Phase diagram of a model protein derived by exhaustive enumeration of the conformations

A. Dinner, A. Šali, M. Karplus, and E. Shakhnovich
Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138

(Received 25 October 1993; accepted 23 March 1994)

An understanding of the various states available to a polypeptide chain is important for a description of the protein folding process. We use a 16-monomer chain on a two-dimensional square lattice to model a protein. This makes it possible to enumerate all self-avoiding conformations from which any equilibrium thermodynamic quantity can be calculated. By varying the external conditions of temperature and average attraction, we construct a phase diagram for the model protein. It is found to have an extended coil state, a homopolymer-like disorganized globule state, and an organized frozen globule state that corresponds to the lowest energy (native) conformation. The exact model results agree well with analytical heteropolymer theory.

A full understanding of protein folding requires a characterization of the phase space accessible to a polypeptide chain.[1] Although much is known about the neighborhood of the native state from experimental and theoretical studies,[2] less information exists concerning the nonnative portion of the conformation space. Both theory[3] and experiment[4] suggest that the space is complex and that there are several states. Two regimes which clearly exist are the random coil and native states; the former consists of a very large number of rapidly interconverting configurations, while the latter fluctuates only in the neighborhood of a single unique fold. In addition, there appears to exist a homopolymer-like globule,[3] where the polypeptide chain is relatively compact, but fluctuates between a large number of collapsed configurations.[5] Many proteins also show evidence for a state, often referred to as a "molten globule,"[6] where the backbone has attributes of the native structure, but the sidechains are still free to rotate. It has been suggested[7–9] that the transition from the molten globule to the state with tightly packed sidechains is the first-order ("all-or-none") transition observed in the protein folding process. In this view, the denatured protein encompasses a state with a relatively well-defined backbone conformation (molten globule), a collapsed disorganized homopolymer-like state (globule), and the extended random coil state. An understanding of the character and stability of these non-native states, as well as the transitions between them, is important for a complete description of the process of protein folding and unfolding.[1,10]

A theory developed for a heteropolymer (bead) model of a protein[3] (see a review in Ref. 1) has shown that such a chain with sufficiently high heterogeneity can have a complex phase diagram with a coil, random globule (this is the same as a homopolymer globule when the multitude of different structures are equally probable) and a "frozen" state where only a few (for selected sequences, only one[11]) conformations are stable. It is the heterogeneity of an amino acid sequence, which allows for the existence of a frozen state, that differentiates a polypeptide chain from a simple homopolymer. The results obtained from this heteropolymer model are in accord with spin glass theory[12] and with phenomenological models of such chains.[13] The frozen state is characterized by a temperature T_c below which there is a transition and the chain acquires a unique structure. It is likely that the resulting frozen state represents an organized molten globule rather than the true native state because of the lack of sidechains. An important prediction of the theory is that the "folding temperature" T_c is different from the "compactization temperature" T_Θ, at which the chain makes the transition from the extended random coil state to the homopolymer-like globule state.

To examine the phase space of proteins in more detail, lattice models can be used. The protein is represented as a string of beads whose positions are restricted to a selected lattice. Some aspects of the thermodynamics of the frozen state predicted by heteropolymer theory have been tested by the full enumeration of all compact self-avoiding conformations for a 27-mer on a 3 by 3 by 3 fragment of a cubic lattice.[11] Although useful information about the folding kinetics has been obtained,[14,15] the total number of configurations, noncompact and compact, is so large that all the nonnative configurations cannot be enumerated. To investigate the full configuration space, we introduce an even simpler model. The system used is a 16-mer on a two-dimensional (2D) square lattice.[16] The total number of conformations of such a chain is 802 075. This includes all noncompact and compact self-avoiding conformations that are not equivalent by symmetry.[16] Although 2D lattices may have limits in their applicability to three-dimensional behavior,[1] they are sufficient for the present problem because it has been shown theoretically[17] that 2D chains exhibit compactization, as well as freezing, transitions. Furthermore, when exhaustive enumeration was previously applied to the study of the native-denatured transition,[18–20] such a freezing transition was observed. In spite of its simplicity, the 2D model has the advantage that a wide range of parameters can be scanned, allowing construction of a detailed phase diagram. This is the subject of the present study.

The energy function was chosen to be of the same form as that used for the 3D lattice studies;[11,14,15] i.e., the energy of a conformation m is

 0021-9606/94/101(2)/1444/8/$6.00

TABLE I. The $\{B_{ij}\}$ values of the two Gaussian sequences used in the computations. Since the matrices are symmetric, only half of each is shown; the first Gaussian sequence is above the diagonal, and the second is below. Due to the constraints of the square lattice, only odd contact orders $(j-i)$ are possible. There is no interaction between residues adjacent in sequence or between a residue and itself.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1				-0.4923		1.0474		-0.5496		1.1895		-2.1095		-0.7512		-0.0469
2					1.3830		-1.2977		-0.1408		0.0940		0.9985		-1.8114	
3						-0.1705		1.1223		-0.3847		-0.2865		0.3293		-1.0435
4	1.9063						-1.0519		0.5629		0.3252		-0.5998		1.7539	
5		-0.9914						-0.2118		1.1253		1.8313		0.0326		0.3624
6	-0.8975		-0.9801						-1.6053		0.4711		0.0228		0.3899	
7		-0.6782		-0.3037						0.2471		-1.3267		-1.4820		-0.3501
8	0.0770		-0.0625		-0.0651						-0.1171		-0.2125		-0.5015	
9		1.1379		0.2299		-1.0613						0.0594		0.6888		-0.2472
10	1.6883		-0.6560		1.5177		-0.4063						0.5166		1.1651	
11		0.0821		-0.9717		-0.8259		1.3644						-0.2958		-1.7203
12	-0.6494		-0.3041		1.3921		0.8901		0.6439						-0.0592	
13		0.5539		-0.8744		-1.7790		1.3071		-1.5915						0.3877
14	0.8742		-0.7576		-0.9366		0.3860		0.8944		0.2272					
15		-0.6158		-0.2275		0.7045		0.3967		-0.7950		0.6108				
16	0.2781		-0.2991		-1.5168		-0.8012		0.1926		-1.1363		-0.3228			

$$E_m = B_0 \sum_{i>j}^{N} \Delta(r_i^m - r_j^m) + \sum_{i>j}^{N} B_{ij} \Delta(r_i^m - r_j^m), \tag{1}$$

where N is the total number of monomeric units, B_0 is the average interaction between monomers, and B_{ij} is the inhomogeneous part of the interaction energy that depends on the type of monomer i and j. The Kronecker delta function $[\Delta(r_i^m - r_j^m) = 1$ if monomers i and j are lattice neighbors and 0 otherwise] reflects the short-range nature of the model potential; only nearest-neighbor (in space) interactions are included. A given sequence is characterized by the set $\{B_{ij}\}$ associated with it. Two models are employed for the B_{ij} values. In the first, B_{ij}'s are independent random values with a Gaussian distribution[11,14,15]

$$P(B_{ij}) = \frac{1}{\sqrt{2\pi B^2}} \exp\left(-\frac{B_{ij}^2}{2B^2} \right), \tag{2}$$

where B is the standard deviation which determines the heterogeneity of the chain, and the mean of B_{ij} is taken to be 0 since the average interaction is included in B_0. This model potential is exactly that used in the theoretical analysis mentioned above.[3] The second model is a "two-letter" random sequence of A- and B-type monomers. For computation, we used two Gaussian-distributed $\{B_{ij}\}$ with $B = 1.0$ (Table I). The sample AB sequence used was chosen from 100 randomly generated sequences of that type; its sequence and interaction matrix are described in the legend to Fig. 1. A similar two-letter code has been used previously in both lattice[16] and analytical models[21] with A and B representing hydrophobic and hydrophilic amino acids, for example.

Any equilibrium thermodynamic quantity for a given sequence can be determined for this model by performing the appropriate average over all conformations. Here we consider the quantities $\langle N_c \rangle$ the thermally averaged number of contacts, which is related to the density, and $\langle Q \rangle$ the thermally averaged overlap of all pairs of conformations. The brackets denote Boltzmann averaging at a given temperature. Thus, the expression for $\langle N_c \rangle$ is

$$\langle N_c \rangle = \frac{\sum_m^M C_m e^{-E_m/kT}}{\sum_m^M e^{-E_m/kT}}, \tag{3}$$

where summation in the numerator and denominator is taken over all M conformations and C_m is the number of contacts in conformation m with energy E_m. Similarly,

$$\langle Q \rangle = \frac{\sum_{n,m}^M Q_{mn} \exp[-(E_m + E_n)/kT]}{(\sum_m^M e^{-E_m/kT})^2}, \tag{4}$$

where

$$Q_{mn} = \sum_{i>j}^{N} \Delta(r_i^m - r_j^m) \Delta(r_i^n - r_j^n) / C_{\max}^{mn}$$

is the normalized "overlap" function which shows how many contacts in a structure m coincide with the contacts in another structure n. $C_{\max}^{mn}$ is the number of contacts in the more compact of the two structures; as a result, $0 \leqslant Q_{mn} \leqslant 1$ and $Q_{mn} = 1$ only for structures with the same contacts. This can be true for the identical structure (Q_{mn}) or for two different structures with less than the maximum number of contacts. In particular, two structures, both with zero contacts, have $Q_{mn} = 1$. The quantity $\langle Q \rangle$ acts as an order parameter which is related to the number of thermodynamically stable configurations. When only the global energy minimum is thermodynamically stable, its probability at temperature T is approximately one, while all other structures have negligible probabilities; it follows from Eq. (4) that $\langle Q \rangle \approx 1$. On the other hand, when the chain is disordered (i.e., many dissimilar configurations have roughly equal probabilities at temperature T), the average is not dominated by one structure and $\langle Q \rangle \ll 1$ ($\langle Q \rangle \approx 0.2$ since random overlaps prevent $\langle Q \rangle$ from reaching zero). The quantity $\langle N_c \rangle$ was calculated with all conformations, while $\langle Q \rangle$ was calculated with only the 1000 lowest energy structures because it is proportional to a product of Boltzmann probabilities, so that only those contribute significantly. Calculations with 5000 structures did not change the numerical results.

We consider the behavior of the system as a function of the temperature T and the homopolymeric interaction energy B_0, which acts to control the overall compactness of the chain. Although the analytical study varied B rather than T, the effective width of the Gaussian interaction distribution can be scaled with either [Eqs. (2)–(4)]. However, the use of

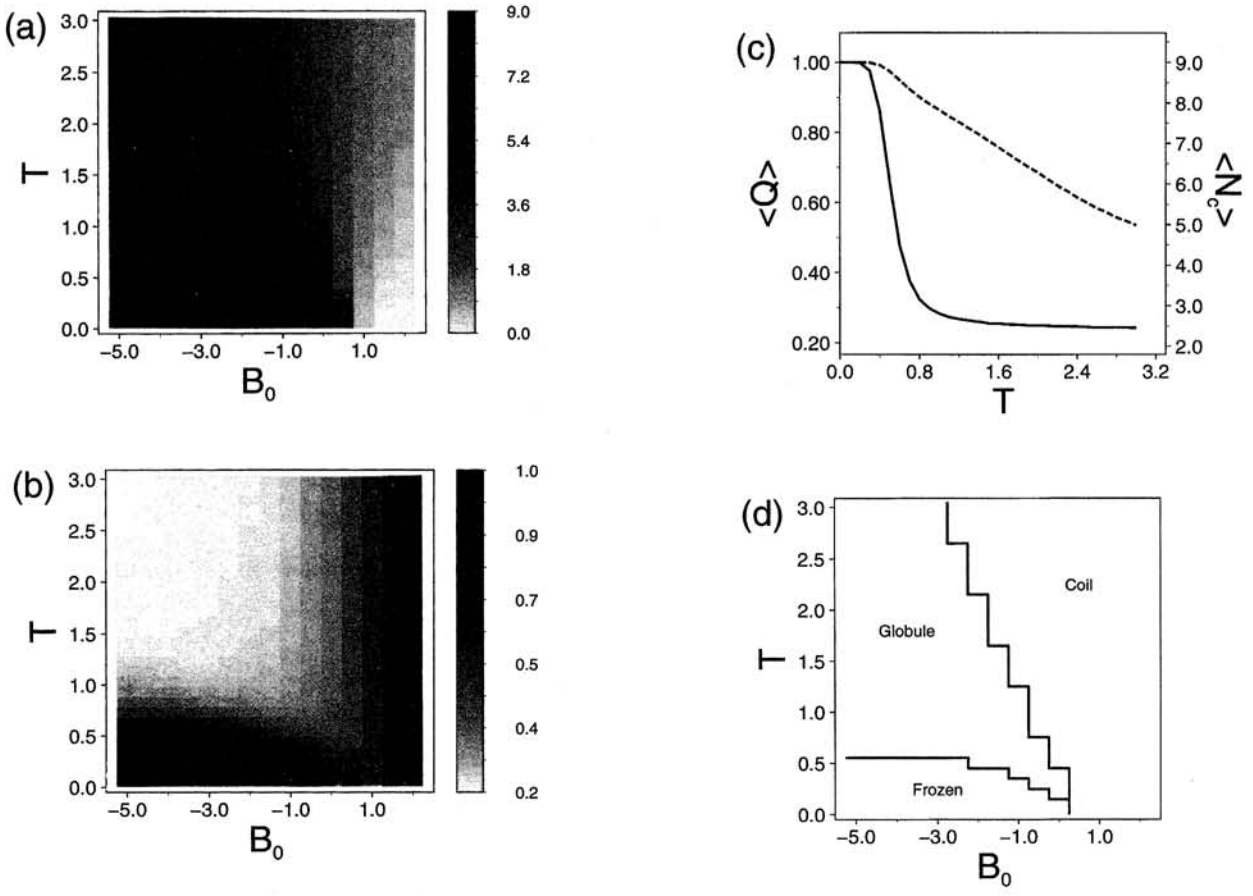

FIG. 1. Phase diagram for a two-letter sequence (a) $\langle N_c \rangle$ and (b) $\langle Q \rangle$. $\langle Q \rangle = 1.0$ for all T with $B_0 \geq 1.0$ because the 1000 lowest energy structures all have zero contacts, so that $Q_{mm} = 1$. $\langle N_c \rangle$ increases as T increases at $B_0 \geq 1.0$ because it is calculated with all structures, and more compact structures generally have higher energy for positive B_0. (c) Superposition of $\langle N_c \rangle$ (dashed) and $\langle Q \rangle$ (solid) at $B_0 = -2.0$. (d) Construction of a phase diagram with the following cutoffs: frozen is $\langle Q \rangle > 0.7$ and $\langle N_c \rangle > 6.0$; globule is $\langle Q \rangle \leq 0.7$ and $\langle N_c \rangle > 6.0$; coil is $\langle Q \rangle \leq 0.7$ and $\langle N_c \rangle \leq 6.0$. Note the triple point at $(B_0, T) = (0.25, 0.15)$. The interaction matrix for this sequence is $B_{AA} = -1$, $B_{AB} = 0$, and $B_{BB} = 0$, and the sequence is *BBABBAABBAABABAA*.

T allows direct comparison with proteins, which have fixed sequences. The variation in $\langle Q \rangle$ and $\langle N_c \rangle$ with B_0 and T is shown in Figs. 1(a), 1(b), 2(a), 2(b), 3(a), and 3(b), respectively, for three different sequences. As shown in Figs. 1(c), 2(c), and 3(c), $\langle N_c \rangle$ has a higher transition temperature than $\langle Q \rangle$ in all cases, so that there exists a regime in which the molecule has lost its unique backbone structure, but is still compact; i.e., the freezing transition occurs at a lower temperature T_c than the disordered globule transition T_Θ. This is because loss of unique structure, characterized by a decrease in $\langle Q \rangle$, involves a much sharper transition than the loss of compactness, characterized by a decrease in $\langle N_c \rangle$. Due to the difference in T_c and T_Θ, there exist three distinct states of the molecule which are stable under different external conditions. They are an extended coil state (C), a compact disorganized globule state (G), and a frozen globule state with a unique structure (F). Comparing (a) and (b) of each figure, we see that there is a region (C), where both $\langle N_c \rangle$ and $\langle Q \rangle$ are large ($T < T_\Theta, T_c$); a region (G), where $\langle N_c \rangle$ is large, while $\langle Q \rangle$ is small ($T_\Theta < T < T_c$); and a region (F), where both $\langle N_c \rangle$ and $\langle Q \rangle$ are small ($T > T_\Theta, T_c$). To make this precise, we show the boundaries of the three regions for reasonable, though arbitrary, criteria on the limits of $\langle N_c \rangle$ and $\langle Q \rangle$ [Figs. 1(d), 2(d) and 3(d)]. The existence of three phases, the general features of the phase boundaries, and the presence of a triple point agree well with the phase diagram predicted by the analytical model.[3] The two primary differences are the existence of the coil state for $B_0 < 0$ and the leveling of the temperature of the globule-frozen boundary at very low B_0. These differences are a result of the shortness of the chains, which allows individual sequence features to play a more important role than expected for infinite chains.

The ability of the chain to collapse to a globule and have a unique native state even when $B_0 > 0$ is due to the heterogeneity of the chain. We show the ground states at different B_0 for the three sequences in Figs. 4, 5, and 6, respectively. For the Gaussian sequences, as B_0 increases, some contacts become positive (repulsive) before others, resulting in a gradual relative increase in the stability of more open structures. At sufficiently high B_0, all contacts are repulsive and the ground state consists of the 116 579 structures with zero contacts. This decrease in the density of the native state also occurs with the two-letter sequence; *BB* (hydrophilic) and *AB* contacts both become repulsive when $B_0 > 0$, while *AA* (hydrophobic) contacts are repulsive only for $B_0 > 1$. However, unlike the Gaussian sequences, all *AA* interactions in

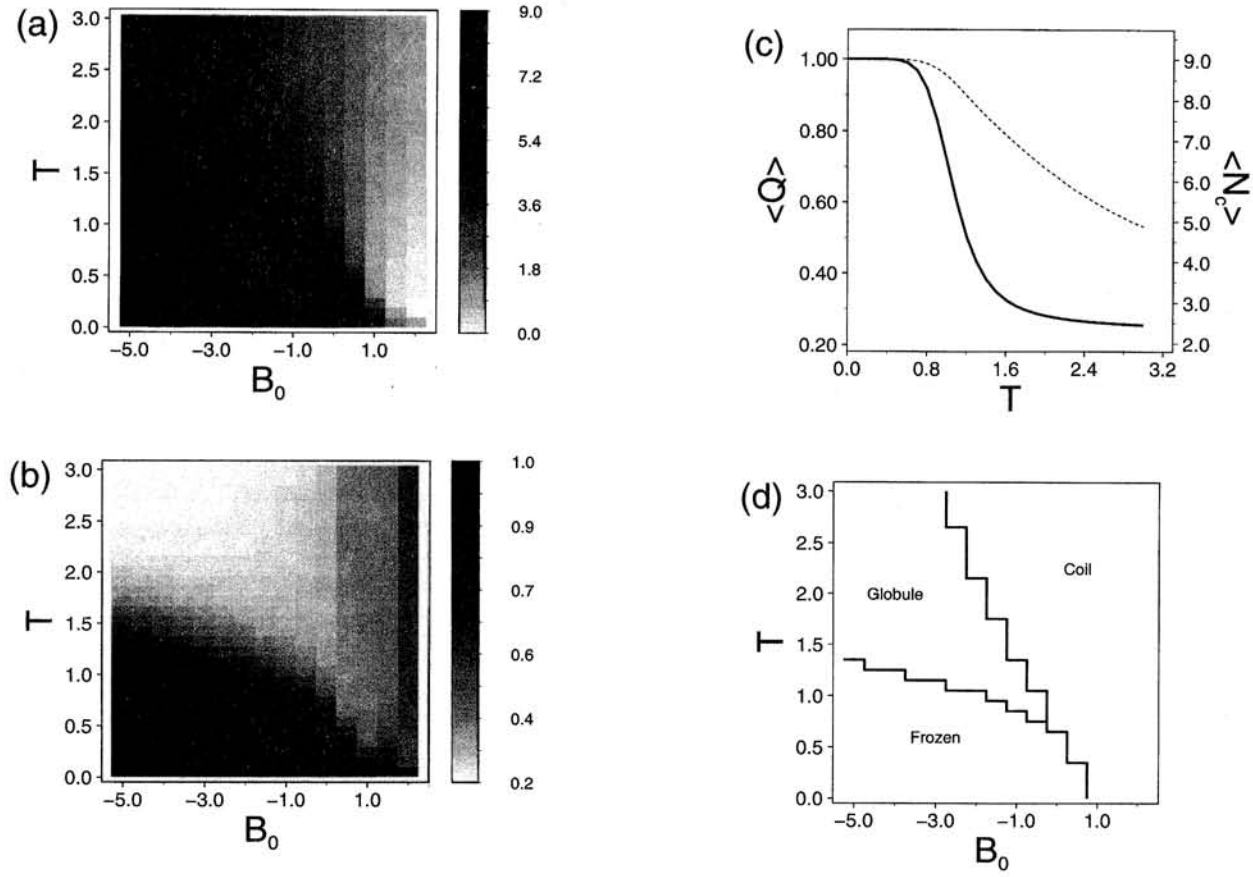

FIG. 2. The same as Fig. 1 for a Gaussian sequence. The triple point for this sequence occurs at $(B_0,T)=(-0.25,0.75)$.

the two letter sequences have the same energy regardless of the positions of the monomers in the primary sequence. As a result, the degeneracy of the two-letter sequence can come from rearrangements of contacts in addition to rearrangements of segments without contacts. This results in a higher degree of degeneracy in the two-letter sequences for the ground states that are not maximally compact.

The shape of the frozen section of the phase diagram is dependent on the size of the plateau in $\langle Q(T) \rangle$ for a given B_0 [Figs. 1(a), 2(a), and 3(a)]. The transition temperature of $\langle Q(T) \rangle$ is dictated by the stability of the ground state relative to the rest of the ensemble. To show this, we present the difference in energy of the ground and first excited states (ΔE_{10}) as a function of B_0 for these sequences; these values are in the captions of Figs. 4, 5, and 6. The quantity ΔE_{10} is closely related to the Boltzmann probability of the native state $[p_0(T)]$, but is temperature independent. As one increases B_0, ΔE_{10} may change for three reasons—the ground state may change, the first excited state may change, or the two structures may have a different number of contacts, causing only the energy gap to change (Figs. 4, 5, and 6). We find that ΔE_{10} decreases as B_0 increases for all three sequences, resulting in a decrease in the temperature range for which formation of the ground state is thermodynamically favorable at higher B_0. This lowering of the transition temperature for $\langle Q(B_0) \rangle$ accounts for the shape of the frozen section of the phase diagram [Figs. 1(d), 2(d), and 3(d)]. Because $\Delta E_{10}(B_0)$ does not change in the same manner for all sequences, different sequences can exhibit qualitatively different configuration space behavior. The first Gaussian sequence has a large separation between the global energy minimum and the first excited state when B_0 is negative (encouraging maximal compactness), while the second sequence has a very small energy difference between these two states. This results in a different dependence of $\langle Q(T) \rangle$ on the other system parameters [Figs. 2(b), 2(c), 3(b), and 3(c)]. The first sequence exhibits a plateau in $\langle Q(T) \rangle$ at low T due to the large relative stability of the global energy minimum with respect to the rest of the ensemble, while the second does not. From detailed folding studies of a 27 bead model in 3D,[15] these variations among phase diagrams are likely to reflect differences in the ability of a sequence to fold. Temperatures which would thermodynamically strongly favor formation of the native state would be too low to allow escape from local minima for the second Gaussian sequence, but the first such sequence is expected to fold. Calculations of ΔE_{10} as well as $p_0(T)$ and the entropy of the system for 100 two-letter sequences suggest that a similar variation of the phase diagram is expected among two-letter sequences. Thus, while there are differences between individual sequences, the Gaussian and two-letter models have the same overall behavior. This is important because the two types of distributions are so different, that if corresponding results are

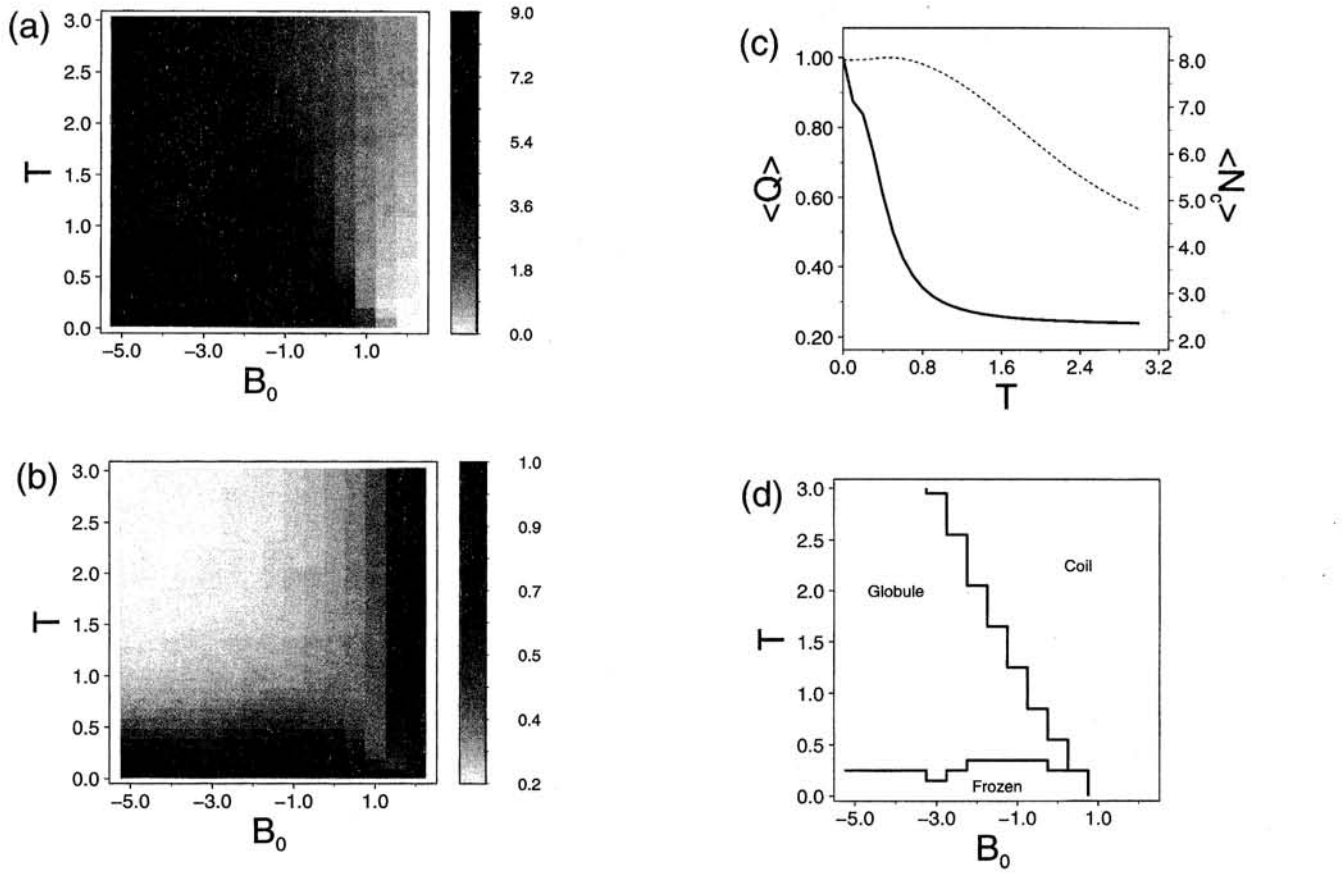

FIG. 3. The same as Fig. 1 for another Gaussian sequence. The triple point for this sequence occurs at $(B_0,T)=(0.25,0.25)$.

obtained for the two of them, they can be expected to be rather general.

The overall behavior of the phase diagram is dependent primarily on the stability of the ground state relative to the remaining part of the conformational ensemble. However, irregularities exhibited by individual sequences arise from excited states with low energies. For example, in Fig. 3(c), there is a kink in $\langle Q(T)\rangle$; this results from two structures in the lower part of the energy spectrum which are almost degenerate and thus become thermodynamically important at the same temperature. Another irregularity exhibited by two sequences is the depression in $\langle N_c\rangle$ at low T for $B_0\approx 1$ in Figs. 2(a) and 3(a). Here, the lowest energy structure has one contact, so that $\langle N_c\rangle=1.0$ at $T=0$. At low $T>0$, zero contact structures play a very significant role in the average, so that $\langle N_c\rangle$ decreases with a small increase in T; however, at even higher temperatures, more compact structures with high energies decrease the weight of the completely open structures, thereby increasing $\langle N_c\rangle$. While such features of the phase diagram illustrate the importance of nonground state structures in the model, they depend on the details of the sequence. Thus, their general implications may be limited.

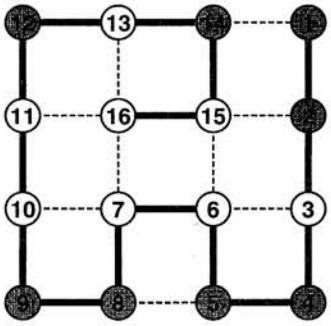

FIG. 4. The nondegenerate native state of the two-letter sequence for $B_0<0$. Monomers are colored according to type with A (white) and B (gray). $(B_0,\Delta E_{10})=\{(-5.0,2.0),(-4.0,2.0),(-3.0,2.0),(-2.0,2.0),(-1.0,1.0)\}$; only one conformation is shown because the ground state remains constant over this B_0 interval in spite of the change of its energetic separation from the rest of the conformational ensemble. For $B_0=0.0$, the ground state is 12-fold degenerate and includes the conformation shown. At $B_0\geq 1.0$, the ground state includes all zero contact structures (116 579-fold degeneracy).

In the diagram for all three sequences, there is a triple point where the three phases coexist. This means that there are conditions (low T and high B_0), where the molecule can make a direct transition from the coil to the frozen state. This transition comes from a transformation in the ground state due to the change in B_0 (Figs. 4–6). For example, the coil to frozen state (C to F) transition of the two-letter sequence takes place in the region in which the ground state gradually goes from a five-fold degenerate state with seven contacts to a maximally compact conformation (nine contacts for $B_0<0$). Likewise, the two Gaussian sequences experience C

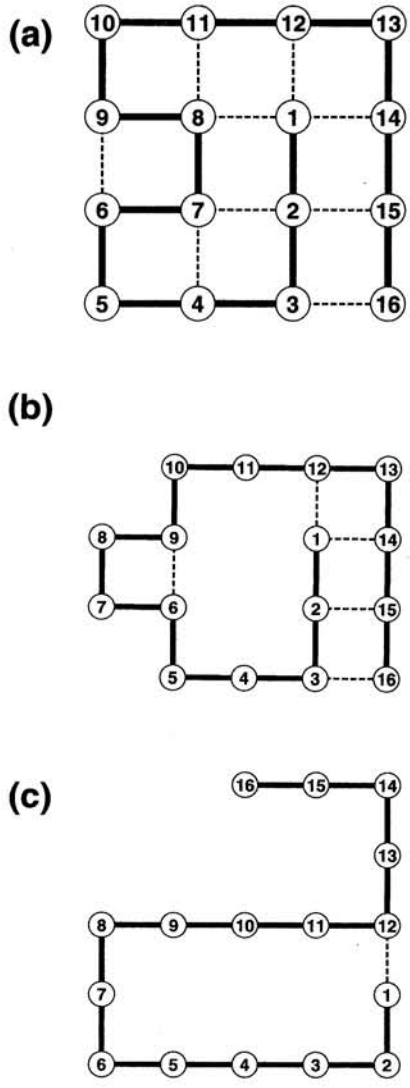

FIG. 5. Ground state of the first Gaussian sequence. (a) The nondegenerate native state for $B_0<1.0$. $(B_0,\Delta E_{10})=\{(-5.0,4.81),\ (-4.0,4.81),\ (-3.0,4.04),\ (-2.0,3.04),\ (-1.0,2.04),\ (0.0,1.04)\}$; only one conformation is shown because the ground state remains constant over this B_0 interval in spite of the change of its energetic separation from the rest of the conformational ensemble. (b) The nondegenerate native state for $B_0=1.0$, $\Delta E_{10}=0.04$. (c) The 314-fold degenerate ground state for $B_0=2.0$ (related to each other by noncontacting tail moves). At $B_0>2.0$, the ground state includes all zero contact structures.

to F transitions when their ground states go from five- to nine-contact and three- to seven-contact structures, respectively. The change of ground state to a maximally compact structure in the second Gaussian sequence can be seen in the phase diagram from the depression in the globule to frozen (G to F) boundary at $B_0=-3.0$ [Figs. 3(d) and 6]. The noncompact and degenerate nature of the ground (F) state for cases where a direct C to F transition occurs suggests that it is not of great physical interest.

To look at the agreement between the present results and the mean-field theory of heteropolymers, we compare the Gaussian sequence phase diagrams [Figs. 2(d) and 3(d)] with the one derived in Ref. 3. Because the latter was plotted in the variables B_0 and B, which are more appropriate for studies using ensembles of sequences, we recalculate the theoretical phase diagram in the variables employed in the present work, B_0 and T (Fig. 7). The line delineating the G to F boundary is determined by the equation

$$T_c=\frac{B\sqrt{\rho(B_0,T_c)}}{2k\sqrt{\ln\gamma}}, \tag{5}$$

where ρ is the average number of nonlocal contacts per monomer and γ is the number of conformations per monomer.[11,22] The dependence of ρ on B_0 and T can be found using the Flory–Huggins approximation, which has been adopted previously to study coil–globule transitions.[23] The equilibrium value of ρ is that which minimizes the free energy, which takes the following form in this approximation:

$$F=B_{\text{eff}}N\rho+\frac{NkT}{\rho}(Z-2)\left(1-\frac{\rho}{Z-2}\right)\ln\left(1-\frac{\rho}{Z-2}\right), \tag{6}$$

where $B_{\text{eff}}=B_0-B$ is the effective interaction energy corrected for heteropolymeric effects,[3] Z is the average coordination number of the lattice (in our example of a 16-mer on a square lattice, $Z=3$), and $\rho/(Z-2)$ is the fraction of sites occupied by monomers. Using the value of ρ obtained from minimization of Eq. (6) with $\gamma=2.6$ (which takes into account excluded volume) and $B=1$, we numerically solve Eq. (5) for T_c, which is the G to F boundary of Fig. 7. We define the other phase boundary (C to G) as a line of effective Θ points (T_Θ) at which the coefficient of the term linear in ρ in the small ρ expansion of Eq. (6) vanishes

$$B_0-B+\frac{kT_\Theta}{2(Z-2)}=0. \tag{7}$$

The resulting $T_\Theta(B_0)$ line is the coil to globule boundary in Fig. 7. The 16-mer phase diagrams [Figs. 2(d) and 3(d)] presented in this study are in clear qualitative agreement with the analytical theory (Fig. 7).

While this relatively simple model of a short chain on a 2D lattice exhibits rich behavior and has three distinct states, determination of the kinetic consequences of the phase diagram would require a separate investigation. From the form of the phase diagram, the C to G transition is likely to precede the G to F transition, particularly for compact native states. The slow smooth decrease in $\langle N_c\rangle$ with temperature [Figs. 1(c), 2(c), and 3(c)] suggests that there is no significant barrier between low and high densities, resulting in a quick collapse. On the other hand, although a 16-mer is too short to provide clear proof that intermediate Q's are unstable, the rapid decrease in $\langle Q(T)\rangle$ [Figs. 1(c), 2(c), and 3(c)] is consistent with the analytical prediction that there is a high free energy barrier between the G and F states. Such a kinetic scheme would imply that the disordered globule may be a real kinetic intermediate in the process of folding. This interpretation is consistent with both kinetic experiments[24] and simulations of folding with a 3D model.[14,15] In the latter, a rapid compactization to the disordered globule is followed by a relatively slow search to find the native state within the set of compact conformations.

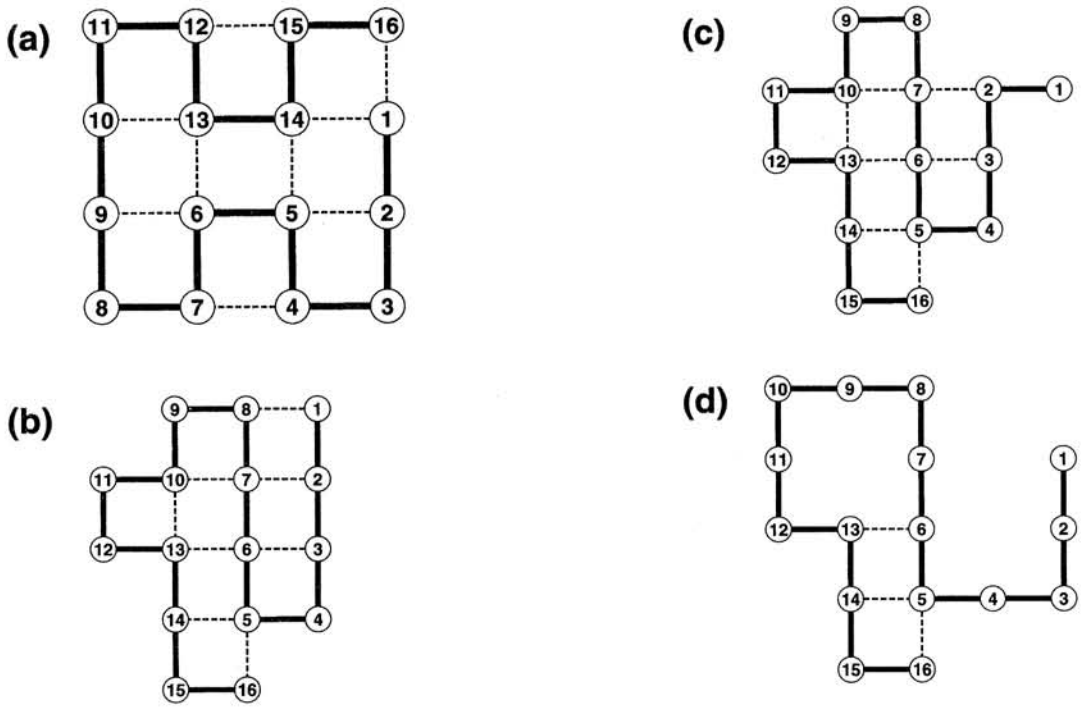

FIG. 6. The ground state of the second Gaussian sequence. (a) The nondegenerate native state for $B_0<-2.0$. $(B_0,\Delta E_{10})=\{(-5.0,0.68),(-4.0,0.68),(-3.0,0.09)\}$; only one conformation is shown because the ground state remains constant over this B_0 interval in spite of the change of its energetic separation from the rest of the conformational ensemble. (b) The nondegenerate native state for $-2.0<B_0<0.0$. $(B_0,\Delta E_{10})=\{(-2.0,0.02),(-1.0,0.02)\}$. (c) The nondegenerate native state for $B_0=0.0$, $\Delta E_{10}=0.02$. (d) The sevenfold degenerate ground state for $B_0=1.0$ (related to each other by noncontacting tail moves). At $B_0>1.0$, the ground state includes all zero contact structures.

Although the present results are clearly of interest for heteropolymer theory in 2D, the applicability to 3D systems, including proteins, has to be assessed. The replica mean-field theory of heteropolymers[3,17] predicts that the properties of the frozen phase will have a strong dependence on spatial dimensionality. This dependence arises primarily from differences in the degree of structural dissimilarity between minima. This energy-structure relationship has a direct bearing on the kinetics of folding, and we believe that kinetic conclusions drawn from 2D models may have limited validity for 3D. Furthermore, numeric details (e.g., transition points) may be quite different in 2D and 3D. Consequently, the thermodynamic analysis given in this paper has only qualitative significance for general systems.

One limitation in comparison with proteins is that the simplicity of the model prevents explicit consideration of the role of secondary structure formation. While this has been shown to be important in experimental studies,[25,26] the introduction of such details in this model might introduce artifacts.[27] Also, the lack of sidechains makes it unclear to what experimentally observed state the frozen globule corresponds. It is tempting to identify it with the molten globule (MG) state[1,28,29] because it has been argued that the MG state can be obtained from the native one by destruction of the sidechain tight packing with little distortion of the backbone.[1,8] In spite of its shortcomings, the simple model provides useful information concerning the phase space of a protein-like heteropolymer. In particular, it makes possible an exhaustive enumeration of all configurations which would be impossible in a more complete description of a protein.

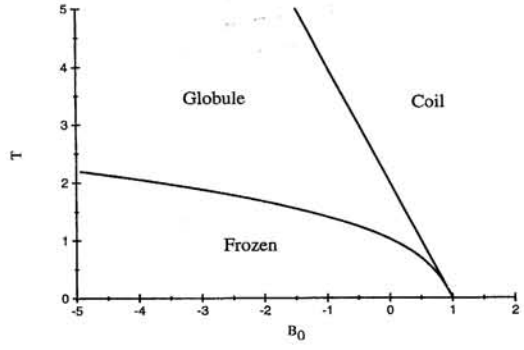

FIG. 7. Phase diagram calculated from the results of the replica mean-field theory of heteropolymers (Ref. 3). See the text for the method of derivation.

ACKNOWLEDGMENTS

The work was supported in part by grants from the National Institutes of Health and National Science Foundation and a gift from Molecular Simulations, Inc. E. S. is supported by a Packard Fellowship. A. Š. is a Fellow of The Jane Coffin Child Memorial Fund for Medical Research. The calculations were done on an IBM RS6000 550 and Silicon Graphics 4D-340.

[1] M. Karplus and E. Shakhnovich, in *Protein Folding*, edited by T. Creighton (Freeman, New York, 1992).
[2] C. L. Brooks III, M. Karplus, and B. M. Pettitt, *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics* (Wiley, New York, 1988).
[3] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[4] P. Privalov, in *Protein Folding*, edited by T. Creighton (Freeman, New York, 1992).
[5] K. Dill, Biochemistry **24**, 1501 (1985).
[6] O. Ptitsyn, in *Protein Folding*, edited by T. Creighton (Freeman, New York, 1992).
[7] E. I. Shakhnovich and A. V. Finkelstein, Dokl. Akad. Nauk. SSSR **243**, 1247 (1982).
[8] E. I. Shakhnovich and A. V. Finkelstein, Biopolymers **28**, 1667 (1989).
[9] J. W. Ponder and F. M. Richards, J. Mol. Biol. **193**, 775 (1987).
[10] P. Kim and R. Baldwin, Annu. Rev. Biochem. **59**, 631 (1990).
[11] E. I. Shakhnovich and A. M. Gutin, Nature **346**, 773 (1990).
[12] D. Stein, Proc. Natl. Acad. Sci. USA **82**, 3670 (1985).
[13] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[14] E. I. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[15] A. Šali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[16] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[17] E. Shakhnovich and A. Gutin, J. Phys. A **22**, 1647 (1989).
[18] D. Shortle, H. S. Chan, and K. A. Dill, Protein Sci. **1**, 201 (1992).
[19] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).
[20] H. S. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116 (1993).
[21] C. Sfatos, A. M. Gutin, and E. I. Shakhnovich, Phys. Rev. E **48**, 465 (1993).
[22] E. I. Shakhnovich and A. M. Gutin, J. Theor. Biol. **149**, 537 (1991).
[23] A. V. Finkelstein and E. I. Shakhnovich, Biopolymers **28**, 1681 (1989).
[24] G. A. Elöve, A. F. Chaffotte, H. Roder, and M. E. Goldberg, Biochemistry **31**, 6876 (1992).
[25] J. Udgaonkar and R. Baldwin, Nature **335**, 694 (1988).
[26] S. Radford, C. Dobson, and P. Evans, Nature **358**, 302 (1992).
[27] L. Gregoret and F. Cohen, J. Mol. Biol. **219**, 109 (1991).
[28] D. A. Dolgikh, R. I. Gilmanshin, E. V. Brazhnikov, V. E. Bychkova, G. V. Semisotnov, S. Yu. Venyaminov, and O. B. Ptitsyn, FEBS Lett. **136**, 311 (1981).
[29] D. A. Dolgikh, R. I. Gilmanshin, E. V. Brazhnikov, V. E. Bychkova, G. V. Semisotnov, S. Yu. Venyaminov, and O. B. Ptitsyn, Eur. Biophys. J. **13**, 109 (1985).