A Minimal Physically Realistic Protein-Like Lattice Model: Designing an Energy Landscape that Ensures All-Or-None Folding to a Unique Native State

Piotr Pokarowski,* Andrzej Kolinski,[†‡] and Jeffrey Skolnick[‡]

*Institute of Applied Mathematics and Mechanics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland; [†]Laboratory of Theory of Biopolymers, Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland; and [‡]Donald Danforth Plant Science Center, Bioinformatics and Computational Genomics, 975 N. Warson Rd., Saint Louis, Missouri 63141 USA

ABSTRACT A simple protein model restricted to the face-centered cubic lattice has been studied. The model interaction scheme includes attractive interactions between hydrophobic (H) residues, repulsive interactions between hydrophobic and polar (P) residues, and orientation-dependent P-P interactions. Additionally, there is a potential that favors extended β-type conformations. A sequence has been designed that adopts a native structure, consisting of an antiparallel, six-member Greek-key β-barrel with protein-like structural degeneracy. It has been shown that the proposed model is a minimal one, i.e., all the above listed types of interactions are necessary for cooperative (all-or-none) type folding to the native state. Simulations were performed via the Replica Exchange Monte Carlo method and the numerical data analyzed via a multihistogram method.

INTRODUCTION

Despite their enormous conformational space, small globular proteins rapidly fold to a well-defined densely packed native structure and with a transition that resembles a first-order phase transition (Ptitsyn, 1987; Anfinsen, 1973; Jackson, 1998). Due to the small size (several hundreds of atoms) of a protein that precludes any notion of the thermodynamic limit, this abrupt and cooperative folding transition is frequently abbreviated as the all-or-none transition to underline the very small population of folding intermediates at the transition temperature (Shakhnovich and Finkelstein, 1989b; Scheraga et al., 1995). In this paper, we attempt to design a minimal protein-like model that in a qualitative way mimics the most pronounced features of globular proteins (Baker, 2000). These features include: the existence of a lowest energy native state that has secondary structure features, a well-defined hydrophobic core, and a unique, quite complicated, Greek-key (Branden and Tooze, 1991) topology. Additionally, the model has to reproduce a cooperative all-or-none folding transition and the cooperative formation of secondary structure upon the collapse (or folding) transition (Shakhnovich and Finkelstein, 1989a). The last features are the major difference between this model and other well-known simple-exact cubic lattice models (Dill et al., 1995; Abkevich et al., 1996; Dinner et al., 1996; Karplus and Sali, 1995). We also provide proof that the designed model is indeed a minimal model, i.e., that one needs all the proposed components of the interaction scheme to ensure the above outlined protein-like features are present.

The protein model we adopt is a face-centered cubic lattice chain, with the chain beads representing the polypeptide amino acid units. Each amino acid residue is characterized by two fundamental properties: its hydrophobicity (that dictates the character of the binary interactions) and its secondary structure propensity (that encodes the tendency to adopt a specific rotational-isomeric state of a chain fragment). As demonstrated in many earlier studies, such an interplay between the short- and long-range interactions leads to cooperative collapse transitions in a finite length polymer (Kolinski and Skolnick, 1996; Kolinski et al., 1986; Kolinski et al., 1996; Post and Zimm, 1979). Here, for the first time, we provide quantitative arguments that the existence of both types of interactions is actually a necessary condition for protein-like behavior.

Submitted June 28, 2002, and accepted for publication October 30, 2002.

Address reprint requests to Andrzej Kolinski, E-mail: Kolinski@chem.uw.edu.pl or Skolnick@buffalo.edu.

0006-3495/03/03/1518/09 $2.00

PROTEIN MODEL

In this section, a detailed description of the model is provided. The purpose of the rigorous math-type definition of the model is to provide a convenient and precise notation for the following section, where the model's energy landscape (Bryngelson et al., 1995; Onuchic et al., 1997) is analyzed in considerable detail.

Representation of protein conformation

The model polypeptide is restricted to a face-centered cubic lattice (fcc). There are 12 orientations of the fcc vectors, which form a *BASE*, base set, of the lattice. This set could be written as:

$$BASE = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{12}\}, \tag{1}$$

where: $\mathbf{e}_1 = (1,1,0)$, $\mathbf{e}_2 = (1,-1,0)$, $\mathbf{e}_3 = (1,0,1)$, $\mathbf{e}_4 = (1,0,-1)$, $\mathbf{e}_5 = (0,1,1)$, $\mathbf{e}_6 = (0,1,-1)$, $\mathbf{e}_7 = (0,-1,1)$, $\mathbf{e}_8 = (0,-1,-1)$, $\mathbf{e}_9 = (-1,0,1)$, $\mathbf{e}_{10} = (-1,0,-1)$, $\mathbf{e}_{11} = (-1,1,0)$, and $\mathbf{e}_{12} = (-1,-1,0)$.

The fcc lattice, *FCC*, may be defined by induction as follows:

$$(0,0,0) \in FCC, \tag{2a}$$

$$\text{and: if } \mathbf{e} \in BASE,\ \mathbf{x} \in FCC,\ \text{then } \mathbf{x} + \mathbf{e} \in FCC. \tag{2b}$$

Points $\mathbf{x}_1, \mathbf{x}_2 \in FCC$ are neighbors on the lattice if there exists $\mathbf{e} \in BASE$, such that $\mathbf{x}_2 = \mathbf{x}_1 + \mathbf{e}$. We write in this case: $\mathbf{x}_1 \sim \mathbf{x}_2$.

Let $CHAIN = \{1, \ldots, N\}$ be a set of residues in a polypeptide chain. A structure of a polypeptide is represented on the lattice by a function $\mathbf{s}$: $CHAIN \rightarrow FCC$, which satisfies the following three conditions:

$$\mathbf{s}(1) = (0,0,0), \tag{3a}$$

$$\text{if } i \in CHAIN,\ i < N,\ \text{then } \mathbf{s}(i+1) \sim \mathbf{s}(i), \tag{3b}$$

$$\text{if } i,j \in CHAIN,\ i \neq j,\ \text{then } \mathbf{s}(i) \neq \mathbf{s}(j). \tag{3c}$$

We will identify a structure with its representation on the lattice, and we denote by S the set of all structures. S will be called the conformational space. It is easily seen that

$$\#\mathrm{S} < 12 \times 11^{\mathrm{N}-2}, \tag{4}$$

where the symbol # denotes the number of elements in a set. The above inequity reflects the ''attrition'' of conformations due to the excluded volume of the chain.

Representation of the polypeptide sequence

A sequence of the chain is defined by its hydrophobic pattern Pat: $CHAIN \rightarrow \{\mathrm{H,P}\}$ and its secondary structure Sec:$CHAIN \rightarrow \{\beta,\mathrm{C}\}$. This means that from the point of view of the long-range pairwise interactions, there are two types of residues (Dill et al., 1995): nonpolar, hydrophobic (H) and polar (P). Moreover, on the level of secondary structure, or chain stiffness, β stands for extended, β-type short-range interactions, and C denotes the flexible coil, or loop, regions. Thus, the model employs a four-letter sequence code.

Interaction scheme

The definition of a model polypeptide sequence implies two main types of molecular interactions. First, the long-range interactions depend on the number of contacts between residues. Let $\mathbf{v}_i(\mathbf{s})$ be a vector from $\mathbf{s}(i)$ to $\mathbf{s}(i+1)$. We will write it simply as $\mathbf{v}_i$. A pair of vectors, $(\mathbf{v}_{i-1},\mathbf{v}_i)$ and $(\mathbf{v}_{j-1},\mathbf{v}_j)$, are called parallel (notation: $\mathbf{v}_{i-1},\mathbf{v}_i \parallel \mathbf{v}_{j-1},\mathbf{v}_j$) if either $\mathbf{v}_{i-1} = \mathbf{v}_{j-1}$ and $\mathbf{v}_i = \mathbf{v}_j$ or $\mathbf{v}_{i-1} = -\mathbf{v}_j$ and $\mathbf{v}_i = -\mathbf{v}_{j-1}$. For a given structure $\mathbf{s}$, we define functions counting three types of long-range contacts between residues:

$$K_{\mathrm{HH}}(\mathbf{s}) = \#\{\{i,j\}:\ |i-j|>1,\ \mathbf{s}(i) \sim \mathbf{s}(j),\ \mathrm{Pat}(i) = \mathrm{Pat}(j) = H\}, \tag{5a}$$

$$K_{\mathrm{HP}}(\mathbf{s}) = \#\{\{i,j\}:\ |i-j|>1,\ \mathbf{s}(i) \sim \mathbf{s}(j),\ \mathrm{Pat}(i) \neq \mathrm{Pat}(j)\}, \tag{5b}$$

$$K_{\mathrm{PP}}(\mathbf{s}) = \#\{\{i,j\}:\ |i-j|>2,\ \mathbf{s}(i) \sim \mathbf{s}(j),\ \mathrm{Pat}(i) = \mathrm{Pat}(j) = P,\ \mathbf{v}_{i-1},\mathbf{v}_i \parallel \mathbf{v}_{j-1},\mathbf{v}_j\}. \tag{5c}$$

Note that PP interactions are counted only for the residues contacting in a parallel fashion, reflecting the tendency of the parallel packing of polar side chains on the surface of a protein (Ilkowski et al., 2000).

The short-range interactions simulate the local conformational stiffness of the polypeptide chains. Here, for illustration, we limited ourselves to the case of β-type proteins. Let us denote by $\mathbf{x} \cdot \mathbf{y}$ the dot product of vectors $\mathbf{x}$, $\mathbf{y}$. The number of residues with preferences to be in β-strands is defined as follows:

$$K_\beta(\mathbf{s}) = \#\{i:\ \mathrm{Sec}(i) = \beta,\ \mathbf{v}_{i-2}\cdot\mathbf{v}_{i-1} = \mathbf{v}_{i-1}\cdot\mathbf{v}_i = 1,\ \mathbf{v}_{i-2}\cdot\mathbf{v}_i = 2\}. \tag{6}$$

The geometric conditions mean that a given three-bond fragment has the most expanded conformation with its planar angles equal to 120°.

Let $\mathbf{K}(\mathbf{s}) = (K_{\mathrm{HH}}(\mathbf{s}),K_{\mathrm{HP}}(\mathbf{s}),K_{\mathrm{PP}}(\mathbf{s}),K_\beta(\mathbf{s}))$ be a vector defining the numbers of various interactions and $\boldsymbol{\varepsilon} = (\varepsilon_{\mathrm{HH}},\varepsilon_{\mathrm{HP}},\varepsilon_{\mathrm{PP}},\varepsilon_\beta)$ be a vector of weights, or the force-field parameters. The conformational energy of a structure $\mathbf{s}$ is, by definition, a linear combination of its contacts:

$$E(\mathbf{s}) = \boldsymbol{\varepsilon}\cdot\mathbf{K}(\mathbf{s}). \tag{7}$$

Recently, we have shown that this model exhibits a highly cooperative all-or-none collapse transition (Gront et al., 2001) into a three-dimensional structure of unique Greek-key topology (Branden and Tooze, 1991). Here, we would like to show that a very similar model is indeed minimal, i.e., that the design of the force field is not accidental and that one needs nonzero values of all the proposed interactions to obtain a protein-like folding transition. The same, quite complex topology of the native state is assumed, which is an antiparallel six-stranded Greek-key β-barrel typical for a significant fraction of real β-type proteins. The force field has been simplified with respect to the previously studied model (Gront et al., 2001). The present model constitutes a highly simplified version of our older studies of a Greek-key folding motif (Kolinski et al., 1995), where the effect of multibody potentials on protein dynamics and thermodynamics were investigated in a framework of high coordination lattice model of polypeptide chain (Kolinski et al., 1996).

Definition of the target native structure

The target structure is an ''ideal,'' six-stranded, antiparallel, β-barrel motif with a Greek-key topology, assumed to be a lattice representation of the ''native structure'' (see Fig. 1 *A*). Using the numbers representing the *BASE* vectors, this structure could be abbreviated as follows:

$$\mathrm{Native} = 11,(6,11)_{\mathrm{L}},9,(7,2)_{\mathrm{L}},7,1,(6,11)_{\mathrm{L}},6,4,3,(7,2)_{\mathrm{L}},1,11,(6,11)_{\mathrm{L}},9,(7,2)_{\mathrm{L}},7, \tag{8}$$

The following sequence has been designed to be consistent with the above six-stranded β-barrel structure:

$$\mathrm{Sec}(CHAIN) = (\mathrm{C}_2\beta_{2\mathrm{L}-1}\mathrm{C})_6, \tag{9a}$$

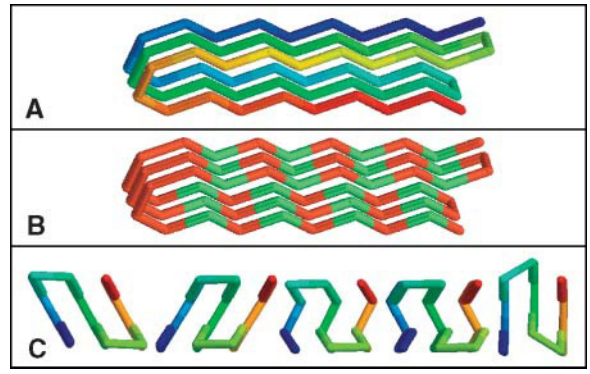

FIGURE 1 Model six-stranded Greek-key β-barrel on the face-centered cubic lattice. (*A*) The N-terminus is shown in blue and the C-terminus in red. (*B*) Illustrates the well-defined hydrophobic core of the barrel (hydrophobic residues shown in *green*). The polar residues (shown in *red*) point outside of the structure. The loops contain longer sequences of the polar residues. (*C*) Shows five distinct conformations of the native state (*viewed from the top of the barrel*). The total number of native conformations is 20 (see the text for details). All have exactly the same conformational energy and pattern of interactions. They differ in small details involving the mutual positions of the β-strands.

$$\mathrm{Pat}(CHAIN) = ((\mathrm{PH})_{\mathrm{L}}\mathrm{P}_3(\mathrm{PH})_{\mathrm{L}}\mathrm{P})_3, \qquad (9b)$$

where L is a number of hydrophobic residues in one strand of the native structure. L is a parameter of the size of the model, and it is easy to verify that the number of residues $N = 12(L + 1)$. The designed model appears to satisfy the principle of ''minimal'' (energetic) frustration, in the sense that the PH pattern and the pattern of secondary propensities are consistent with the structure shown in Fig. 1 *B*. The native structure for this model is degenerate, i.e., several very similar (but not the same) structures have exactly the same pattern of interactions with a well-defined hydrophobic core, polar surface, and clearly defined secondary structure. As shown in Fig. 1 *C*, there are five basic variants of the native structure depending on the projection of the structure onto the plane containing the ends of the chain and the tops of the β-hairpins. Each of these structures has two substructures that differ with the location of a single 90° planar angle near the hairpin end. Finally, each substructure has a mirror image topology, due to the lack of any chiral interactions in the model. All together, the native structure appears in 20 forms. The forms have exactly the same (the same contribution from various components of the interaction scheme) interactions and the same topology (modulo their mirror image) but have a slightly different detailed geometry. This conformational degeneracy of the model's native state is probably quite physical. Indeed, in real proteins the native structure could be quite mobile, with (at least) changes of side-chain rotamers. In both cases (in the model and real proteins), the degeneracy of the native state provides an entropic stabilization of the native structure.

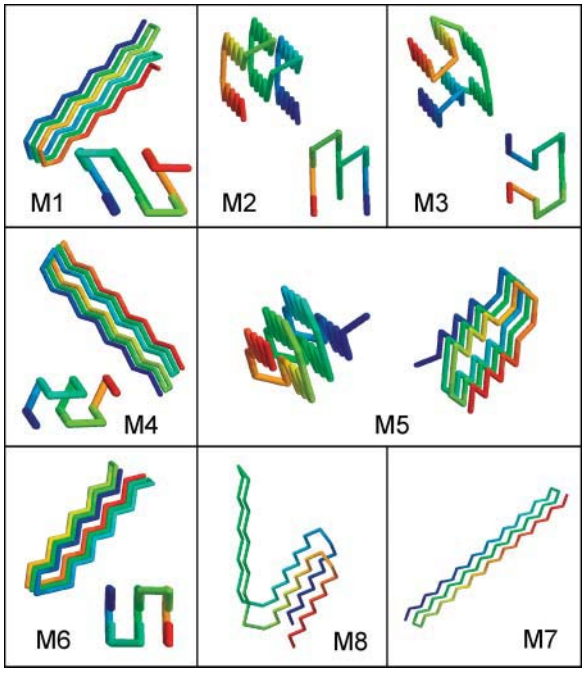

FIGURE 2 Snapshots of the eight low-energy conformations (for some values of interaction parameters) that might compete with the native structure. More complex structures are shown in two alternative projections.

RESULTS

Native state and alternative low-energy conformations

To explore the conformational space of the model, we performed a large number of Replica Exchange Monte Carlo simulations (Hansmann, 1997; Hansmann and Okamoto, 1997; Hansmann and Okamoto, 1999; Gront et al., 2000; Gront et al., 2001; Hukushima and Nemoto, 1996; Sugita and Okamoto, 1999; Swendsen and Wang, 1986) for various values of $L = 2,3,4$ (only the results for $L = 4$ are discussed in detail) and different vectors of interaction parameters $\boldsymbol{\varepsilon}$. We found all 20 different forms of our native structure and a collection of regular, nonnative structures which, depending on the model parameters, were the global minimum of energy. These competitive structures are schematically shown in Fig. 2, and their lattice representations are listed in Table 1. The interaction patterns in these structures were analyzed in detail, and the results are given in Table 2. While competitive structures (and the native structures) were found many times in various simulations, no other lower energy structure was ever recorded, regardless of the very broad range of the interaction parameters explored.

According to Anfinsen's hypothesis (Anfinsen, 1973), for a meaningful model the native structure has to be of minimum conformational energy. To have ''protein-like'' energy landscape the conformational energy of the native state needs to be lower than the energy of all competitive structures (misfolds M_1–M_8). This implies the following system of linear inequalities:

$$\mathrm{E(Native)} < \mathrm{E(M_i)} \quad i = 1, 2 \ldots 8 \qquad (10)$$

According to the interaction patterns provided in Table 2, for $i = 1,2,\ldots 8$, the above inequalities imply the following set of relations between the parameters of the models:

$$\begin{aligned}
&(11.1) \quad \varepsilon_{\mathrm{HP}} < -\varepsilon_\beta \\
&(11.2) \quad -\varepsilon_{\mathrm{PP}} < (\varepsilon_{\mathrm{HP}} - \varepsilon_\beta)/2 \\
&(11.3) \quad -4\varepsilon_{\mathrm{PP}} > 0 \\
&(11.4) \quad -\varepsilon_{\mathrm{PP}} < 7L\varepsilon_{\mathrm{HP}}/(5L + 2) \\
&(11.5) \quad -\varepsilon_{\mathrm{HH}} < (6\varepsilon_{\mathrm{HP}} + (4\mathrm{L} + 1)\varepsilon_{\mathrm{PP}} \\
&\qquad\qquad - (12L - 9)\varepsilon_\beta)/(12L - 12) \\
&(11.6) \quad -\varepsilon_{\mathrm{HH}} < (2L + 1)\varepsilon_{\mathrm{HP}}/(L - 1) \\
&(11.7) \quad -5L\varepsilon_{\mathrm{HH}} > 0 \\
&(11.8) \quad -\varepsilon_{\mathrm{HH}} > (4L\varepsilon_{\mathrm{HP}} + 4(L + 1)\varepsilon_{\mathrm{PP}})/(5L)
\end{aligned} \qquad (11)$$

A simple consequence of the above inequalities is that our force field is minimal. It is easy to see that $\varepsilon_{\mathrm{HP}}$, $-\varepsilon_{\mathrm{HH}}$, $-\varepsilon_{\mathrm{PP}}$, $-\varepsilon_\beta > 0$. Indeed, inequalities (11.3) and (11.7) trivially mean that $-\varepsilon_{\mathrm{PP}} > 0$ and $-\varepsilon_{\mathrm{HH}} > 0$, respectively. The last condition, together with (11.6), gives $\varepsilon_{\mathrm{HP}} > 0$. Similarly, from (11.1) one obtains $-\varepsilon_\beta > 0$. Let us also note that the requirement of the parallel contacts of the polar residues in the definition of K_{PP} is a necessary one. Without such an assumption, the competitive structure M_3 would have exactly the same pattern of interactions $\mathbf{K}$ as the native ones. Consequently, the force field seems to be the simplest one able to satisfy the thermodynamic hypothesis in the context of our lattice model. Thus we have shown that the

TABLE 1 Formulas for native and all competitive structures

Name	Formula
Native	$11,(6,11)_L,9,(7,2)_L,7,1,(6,11)_L,6,4,3,(7,2)_L,1,11,(6,11)_L,9,(7,2)_L,7$
M1	$11,(6,11)_L,9,(7,2)_L,7,1,(6,11)_L,6,4,3,(7,2)_L,1,11,(6,11)_L,9,(7,2)_L,1$
M2	$8,(10,8)_L,2,3,(5,3)_L,6,10,(10,8)_L,2,3,(5,3)_L,6,(10,8)_L,10,11,(5,3)_L,5$
M3	$11,(6,11)_L,12,2,(7,2),8,11,(6,11)_L,9,(7,2),7,5,(6,11)_L,6,1,(7,2),7$
M4	$11,(9,11)_L,10,(2,4)_L,2,12,11,(9,11)_L,3,(2,4)_L,2,12,11,(9,11)_L,10,(2,4)_L,2$
M5	$11,(5,12)_{L-1},9,5,10,8,4,(8,1)_L,5,(5,12)_{L-1},9,5,10,8,4,(8,1)_L,5,(5,12)_{L-1},9,5,10,8,4,(8,1)_{L-1},8$
M6	$1,(3,1)_L,10,(10,12)_L,10,8,(3,1)_L,3,3,12,(10,12)_L,8,1,(3,1)_L,10,(10,12)_L,10$
M7	$11,(6,11)_{2L+1},1,2,(7,2)_{2L+1},1,11,(6,11)_{2L+1}$
M8	$6,(10,6)_L,5,3,(7,3)_L,1,(6,11)_L,6,8,(2,7)_L,2,12,(10,6)_L,10,8,(7,3)_L,7$

As explained in the text the native structure exists in twenty conformations (given the first vector fixed). Similarly, the competing structures have multiple conformations with the same pattern of interactions.

model force field is minimal, i.e., to have a protein-like model, one needs all types of interactions considered in this work.

The above statement about a minimal character of the interaction scheme relies on our definition of the native state and the assumption that the other low-energy states found in a broad range of interaction parameters are nonnative, misfolded structures. Let us discuss these misfolds in more detail, pointing out their non protein-like features. They are abbreviated with the symbols M1–M8 in Fig. 2. M1 differs from the native only with the orientation of a single residue on the C-terminus. Instead of pointing along the barrel, it points sidewise. It was assumed that these kinds of conformations deviate from the regular Greek-key topology and are less ''protein-like'' than the native target structure. Structure M2, although a compact and regular one, has a different topology. Two loops placed on top of each other are not typical of globular proteins. The topology of M3 is wrong, and some of its P-P contacts are not parallel and, therefore, are not counted. The M4 and M6 structures have a poorly defined hydrophobic core. Interestingly, structures M4-M6 are more compact than the native one. Simply, a number of polar residues are buried. Additionally M5 lacks most of extended β-type secondary structure. Structures M7 and M8 do not have well-defined hydrophobic core and are not completely folded. As one might intuitively expect (and it is apparent from the quantitative analysis of the following sections), these misfolds result from a wrongly balanced strength of various (short-range and long-range) interactions.

TABLE 2 Number of contacts in the native and competitive structures

Name	K_{HH}	K_{HP}	K_{PP}	K_β
Native	9L	4L	4L+4	12L−6
M1	9L	4L−1	4L+4	12L−7
M2	9L	4L+1	4L+6	12L−7
M3	9L	4L	4L	12L−6
M4	9L	11L	9L+6	12L−6
M5	21L−12	4L+6	8L+5	3
M6	14L−5	14L+5	4L+4	12L−6
M7	4L	4L	4L+4	12L−6
M8	4L	0	0	12L−6

An upper bound for a set of good parameters

Let us denote by E a set of good parameters, i.e., a set of such $\boldsymbol{\varepsilon}$, for which the native structure corresponds to the global minimum of conformational energy. Obviously, for every pair of structures $\mathbf{s}_1, \mathbf{s}_2 \in \mathsf{S}$ and a positive number a, conditions $E(\mathbf{s}_1) < E(\mathbf{s}_2)$ and $aE(\mathbf{s}_1) < aE(\mathbf{s}_2)$ are equivalent. Therefore, without loss of generality we can assume that $-\varepsilon_\beta = 1$ and identify $\boldsymbol{\varepsilon} \in \mathsf{E}$ with the restrictions on $(\varepsilon_{HH},\varepsilon_{HP},\varepsilon_{PP})$. It is easy to see that the system of inequalities (11) is satisfied if and only if $\boldsymbol{\varepsilon} \in \mathsf{E}_U$, where E_U is the convex polyhedron given by the vertices listed in Table 3. Obviously E_U is an upper bound for E, which means that $\mathsf{E} \subseteq \mathsf{E}_U$. The shape of the E_U is schematically drawn in Fig. 3. Of course, the specific shape of E_U depends on the choice of the set of competitive structures, which define the set of inequalities as the one given in Eq. 11. On the other hand, the competitive structures were selected very carefully, and they appear to be representative. We searched for the lowest energy conformations over a broad range of interaction parameters and found no other low energy structures. Therefore, it is very unlikely that adding more structures to the set of competitive structures could significantly change the estimation of the upper bound for the set of good parameters of the model.

A lower bound for a set of good parameters

In this section we are concerned with a lower bound for E. Let $\boldsymbol{\varepsilon}_0$ be the center of mass of E_U. By the definition $\boldsymbol{\varepsilon}_0 = (\boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_2 + \cdots + \boldsymbol{\varepsilon}_{10})/10$. Let us also define the set $\boldsymbol{\varepsilon}_i(\delta) = \delta\boldsymbol{\varepsilon}_i + (1 - \delta)\boldsymbol{\varepsilon}_0$, where $i = 1,2,\ldots,10$ (enumerating the vertices

TABLE 3 Vertices of E_U under the assumption that: $-\varepsilon_\beta = 1$

Vertex	$(-\varepsilon_{HH}, \varepsilon_{HP}, -\varepsilon_{PP})$
$\boldsymbol{\varepsilon}_1$	$((4L-3)/(8L+2)) \cdot ((2L+1)/(L-1), 1, 0)$
$\boldsymbol{\varepsilon}_2$	$((4L-3)/((37L+12)(4L+1))) \cdot ((6L+3)(5L+2)/(L-1), 15L+6, 21L)$
$\boldsymbol{\varepsilon}_3$	$((40L^2-41L+15)/(6L-6), 5L+2, 7L)/(9L-2)$
$\boldsymbol{\varepsilon}_4$	$((2L-1)/(3L-3), 1, 1)$
$\boldsymbol{\varepsilon}_5$	$((12L-3)/(12L-12), 1, 0)$
$\boldsymbol{\varepsilon}_6$	$(0, 0, 0)$
$\boldsymbol{\varepsilon}_7$	$(0, 5L+2, 7L)/(9L-2)$
$\boldsymbol{\varepsilon}_8$	$(0, 1, 1)$
$\boldsymbol{\varepsilon}_9$	$(0, 1, L/(L+1))$
$\boldsymbol{\varepsilon}_{10}$	$(4/5, 1, 0)$

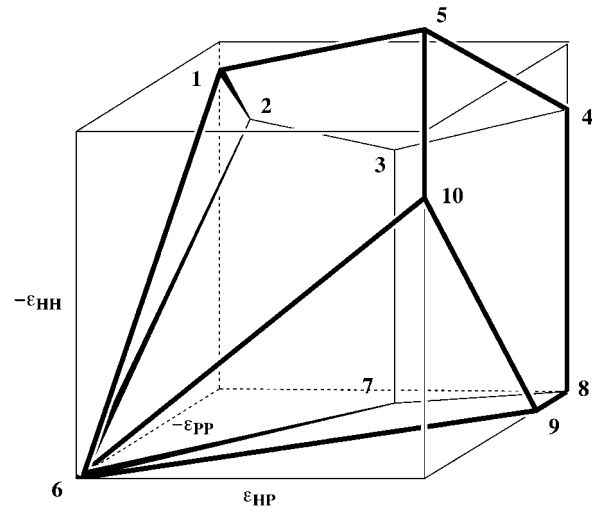

FIGURE 3 Illustration of the upper-bond estimation of the set of ''good parameters'' of the model, where the native conformation has the lowest conformational energy. It is assumed that $\varepsilon_\beta = -1$. The coordinates of the vertices 1–10 could be read from Table 3 for a given value of the barrel size parameter $L = 4$. See the text for details.

of the polyhedron E_U as shown in Fig. 3) and $0 < \delta < 1$. Suppose, for a moment, that the native structure is a unique, global minimum of E in $\boldsymbol{\varepsilon}_i(\delta)$, $i = 1,2,\ldots,10$. Therefore, by the fact that a convex combination of interaction parameters does not change the energy order of structures, the native is a unique, global minimum of E in the convex polyhedron $\mathsf{E}_U(\delta)$ defined by vertices $\boldsymbol{\varepsilon}_i(\delta)$, $i = 1,2,\ldots,10$. Obviously $\mathsf{E}_U(\delta) \subseteq \mathsf{E} \subseteq \mathsf{E}_U$ and $\mathsf{E}_U(\delta) \rightarrow \mathsf{E}_U$ if $\delta \rightarrow 1$.

Unfortunately, we are not able to prove $\mathsf{E}_U(\delta) \subseteq \mathsf{E}$ for some value(s) of δ. However, a credible estimation of the lower bound of the parameter space could be obtained from computer experiments. In many Monte Carlo simulations, we obtained the native structure as a unique, global minimum of E in large sets of structures visited during the simulations where Replica Exchange Monte Carlo was employed as a sampling scheme.

Thermodynamics of the model

Replica Exchange Monte Carlo sampling combined with the histogram method provided data for analysis of the thermodynamic properties of the model. Each computational experiment consisted of two parts. The first stage employed 16 replicas with 10^6 attempts to replica exchange (per replica) and 10^3 local moves (also per replica) between the exchanges. In the next stage we employed 3–5 replicas with 3×10^6 replica exchanges and 10^3 micromodifications between exchanges. The temperatures of particular replicas were linearly distributed around estimated (in preliminary simulations) transition temperature. A modified multihistogram method of Ferrenberg and Swendsen was employed for analysis of the system thermodynamics (Ferrenberg and Swendsen, 1988; Ferrenberg and Swendsen, 1989; Newman and Barkema, 1999).

TABLE 4 Selected points from E_U for our Monte Carlo simulations

Point	$-\varepsilon_{HH}$	ε_{HP}	$-\varepsilon_{PP}$	%
ε_0	0.57	0.70	0.48	31.1
ε_1	1.15	0.38	0.0	45.7
$\varepsilon_1(0.95)$	1.12	0.40	0.02	45.0
$\varepsilon_1(0.75)$	1.0	0.46	0.12	42.5
$\varepsilon_1(0.5)$	0.86	0.54	0.24	39.2
ε_2	0.95	0.32	0.40	46.9
ε_2 (0.95)	0.93	0.33	0.41	46.4
ε_2 (0.75)	0.85	0.41	0.42	43.6
ε_2 (0.5)	0.76	0.51	0.44	40.0
ε_3	0.80	0.65	0.82	45.3
ε_3 (0.95)	0.79	0.65	0.81	44.9
ε_3 (0.75)	0.74	0.66	0.74	42.4
ε_3 (0.5)	0.69	0.67	0.65	39.2
ε_4	0.78	1.0	1.0	43.3
ε_4 (0.95)	0.77	0.98	0.97	42.8
ε_4 (0.75)	0.73	0.92	0.87	40.8
ε_4 (0.5)	0.68	0.85	0.74	37.9
ε_5	1.25	1.0	0.0	40.8
ε_5 (0.95)	1.22	0.98	0.02	40.5
ε_5 (0.75)	1.08	0.92	0.12	38.7
ε_5 (0.5)	0.91	0.85	0.24	36.3
ε_6	0.0	0.0	0.0	0.0
ε_6 (0.95)	0.03	0.03	0.02	2.3
ε_6 (0.75)	0.14	0.17	0.12	10.1
ε_6 (0.5)	0.29	0.35	0.24	18.7
ε_7	0.0	0.65	0.82	12.5
ε_7 (0.95)	0.03	0.65	0.81	14.1
ε_7 (0.75)	0.14	0.66	0.74	18.1
ε_7 (0.5)	0.29	0.67	0.65	23.2
ε_8	0.0	1.0	1.0	8.7
ε_8 (0.95)	0.03	0.98	0.97	10.3
ε_8 (0.75)	0.14	0.92	0.87	15.5
ε_8 (0.5)	0.29	0.85	0.74	21.7
ε_9	0.0	1.0	0.80	0.0
ε_9 (0.95)	0.03	0.98	0.78	2.3
ε_9 (0.75)	0.14	0.92	0.72	10.1
ε_9 (0.5)	0.29	0.85	0.64	18.7
ε_{10}	0.80	1.0	0.0	23.4
ε_{10} (0.95)	0.79	0.98	0.02	23.9
ε_{10} (0.75)	0.74	0.92	0.12	25.4
ε_{10} (0.5)	0.69	0.85	0.24	27.6

It is assumed that $-\varepsilon_\beta = 1$, $L = 4$, and in the last column there is a percent of energy of long-range interactions.

The thermodynamics of the model system is analyzed in terms of the density of states; this enables us to define the distribution of states for the model system.

$$w(E') = \#\{\mathbf{s} \in \mathsf{S}\colon E(\mathbf{s}) = E'\} \tag{12}$$

and

$$p_T(E') = Z_T^{-1} w(E') \exp(-E'/k_B T), \tag{13}$$

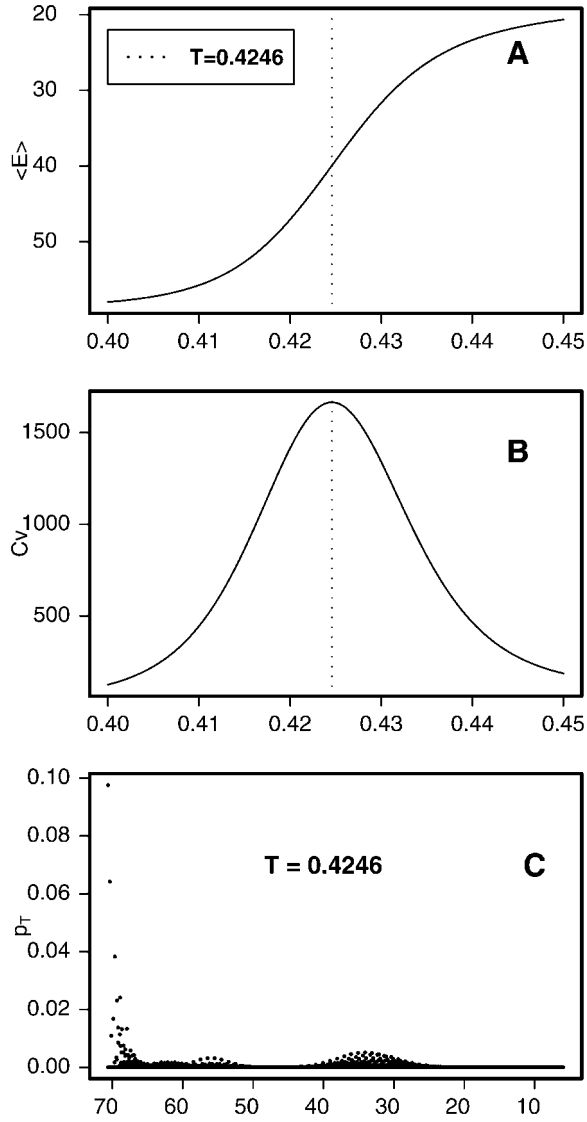

FIGURE 4 The thermodynamics of the model system for the interaction parameters corresponding to the center of gravity of the E_U set. The (*A*) and (*B*) panels show the conformational energy and heat capacity curves respectively. The vertical dotted line indicates the transition temperature (maximum of Cv) which is equal to $T = 0.4246$. The (*C*) panel shows the Boltzmann distribution of states.

where Z_T is the partition function $\Sigma_{E'}\, w(E')\exp(-E'/k_BT)$.

This allows the definition of entropy and free energy of the system to be:

$$S(E') = k_B\log(w(E')) \tag{14}$$

$$F_T(E') = E' - TS(E') = -k_BT(\log(p_T(E')) + \log(Z_T)). \tag{15}$$

At an infinite temperature the system energy can be estimated as:

$$\langle E\rangle_\infty = \Sigma_{E'}E'w(E')/\Sigma_{E'}w(E'). \tag{16}$$

This enables a definition of an equivalent of the system calorimetric enthalpy:

$$\Delta E_{cal} = \langle E\rangle_\infty - E_{native} \tag{17}$$

The ratio of van't Hoff and calorimetric enthalpy is a conventional way to measure the transition cooperativity:

$$\kappa = 2 \times T_{max}[k_BC_v(T_{max})]^{1/2}/\Delta E_{cal} \tag{18}$$

Cooperativity coefficient κ assumes value 1 for strictly two-state all-or-none folding transition. T_{max} is the temperature corresponding to the maximum $C_v(T_{max})$ of the heat capacity curve. The heat capacity is measured in a standard way from the fluctuations of the system conformational energy. This analysis follows the approach employed previously by Chan and co-workers (Chan, 2000; Kaya and Chan, 2000a,b).

Fig. 4 shows the plots of the average system energy (*A*) and the average heat capacity (*B*) as the functions of the dimensionless absolute temperature for the central point of the E_U set. These quantities were calculated via canonical averaging (with the free energy given in Eq. 15). The transition temperature $T = 0.4246$ is very well-defined by the maximum of the heat capacity at constant volume, *C*v, plot. A very narrow range of the system temperature indicates a very abrupt folding transition. Fig. 4 *C* shows the Boltzmann distribution of states (Eq. 3) at the transition temperature. Clearly, the highest density of states could be observed at the low-energy end of the spectrum and in the high-energy region. There is a gap in the intermediate energy range, suggesting a cooperative two-state transition. The free energy plot at the transition temperature for the central point of the E_U set is shown in more detail in Fig. 5. Several interesting features can be seen from this plot. First, due to the discrete character of the model, there is no single line; instead, for almost the entire range of system energies, the free energy can assume various scattered values. Interestingly, near the native state the free energy becomes very well defined and reaches a deep minimum at the native state. In spite of the scattered character of the plot, the free energy barrier between the low and the high energy states is well pronounced and provides the signature of an all-or-none, protein-like folding transition. The large symbols in the plot indicate the native conformation (star) and the selected competitive structures. Those similar to the native structure are marked by black symbols (structure M2 was not observed in this trajectory), and the open symbols indicate more exotic misfolds. Interestingly, these appear in the higher energy range; however, they are on the native side of the free energy barrier.

To find out how the properties of the model are dependent on the interaction parameters, we performed simulations for the various vectors of parameters including these as close as possible to the corners of the E_U set (see Table 4). Somewhat arbitrarily, we assumed that for a dependable estimation of

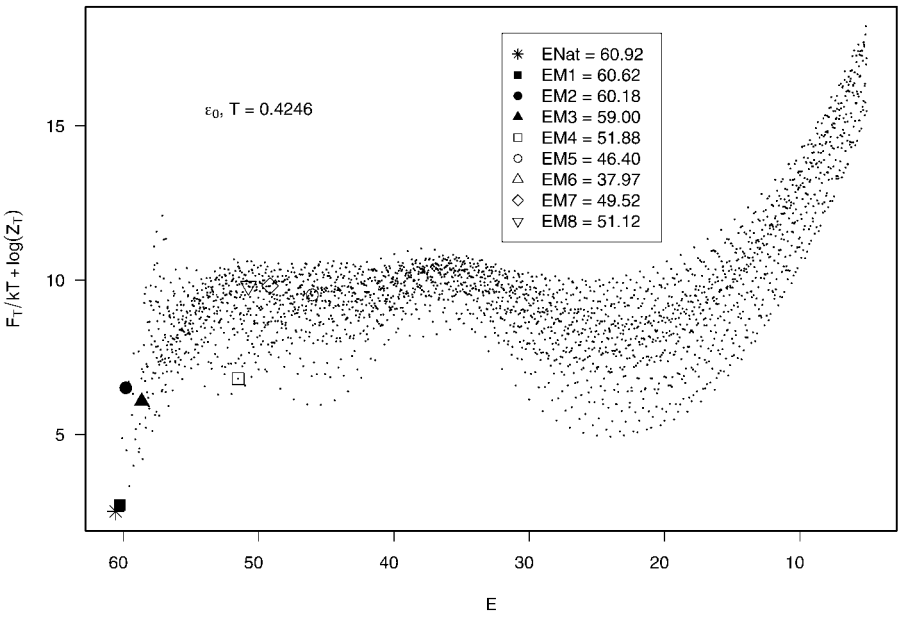

FIGURE 5 Reduced free energy as a function of energy for the center of gravity of the E_U set. The star symbol indicates the native state which corresponds to the minimum of the conformational energy and the minimum of free energy. The inset gives the values of energy for the native state and eight selected misfolded structures. Slightly misfolded structures (*black symbols*) have energy and free energy close to the native one. More distant structures are further away from the native structure parameters. The M2 misfold was not observed in this long trajectory, suggesting its high free energy, consistent with the rest of the plot.

the thermodynamic quantities, the native structure should appear in the first 20% of the simulation time. Thus, we analyze the results for only six (rather representative) sets of interactions parameters: $\boldsymbol{\varepsilon}_0$, $\boldsymbol{\varepsilon}_1$ (0.85), $\boldsymbol{\varepsilon}_2$ (0.95), $\boldsymbol{\varepsilon}_3$ (0.95), $\boldsymbol{\varepsilon}_4$ (0.95), $\boldsymbol{\varepsilon}_5$ (0.95), where $\boldsymbol{\varepsilon}_5$ (0.95) means that the values of the parameters correspond to the following $(0.05\boldsymbol{\varepsilon}_0 + 0.95\boldsymbol{\varepsilon}_5)$ combination of the vectors $\boldsymbol{\varepsilon}_0$ and $\boldsymbol{\varepsilon}_5$. These parameters are close to the corners of the E_U set, confirming that our selection of competitive structures led to a reasonable estimation of the range of good parameters of the model interaction scheme. Near the other vertices of the set the folding is slow and the native structure appears less frequently in the MC trajectories. For easy comparison the Boltzmann distribution was subject to a smoothing procedure. The averaged quantity is defined below:

$$p^*_T(E) = \Delta E_i^{-1} \, \Sigma_{E' \in \Delta Ei} \, p_T(E'). \tag{19}$$

where $\Delta E_i = 1.0$ is a small (however, containing a large number of states) energy interval. The results are compared in Fig. 6. The values of cooperativity parameters κ are included in all panels. First, it is easy to note that the stronger interactions between polar groups led to a wider gap in the distributions of states. Indeed, the clearly manifested cooperative folding transitions are for $\boldsymbol{\varepsilon}_3$ (0.95), and $\boldsymbol{\varepsilon}_4$ (0.95), where the $-\varepsilon_{PP}$ parameters describing the polar contacts have the largest values. It could also be noted that the all-or-none transition is well pronounced in these systems where the contribution from all types of long-range interactions is relatively large. The κ values given in Fig. 6 are obtained without empirical baseline subtractions. As it was demonstrated by Kaya and Chan (Kaya and Chan, 2000a,b) when κ value obtained without baseline subtraction approaches 0.7, the true van't Hoff calorimetric enthalpy ratio should be close to one. This applies to the examples given in Fig. 6, *D* and *E*. For these systems the transition is clearly very close to the ideal two-state folding. In these cases, the height of the free-energy barrier is in the range of 5–10 k_BT, which implies a negligible population of folding intermediates. As it was shown in the previous sections, some contribution from the short-range interactions is necessary for the uniqueness of the native state; however, the systems dominated by these short-range interactions are very poor folders. In such cases, the transition is slow and the energy gap (or the free energy barrier) is low.

The model studied here differs significantly from other minimal models (Kaya and Chan, 2000a,b, 2002; Jang et al., 2002) that exhibit a cooperative two-state folding transition. First, we employ only a four-letter code for the sequence of the model chain (the code describes a combination of secondary preferences and hydrophobicity of the chain segments). Second, the geometry of our model seems to be more realistic (but not more complicated) than the geometry of the cubic lattice. Moreover, in contrast to many models based on target-type potentials of long range interactions, the present model allows nonnative interactions. Very cooperative folding, close to an ideal two-state transition, could be achieved for properly balanced contributions of the short-range and long-range interactions. Earlier findings (Kaya and Chan, 2000a,b; Kaya and Chan, 2002) that a more complex than two-letter (or three-letter) code for chain sequence in the presence of repulsive interactions is necessary for a highly cooperative transition to a unique native state were confirmed in this work.

CONCLUSIONS

A very simple lattice model of globular proteins was studied

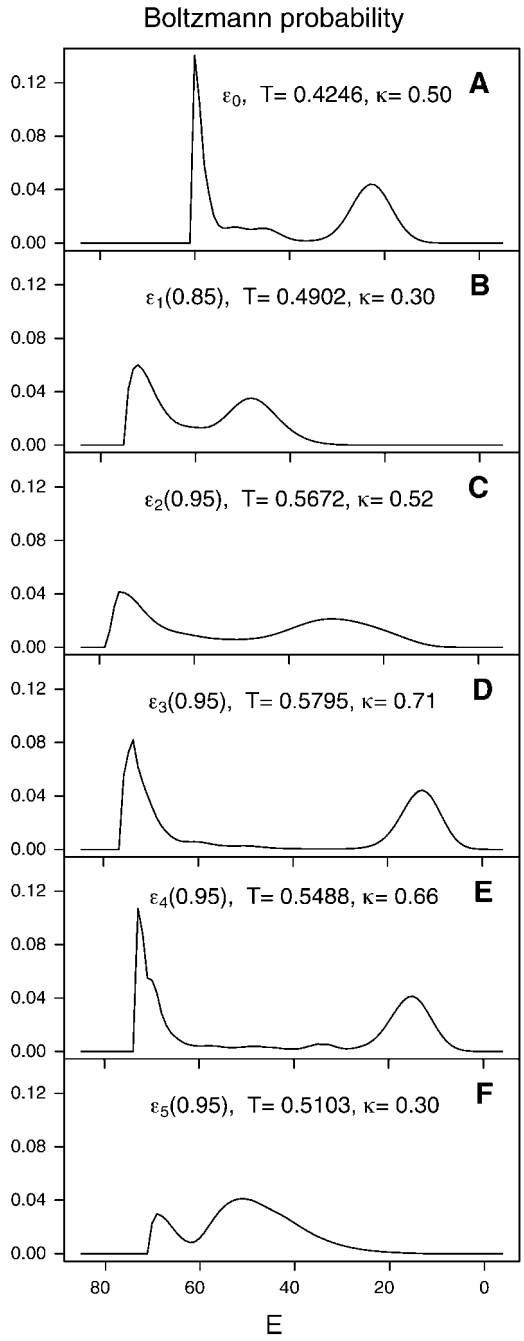

FIGURE 6 Averaged (see the text for details) Boltzmann distribution of states for representative sets of the parameters of the interaction scheme. T indicates the folding transition temperature and κ the cooperativity parameter. The values of interaction parameters for particular plots could be extracted from Table 4 assuming that ε_5 (0.95) means the $(0.05\varepsilon_0 + 0.95\varepsilon_5)$ combination.

both theoretically and computationally by means of the Replica Exchange Monte Carlo method combined with a multihistogram analysis (Newman and Barkema, 1999). The interaction scheme for the fcc lattice chain included short- and long-range interactions. The short-range interactions mimic a propensity to extended conformations, typical for β-type proteins. The long-range interactions are controlled by a pattern of polar and hydrophobic residues. The pairs of contacting hydrophobic residues decrease the system energy, whereas HP pairs are repulsive. The PP pairs are attractive, provided that the contacting chain fragments mimic the geometry typical of the parallel orientation of polar side chains in real proteins. Other types of PP contacts are ignored. The sequence of the model chain was designed to be consistent with a two-sheet, six-stranded antiparallel Greek-key β-barrel. Here, it was demonstrated that the proposed interaction scheme leads to a highly cooperative all-or-none (i.e., pseudo first order) transition to the expected folded state. Interestingly, the folded state is energetically degenerate; twenty slightly different geometrical realizations of the barrel have exactly the same contributions from various components of the force field (and consequently, the same conformational energy). In an approximate way, this mimics conformational mobility of the native structure of real proteins. Thus, this is probably the first simple lattice model that undergoes a protein-like discontinuous transition to the folded state that exhibits limited conformational degeneracy, a compact hydrophobic core, a protein-like fold topology and well-defined secondary structure. Moreover, it has been shown that the proposed interaction scheme is a minimal one. The system requires nonzero contributions of all potentials. The range of good parameters (that led to the above outlined protein-like behavior) was estimated both analytically and in Monte Carlo computational experiments.

The present work focused on β-type systems. Studies of minimal α-type and α/β-type model polypeptides are now in progress.

The assistance of Dr. Michal Boniecki in preparation of the figures is gratefully acknowledged.

This research was supported in part by the Division of General Medical Sciences of the National Institutes of Health (GM 37408). Piotr Pokarowski acknowledges partial support from the Polish Research Council KBN (7-T11F-016-21).

REFERENCES

Abkevich, V. I., A. M. Gutin, and E. I. Shakhnovich. 1996. Improved design of stable and fast-folding model proteins. *Fold. Des.* 1:221–230.

Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science*. 181:223–230.

Baker, D. 2000. A surprising simplicity to protein folding. *Nature*. 405: 39–42.

Branden, C., and J. Tooze. 1991. Introduction to Protein Structure. Garland Publishing, Inc., New York and London.

Bryngelson, J. D., J. N. Onuchic, N. D. Socci, and P. G. Wolynes. 1995. Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins*. 21:167–195.

Chan, H. S. 2000. Modeling protein density of states: additive hydrophobic effects are insufficient for calorimetric two-state cooperativity. *Proteins*. 40:543–571.

Dill, K. A., S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. 1995. Principles of protein folding — a perspective from simple exact models. *Prot. Sci.* 4:561–602.

Dinner, A. R., A. Sali, and M. Karplus. 1996. The folding mechanism of larger proteins: role of native structure. *Proc. Natl. Acad. Sci. USA.* 93:8356–8361.

Ferrenberg, A. M., and R. H. Swendsen. 1988. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.* 61:2635–2637.

Ferrenberg, A. M., and R. H. Swendsen. 1989. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* 63:1195–1198.

Gront, D., A. Kolinski, and J. Skolnick. 2000. Comparison of three Monte Carlo search strategies for a proteinlike homopolymer model: Folding thermodynamics and identification of low-energy structures. *J. Chem. Phys.* 113:5065–5071.

Gront, D., A. Kolinski, and J. Skolnick. 2001. A new combination of replica exchange Monte Carlo and histogram analysis for protein folding and thermodynamics. *J. Chem. Phys.* 115:1569–1574.

Hansmann, U. H. E. 1997. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* 281:140–150.

Hansmann, U. H. E., and Y. Okamoto. 1997. Numerical comparison of three recently proposed algorithms in the protein folding problem. *J. Comput. Chem.* 18:920–933.

Hansmann, U. H. E., and Y. Okamoto. 1999. New Monte Carlo algorithms for protein folding. *Curr. Opin. Struct. Biol.* 9:177–181.

Hukushima, K., and K. Nemoto. 1996. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jap.* 65:1604–1608.

Ilkowski, B., J. Skolnick, and A. Kolinski. 2000. Helix-coil and sheet-coil transitions in a simplified yet realistic protein model. *Macromol. Theory Simul.* 9:523–533.

Jackson, S. E. 1998. How do small single-domain protein fold? *Fold. Des.* 3:R81–R91.

Jang, H., C. K. Hall, and Y. Zhou. 2002. Folding thermodynamics of model four-strand antiparallel β-sheet proteins. *Biophys. J.* 82:646–659.

Karplus, M., and A. Sali. 1995. Theoretical studies of protein folding and unfolding. *Curr. Opin. Struct. Biol.* 5:58–73.

Kaya, H., and H. S. Chan. 2000a. Polymer principles of protein calorimetric two-state cooperativity. *Proteins.* 40:637–661.

Kaya, H., and H. S. Chan. 2000b. Energetic components of cooperative protein folding. *Phys. Rev. Lett.* 85:4823–4826.

Kaya, H., and H. S. Chan. 2002. Towards a consistent modeling protein thermodynamic and kinetic cooperativity: how applicable is the transition state picture to folding and unfolding? *J. Mol. Biol.* 315:899–909.

Kolinski, A., W. Galazka, and J. Skolnick. 1995. Computer design of idealized β-motifs. *J. Chem. Phys.* 103:10286–10297.

Kolinski, A., W. Galazka, and J. Skolnick. 1996. On the origin of the cooperativity of protein folding. Implications from model simulations. *Proteins.* 26:271–287.

Kolinski, A., and J. Skolnick. 1996. Lattice Models of Protein Folding, Dynamics and Thermodynamics. R. G. Landes, Austin.

Kolinski, A., J. Skolnick, and R. Yaris. 1986. The collapse transition of semiflexible polymers. A Monte Carlo simulation of a model system. *J. Chem. Phys.* 85:3585–3597.

Newman, M. E. J., and G. T. Barkema. 1999. Monte Carlo Methods in Statistical Physics. Clarendon Press, Oxford.

Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes. 1997. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.

Post, C. B., and B. H. Zimm. 1979. Internal condensation of single DNA molecule. *Biopolymers.* 18:1487–1501.

Ptitsyn, O. B. 1987. Protein folding: Hypotheses and experiments. *J. Protein Chem.* 6:273–293.

Scheraga, H. A., M.-H. Hao, and J. Kostrowicki. 1995. Theoretical studies of protein folding. *In* Methods in Protein Structure Analysis. M. Z. Atassi and E. Appela, editors. Plenum Press, New York.

Shakhnovich, E. I., and A. V. Finkelstein. 1989a. Theory of cooperative transitions in protein molecules. II. Phase diagram for a protein molecule in solution. *Biopolymers.* 26:1681–1694.

Shakhnovich, E. I., and A. V. Finkelstein. 1989b. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers.* 28:1667–1680.

Sugita, Y., and Y. Okamoto. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314:141–151.

Swendsen, R. H., and J. S. Wang. 1986. Replica Monte Carlo simulations of spin glasses. *Phys. Rev. Lett.* 57:2607–2609.