4) but δ_o reads as follows [8]:

$$\delta_o = 1 - \frac{2}{N^2 - 3N + 2} \sum_{i \neq j, j \pm 1} \delta(d_{ij} - d_{ij}^{NAT}). \tag{13}$$

For sequence L_{27} the peaks of all quantities are located at $T = T_f$. The fact that the maxima of χ_o and D_g are located at the same position has also been observed for some on-lattice sequences [17]. Our new result is that, similar to χ_o, the susceptibilities based on the fluctuations of the distances to the native conformation and D also give a correct position for T_f. According to the thermodynamic criterion [8, 16], L_{27} should be a good folder because T_f coincides with the collapse temperature T_θ (T_θ is defined as a temperature where C_v develops a peak).

In order to study the time evolution of the off-lattice sequences G, R' and H, we use the equations of motion for the Langevin uncorrelated noise terms

$$m\ddot{\boldsymbol{r}} = -\Gamma\dot{\boldsymbol{r}} + F_c + \eta \tag{14}$$

where $F_c = -\nabla_r E_p$ and

$$\langle \eta(0)\eta(t) \rangle = 2\Gamma k_B T \delta(t). \tag{15}$$

Here k_B and Γ are the Boltzmann constant and the kinetic coefficient, respectively. Equation (14) is integrated by the fifth-order predictor–corrector method [18]. The integration step is chosen to be 0.005τ, where $\tau = m\sigma^2/\epsilon$ is the characteristic time unit and m is the mass

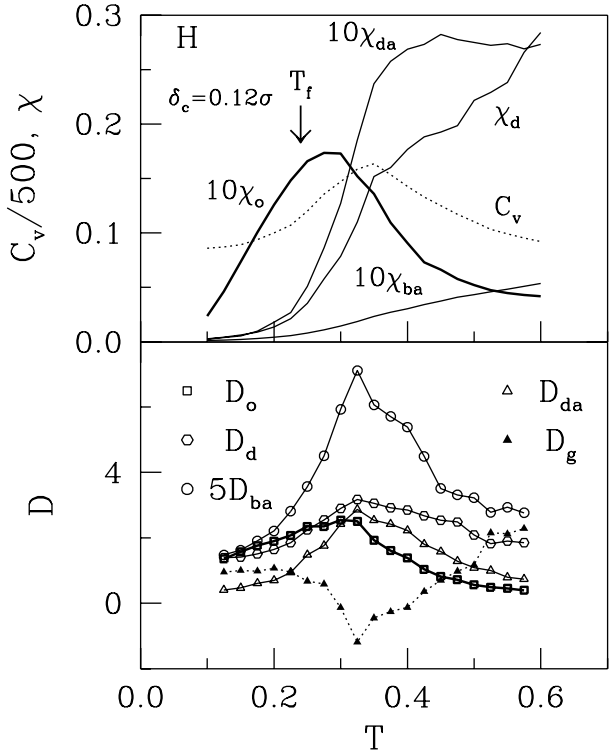

Figure 5. The temperature dependence of C_v, χ and D for sequence H. The native basin is equal to $\delta_c = 0.12\sigma$. The results are averaged over 200 molecular dynamic trajectories.

of a bead. We take $\Gamma = 2$. In the following, the temperature will be measured in the reduced units of ϵ/k_B.

The folding properties of G, R' and H were characterized in detail previously [12,13]. One of them, R', is a bad folder and two others are good folders. We calculate the thermodynamic quantities of G, R' and H by averaging over many molecular dynamic trajectorjes using the native state as the starting configuration to make sure that the evolution takes place in the right part of the phase space [12]. For all of these sequences, the time used for averaging in each trajectory is 4000τ for each temperature. The first 2000τ are discarded.

Figures 3 and 4 show the results for G and R'. Since these sequences are two-dimensional, χ_{ha} and D_{ha} corresponding to the dihedral angles do not appear. The basin was obtained by the shape distortion approach [7] and is equal to $\delta_c = 0.2\sigma$ and $\delta_c = 0.09\sigma$ for G and R', respectively. Within the error bars of 0.02 all of the maxima of χ, D and C_v are located at the folding temperature T_f ($T_f = 0.24 \pm 0.02$ and 0.10 ± 0.02 for G and R', respectively). Therefore, the determination of T_f does not require the native basin because it is enough to find the peak of χ (or of D) in which δ_c is not involved.

For R', χ_o has only one peak at T_f whereas χ_d and χ_{ba} have an additional maximum at $T = T_\theta$. Therefore, the advantage of χ_d and χ_{ba} compared with χ_o is that they allow us to find not only T_f but also T_θ. Since the maximum of D_g is broad around the folding temperature, it is better to locate T_f as a second peak of χ_d (D_d) or χ_{ba} (D_{ba}). This demonstrates another advantage of the new quantities compared with the standard quantity D_g.

It should be noted that the behaviour of $\chi_{\rm d}$ and $\chi_{\rm ba}$ is qualitatively the same but there is a quantitative difference in the sharpness of their peaks. It is clear from figure 4 that at $T = T_{\rm f}$ the maximum of $\chi_{\rm ba}$ is more pronounced compared with that of $\chi_{\rm d}$. An opposite situation takes place at $T = T_\theta$. So, the study of all of susceptibilities would help us to isolate peaks better.

The fact that $\chi_{\rm o}$ has only one peak, but the others have two may be explained in the following way. Since $\chi_{\rm o}$ is a fluctuation of the overlap with the native state it reflects the behaviour of the system in the vicinity of the native basin and it should have, therefore, only one peak at $T_{\rm f}$. The remaining susceptibilities related to the chain compactness would have two maxima at $T_{\rm f}$ and T_θ where the topology changes abruptly.

The temperature dependence of χ, D and C_v for the three-dimensional sequence H is shown in figure 5. In this case we have the basin $\delta_{\rm c} = 0.12\sigma$ and $T_{\rm f} = 0.24 \pm 0.02$ [13, 19]. Since $\chi_{\rm d}$, $\chi_{\rm ba}$ and $\chi_{\rm da}$ do not display any peak in the relevant temperature interval, they cannot be used to determine $T_{\rm f}$. It is also true for D (their extremal points are located at $T = 0.325 \pm 0.025$ which is far from $T_{\rm f}$). The overlap susceptibility $\chi_{\rm o}$ has its maximum at $T_{\chi_{\rm o}} = 0.275 \pm 0.025$. Within the error bars, $T_{\chi_{\rm o}}$ may be identified as $T_{\rm f}$ but such an estimate is less accurate compared with the case of G and R'. Furthermore its computation involves the native basin $\delta_{\rm c}$.

From the results presented in figures 2–5 we propose the following criterion for foldability: a good folder would have only one peak in the derivatives of distances to the native state with respect to temperature, whereas a bad folder has two. Our criterion is compatible with the fact that for the good folders the folding takes place just after the collapse transition. A three-state scenario of folding is, however, more suitable for the bad folders [9]. Thus, one can still gain information about the foldability for H without the native basin $\delta_{\rm c}$.

The question we ask now is why H is so different from the other sequences. The main difference is that its native conformation is not compact. It results in the non-trivial dependence of $R_{\rm g}$ on T: $D_{\rm g}$ does not develop a maximum but rather a minimum around the collapse transition. This leads to the anormal behaviour of $\chi_{\rm d}$, $\chi_{\rm ba}$ and $\chi_{\rm ha}$.

In conclusion, we have introduced several new structural susceptibilities as fluctuations of distances to the native conformation. If the native conformation of proteins is compact $T_{\rm f}$ may be obtained from the peak of χ and the native basin is not required. For sequences with non-compact native conformation $\delta_{\rm c}$ is not needed to establish the foldability but the accurate estimate of $T_{\rm f}$ should involve it. The number of peaks in the derivatives of distances to the native state with respect to temperature, when plotted against T, may serve as a tool to distinguish between good and bad folders. The question of why the susceptibilities χ and the derivatives D behave in the same way if the native conformations are compact remains to be elucidated. Nevertheless, in agreement with other studies (see, for instance, [20] and references therein), our results indicate that the topology of the native state plays a crucial role in the folding process.

Acknowledgments

The author is grateful to M Cieplak for useful discussions. This work was supported by Komitet Badan Naukowych (Poland; grant number 2P03B-146 18).

References

[1] Dill K A, Bromberg S, Yue S, Fiebig K, Yee K M, Thomas D P and Chan H S 1995 *Protein Sci.* **4** 561
[2] Pande V S, Grosberg A Yu and Tanaka T 2000 *Rev. Mod. Phys.* **72** 259

[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
[5] Veitshans T, Klimov D K and Thirumalai D 1997 *Folding and Design* **2** 1
[6] Clementi C, Nymeyer H and Onuchic J N 2000 *J. Mol. Biol.* **298** 937
(Clementi C, Nymeyer H and Onuchic J N 2000 *Preprint* cond-mat/0003460)
Onuchic J N, Nymeyer H, Garcia H, Chahine A E and Socci N D 2000 *Adv. Protein Chem.* **53** 87
[7] Li M S and Cieplak M 1999 *J. Phys. A: Math. Gen.* **32** 5577
[8] Camacho C J and Thirumalai D 1993 *Proc. Natl Acad. Sci. USA* **90** 6369
[9] Thirumalai D 1995 *J. Physique (Paris)* I **5** 1457
[10] Go N and Abe H 1981 *Biopolymers* **20** 991
[11] Iori G, Marinari E and Parisi G 1991 *J. Phys. A: Math. Gen.* **24** 5349
[12] Li M S and Cieplak M 1999 *Phys. Rev.* E **59** 970
[13] Hoang T X and Cieplak M 2000 *J. Chem. Phys.* **112** 6851
[14] Cieplak M, Henkel M, Karbowski J and Banavar J R 1998 *Phys. Rev. Lett.* **80** 3654
Cieplak M, Henkel M and Banavar J R 1999 *Cond. Mat. Phys. (Ukraine)* **2** 369
[15] Chan H S and Dill K A 1994 *J. Chem. Phys.* **99** 2116
Chan H S and Dill K A 1994 *J. Chem. Phys.* **100** 9238
[16] Klimov D K and Thirumalai D 1996 *Phys. Rev. Lett.* **76** 4070
[17] Klimov D K and Thirumalai D 1998 *J. Chem. Phys.* **109** 4119
[18] Allen M P and Tildesley D J 1987 *Computer Simulation of Liquids* (New York: Oxford University Press)
[19] Li M S, Cieplak M and Sushko N 2000 *Phys. Rev.* E **62** 4025
[20] Baker D 2000 *Nature* **405** 39