2007

Generation and enumeration of compact conformations on the two-dimensional triangular and three-dimensional fcc lattices

Myron Peto
Iowa State University

Taner Z. Sen
Iowa State University, taner@iastate.edu

Robert L. Jernigan
Iowa State University, jernigan@iastate.edu

Andrzej Kloczkowski
Iowa State University

Generation and enumeration of compact conformations on the two-dimensional triangular and three-dimensional fcc lattices

Abstract

We enumerated all compact conformations within simple geometries on the two-dimensional (2D) triangular and three-dimensional (3D) face centered cubic (fcc) lattice. These compact conformations correspond mathematically to Hamiltonian paths and Hamiltonian circuits and are frequently used as simple models of proteins. The shapes that were studied for the 2D triangular lattice included m×n" role="presentation" style="display: inline; line-height: normal; word-spacing: normal; word-wrap: normal; white-space: nowrap; float: none; direction: ltr; max-width: none; max-height: none; min-width: 0px; min-height: 0px; border: 0px; padding: 0px; margin: 0px; position: relative;">m×nm×n parallelograms, regular equilateral triangles, and various hexagons. On the 3D fcc lattice we generated conformations for a limited class of skewed parallelepipeds. Symmetries of the shape were exploited to reduce the number of conformations. We compared surface to volume ratios against protein length for compact conformations on the 3D cubic lattice and for a selected set of real proteins. We also show preliminary work in extending the transfer matrix method, previously developed by us for the 2D square and the 3D cubic lattices, to the 2D triangular lattice. The transfer matrix method offers a superior way of generating all conformations within a given geometry on a lattice by completely avoiding attrition and reducing this highly complicated geometrical problem to a simple algebraic problem of matrix multiplication.

Disciplines

Biochemistry, Biophysics, and Structural Biology | Bioinformatics | Computational Biology | Molecular Biology

Comments

This article is published as Peto, Myron, Taner Z. Sen, Robert L. Jernigan, and Andrzej Kloczkowski. "Generation and enumeration of compact conformations on the two-dimensional triangular and three-dimensional fcc lattices." The Journal of chemical physics 127, no. 4 (2007):044101. doi: 10.1063/1.2751169. Posted with permission.

Generation and enumeration of compact conformations on the two-dimensional triangular and three-dimensional fcc lattices

Myron Peto, Taner Z. Sen, Robert L. Jernigan, and Andrzej Kloczkowski

Citation: The Journal of Chemical Physics **127**, 044101 (2007);
View online: https://doi.org/10.1063/1.2751169
View Table of Contents: http://aip.scitation.org/toc/jcp/127/4
Published by the American Institute of Physics

Generation and enumeration of compact conformations on the two-dimensional triangular and three-dimensional fcc lattices

Myron Peto
Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011-3020

Taner Z. Sen, Robert L. Jernigan, and Andrzej Kloczkowski
Laurence H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa 50011-3020 and Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011-3020

(Received 30 April 2007; accepted 29 May 2007; published online 23 July 2007)

We enumerated all compact conformations within simple geometries on the two-dimensional (2D) triangular and three-dimensional (3D) face centered cubic (fcc) lattice. These compact conformations correspond mathematically to Hamiltonian paths and Hamiltonian circuits and are frequently used as simple models of proteins. The shapes that were studied for the 2D triangular lattice included $m \times n$ parallelograms, regular equilateral triangles, and various hexagons. On the 3D fcc lattice we generated conformations for a limited class of skewed parallelepipeds. Symmetries of the shape were exploited to reduce the number of conformations. We compared surface to volume ratios against protein length for compact conformations on the 3D cubic lattice and for a selected set of real proteins. We also show preliminary work in extending the transfer matrix method, previously developed by us for the 2D square and the 3D cubic lattices, to the 2D triangular lattice. The transfer matrix method offers a superior way of generating all conformations within a given geometry on a lattice by completely avoiding attrition and reducing this highly complicated geometrical problem to a simple algebraic problem of matrix multiplication. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2751169]

INTRODUCTION

Despite recent advances in computational biology, reduced models of proteins still enjoy considerable interest and applicability for studying protein structure, function, and dynamics. Globular proteins have compact structures with very tight packing of amino acids inside protein cores due in large part to the segregation between hydrophobic and polar residues. Additionally amino acids in proteins are covalently bonded forming relatively long sequences, containing on average between several tens to several hundreds of residues. The simplest mathematical models that mimic the linear nature of the protein sequence, its tight packing in the native state, and the excluded volume effect are compact self-avoiding walks on simple lattices of finite size. A compact self-avoiding walk requires each of the lattice points to be visited once and only once with no voids. Mathematically such walks are named Hamiltonian paths (an alternative nomenclature sometimes used in the literature is Hamilton paths). For regular (noncompact) self-avoiding walks, some points on the lattice may be left unvisited, creating voids. A compact self-avoiding walk that begins and ends at the same site is called a Hamiltonian circuit. In lattice models of proteins each residue is usually represented by a single lattice node. Much work has been done in the past with such models to represent collapsed polymers and proteins.[1–12]

Native conformations of globular proteins are compact and unique. The essence of comprehending protein folding is to determine, for a given sequence of amino acids, the most energetically favorable conformation. This is an extremely difficult computational problem, since the number of possible conformations grows geometrically with the length of the chain. Random search methods frequently fail to identify this single unique structure; whereas complete enumerations, whenever feasible, are better suited and preferable for this task.

The complete enumerations of compact conformations (for both paths and circuits) within rectangles of varying sizes $n \times m$ on the square lattice in two dimensions (2D) and parallelepipeds of the size $l \times n \times m$ on the cubic lattice in three dimensions (3D) have been studied by us and other authors. A major obstacle for such computations on longer chains is attrition as it becomes more and more difficult to locate unoccupied neighboring sites for the continuation of the walk. To overcome this problem we previously developed the transfer matrix method[13–17] to grow the chain not in the traditional linear way but in a piecewise way cross section by cross section, in order to avoid attrition. That approach enabled us to compute all possible Hamiltonian walks and Hamiltonian circuits within rectangles of varying sizes on the square lattice and parallelepipeds on the cubic lattice.

The aim of the present work is the extension of these results to other popular lattices. The triangular lattice in 2D and the face centered cubic (fcc) lattice in 3D are especially well suited for the modeling of proteins. The coordination numbers z for these lattices are 6 and 12, respectively, and because of this the protein conformations generated on such

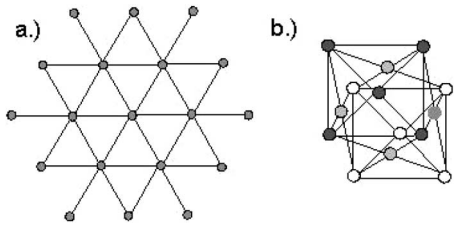

FIG. 1. The 2D triangular lattice (a) and a unit cell in the face centered cubic lattice (b).

lattices are more realistic than for the square ($z=4$) and the cubic ($z=6$) lattices. It is well known that the packing of residues inside globular proteins best fits the fcc lattice among all other lattices.[18] Additionally the distribution of angles between vectors connecting centers of side chains of spatially neighboring residues is best fitted to the 12 directional vectors of the fcc lattice.[18]

For various simple geometric shapes on the 2D triangular and 3D fcc lattice we enumerate all possible compact self-avoiding walks and circuits. Figure 1 shows examples of geometries studied for the triangular [Fig. 1(a)] and the fcc [Fig. 1(b)] lattice.

Figure 2 shows an example of a Hamiltonian circuit [Fig. 2(a)] and a Hamiltonian path [Fig. 2(b)] on the 2D triangular lattice. We enumerate all possible Hamiltonian walks and circuits within several simple geometries such as $n\times m$ parallelograms, equilateral triangles, and several classes of hexagons of varying sizes (Fig. 3). For the 3D fcc lattice we enumerate all possible walks and circuits within a limited class of skewed parallelepipeds (Fig. 4).

We take advantage of the fact that the shapes studied here exhibit symmetries. By excluding paths related by the symmetry of the shape we reduce the computer time necessary for generation. A similar approach was used earlier by us for the generation of compact conformations on the square and the cubic lattices. In the case of the 2D square and 3D cubic lattices other reductions are possible based on parity considerations. This is related to the chessboardlike nature of these lattices that can be exploited to reduce the total computational time required for the generation of compact conformations. Such a reduction is not possible for the presently studied triangular and fcc lattices. For the square and the cubic lattice, any two nodes of a given path (that represents a polypeptide chain) that are lattice neighbors must be separated by an even number of nodes along the path. Because of this it is impossible to have a Hamiltonian circuit composed of an odd number of nodes. Indeed, one aim of our studies on the triangular and the fcc lattices is to utilize protein lattice representations that do not have such parity restrictions.

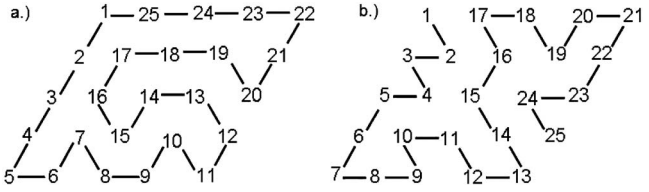

FIG. 2. Examples of a Hamiltonian circuit (a) and a Hamiltonian path (b) within a parallelogram of size 5×5 on the 2D triangular lattice.

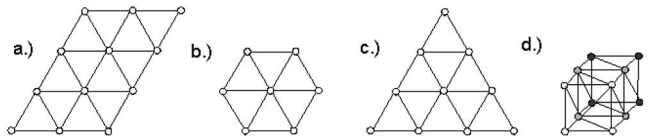

FIG. 3. Examples of protein shapes on the triangular and the fcc lattices studied in the present work. (a) a 3×4 parallelogram, (b) a regular (equilateral) hexagon having side lengths of 1 (in lattice units), (c) a regular (equilateral) triangle with sides of length 3, and (d) a $2\times2\times3$ skewed parallelepiped on the fcc lattice.

In addition, the fcc lattice closely approximates the dihedral angles of real proteins.[18] The fcc lattice allows for the densest packing of hard spheres and thus the dihedral angles in densely packed proteins can be associated with the fcc geometry.

The standard method of enumerating walks, which we employ here, uses a naïve treelike growth algorithm. Paths are generated by adding one bond (step) in each possible way at a time and checking for possible overlaps that are not allowed. The procedure is continued until every node of the graph is visited (i.e., a Hamiltonian path is completed) or until a dead end is reached, at which point the algorithm backs up to a node where a different path (along the branches of a tree starting from that node) might be possible. This is relatively simple to program but, especially for graphs of increasing size, suffers from the serious problem of attrition. As the number of nodes increases, fewer and fewer steps in the path generation will eventually lead to a completed path. In previous work we have developed the transfer matrix method for generating all Hamiltonian paths and Hamiltonian circuits within rectangles on the 2D square lattice and parallelepipeds on the 3D cubic lattice. Here we extend that work to the 2D triangular lattice, suggesting that it may also be possible to develop this method for the fcc lattice and other simple lattices.

Symmetries

We exploit symmetries of shapes in order to reduce the computational costs of conformation generation. The total numbers of symmetries for all of the shapes studied in the present work are given in Table I. For the 2D triangular lattice the regular (equilateral) hexagon has the most symmetries (12), the regular (equilateral) triangle has 6 symmetries,

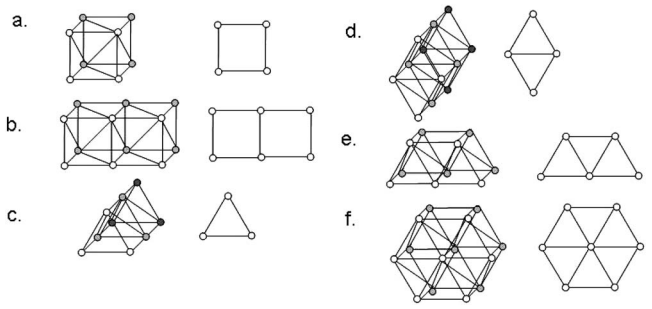

FIG. 4. Shapes embedded in the 3D fcc lattice. (a) $2\times2\times n$, (b) $2\times3\times n$, (c) 1 tri$\times n$, (d) 2 tri$\times n$, (e) 3 tri$\times n$, and (f) hex$\times n$. The symmetries of each are referred to in the preceding table. Cross sections are shown on the right side.

TABLE I. Symmetries for several classes of shapes on the 2D triangular (top) and 3D face centered cubic (fcc) (bottom) lattices.

	Type of symmetry				
	Regular hexagon	Regular triangle	$n\times n$ parallelogram	$n\times m(n\neq m)$ parallelogram	Near-regular hexagon
Identity	1	1	1	1	1
±60° rotation	2	0	0	0	0
±120° rotation	2	2	0	0	0
180° rotation	1	0	1	1	1
Reflection axial	3	0	0	0	1
Reflection diagonal	3	3	2	0	1
Total	12	6	4	2	4

	Type of symmetry					
	Skewed $2\times2\times n$ parallelepiped	Skewed $2\times3\times n$ parallelepiped	Skewed 1 tri$\times n$ parallelepiped	Skewed 2 tri$\times n$ parallelepiped	Skewed 3 tri$\times n$ parallelepiped	Skewed hex$\times n$ parallelepiped
Identity	1	1	1	1	1	1
180° rotation-facial axis	1	0	0	1	0	1
Reflection in an axial plane	1	0	1	1	1	1
Inversion	1	1	0	1	0	1
Total	4	2	2	4	2	4

and a parallelogram with 4 equal sides (rhomb), as well as a near equilateral hexagon with 4 sides of equal length and 2 other sides of equal length both have 4 symmetries. In three dimensions we enumerated conformations within certain classes of skewed parallelepipeds and, depending on the class, there can be either two or four symmetries. The use of symmetries reduces the total numbers of paths by a constant factor σ corresponding to the number of symmetries of the shape. Thus, if there are N_{tot} paths without eliminating paths related by symmetry, then $N=N_{\mathrm{tot}}/\sigma$ is the number of paths after removing paths related by symmetry.

We note that, in addition to symmetries of the shape, there are also symmetries of the sequence if the graph representing the sequence is undirected. Real proteins correspond to directed graphs because of the distinction between their N terminal and the C terminal. It is useful, however, to consider undirected Hamiltonian walks on the lattice. A conformation exhibits head-tail symmetry if starting at either end of the conformation produces the same undirected graph. Figure 5(a) shows an example of two conformations on the triangular lattice related by the head-tail symmetry. If the number of graphs with head-tail symmetry is N_s, then the total number of distinct directed graphs N is related to the number of distinct undirected graphs N_u by $N_u=(N+N_s)/2$.

We use the same method outlined in our previous work to remove symmetries. Specifically, we fix the first few steps of a path until the symmetry of the shape is broken. For example, when starting from the middle node of a regular equilateral hexagon we only need to enumerate paths with the fixed direction of the first step, since five other directions are equivalent [see Fig. 5(b)]. In addition, the second step is also fixed to break the symmetry of the shape.

Extension of the transfer matrix method to the triangular lattice

The self-avoiding walks allow the generation and enumeration of all possible compact conformations; however, due to the attrition mentioned above, the time required for these computations grows geometrically with the length of the chain. Attrition arises from the excluded volume condition together with the requirement of complete occupancy. Because of this (if the chain is grown in a traditional linear way) it becomes more and more difficult to find an unoccupied neighboring site for the subsequent step of the walk. The traditional linear chain growth method is therefore not the most efficient method for growing a chain for a compact dense system. A better approach for growing a chain is one that utilizes a piecewise method to grow it cross section by cross section, using a transfer matrix method.[13–17] This method was first proposed in 1984 by Schmaltz *et al.*[17] for enumerations of Hamiltonian circuits within rectangles in 2D

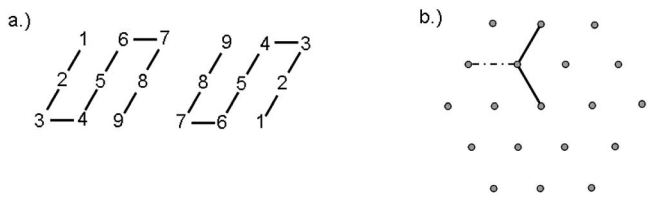

FIG. 5. Dealing with symmetric conformations. (a) An example of two conformations exhibiting head-tail symmetry. The two structures are equivalent upon rotation by 180° in the plane. Shown in (b) is the method we use to eliminate symmetries. If we start our path from the central node then only one of the six equivalent nodes is chosen as the first step and only one of the two equivalent nodes as the second step (the first two steps are shown as dark lines). The other step, shown as a broken line, would produce conformations symmetrical to the first one.

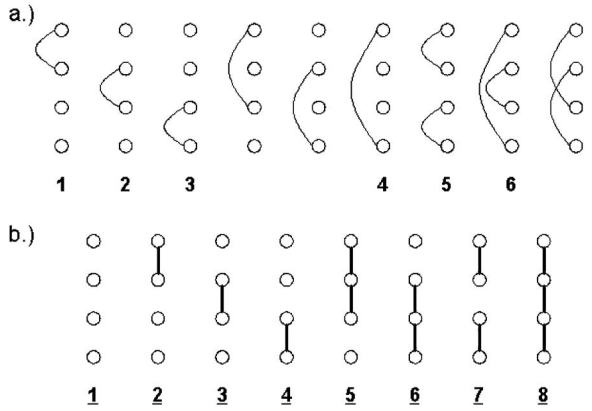

FIG. 6. All possible connectivity states (only physically acceptable states are labeled from 1 to 6) (a) and bond distributions (b) for generation of Hamiltonian circuits within rectangles of size $4\times n$.

on the square lattice. The Hamiltonian circuit [Fig. 2(a)] is defined as a walk through all available lattice points, subject to the conditions that each site can be visited only once and that we return in the last step back to the starting point.

The regular Hamiltonian path [Fig. 2(b)] does not need to satisfy the second condition, and the walk (chain) has two ends. In the past work we have extended this transfer matrix method to Hamiltonian circuits in three dimensions on the cubic lattice and to Hamiltonian paths (chains), both in two dimensions on the square lattice[16] and in three dimensions on the cubic lattice.[15] To briefly illustrate this method let us consider the enumeration of Hamiltonian circuits on a square lattice constrained to the $m\times n$ rectangular strip of width $m=4$ and variable length n. Figure 5(a) defines all possible external connectivities to one side of the four points on a line. Figure 5(b) shows all possible distributions of bonds among the four points on a line, including the case with no bonds (No. 1 where all bonds would be to the neighboring lines). We note that intersecting connectivities such as the last one in Fig. 5(a) are not allowed. Additionally two unlabeled connectivities in the middle of Fig. 5(a) are not allowed due to the parity reasons, so that the total number of the possible connectivity states is only 6 in this simple example.

The transfer matrix $\mathbf{T}$ is constructed by combining all connectivity states [Fig. 6(a)] with all bond distributions [Fig. 6(b)] and finding the resulting connectivity states formed by their combinations. The combinations, which lead to unoccupied sites, triple connections, or the formation of small loops, are not allowed. The element T_{ij} of the transfer matrix is zero if there is no possible transition from connectivity state i to state j. If there are possible transitions from state i to state j, then T_{ij} indicates the number of different ways to realize this transition. (For Hamiltonian circuits on the square lattice the elements T_{ij} of the matrix $\mathbf{T}$ are either 0 or 1, but in general T_{ij} can be larger than 1.) We construct the vector $\mathbf{u}$ of the starting states with elements u_i, for each connectivity state i [such as in Fig. 6(a)], as the first state on the left in the process of building a circuit (we use a left to right convention). The number u_i identifies the number of different ways in which this may be realized. As starting states, we use the distributions of bonds [such as in Fig. 6(b)] that do not contain any unoccupied sites [Nos. 7 and 8 in Fig. 6(b)]. We then determine the connectivity state to which the given distribution of vertical bonds transforms if (1) the horizontal bonds connecting to vertical bonds in the neighboring column on the right side are added and (2) a vertical cross section passing through these newly added horizontal bonds is taken. [The distribution of bond No. 7 in Fig. 6(b) leads to the connectivity state No. 5 in Fig. 6(a), while the distribution No. 8 leads to the connectivity state No. 4.] We also construct the vector $\mathbf{v}$ of the ending states with elements v_i determining if a given connectivity state i may form a closed circuit by combining it with the distribution of vertical bonds. The exact counting of the number N_c of all possible Hamiltonian circuits on the rectangle of size $m\times n$ on the square lattice is then given by the simple formula

$$N_c = \mathbf{u}^T(\mathbf{T})^{n-2}\mathbf{v},$$

with the superscript T denoting the transpose of vector $\mathbf{u}$. If we do not include the impossible unlabeled states in Fig. 1(a) and number physically acceptable states from 1 to 6 then the transfer matrix $\mathbf{T}$, the vectors of the starting states $\mathbf{u}$ and the ending states $\mathbf{v}$, are

$$T=\begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \quad u=\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad v=\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

In our previous work we have extended the method to Hamiltonian chains (with two ends) in two dimensions on the square lattice by generalizing the definition of the connectivity state to include the connectivities with up to two ends and by generalizing bond distributions by including up to two ends.[16] We have also generalized the transfer matrix method to 3D on the cubic lattice both for Hamiltonian circuits and Hamiltonian paths.[15] In 2D, the cross sections used for the generation and enumerations of Hamiltonian paths (or circuits) were lines. In 3D the cross sections are planes. We have written computer programs that automatically calculate the transfer matrices for paths and circuits in 2D and 3D. The only limitation is the computer memory associated with the size of the transfer matrix. The program was used to calculate transfer matrices as large as 3104×3104.

The transfer matrix method for generating and enumerating compact conformations is extremely efficient. The main advantage is that the piecewise generation of conformations is attrition-free. Once the transfer matrix for a given cross section is defined, the more complicated geometrical problem of conformation generation (or calculation of averages such as average energy) becomes a simple problem of matrix algebra that can easily be performed even for extremely long rectangles (parallelepipeds). The main difficulty of this method lies in the rapidly growing number of connectivity states for the increasing size of the cross section, but the development of the transition matrices will be automated

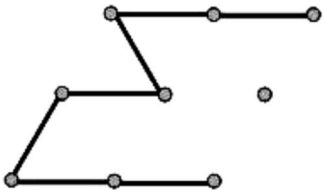

FIG. 7. The extension of the transfer matrix method to the triangular lattice must take into account nodes (such as the central one in the figure) that are already occupied during the process of piecewise building of the chain. The consideration of such nodes on the triangular lattice leads to an extension of the definition of connectivity states compared to the square lattice.

in order to access larger structures. Because calculations of transfer matrices are generated with a computer program, we are only limited by storage of large matrices. The algebraic formulation of the highly complicated compact self-avoiding walk problem is the principal beauty and power of this method. In the present paper we will extend the transfer matrix method to the triangular lattice.

The extension of the transfer matrix method to the triangular lattice

The triangular lattice is more difficult for studies of self-avoiding walks because its coordination number $z=6$ is larger than the coordination number $z=4$ of the square lattice. Because of this the number of possible Hamiltonian walks and Hamiltonian circuits on the triangular lattice grows much faster with the length of the chain than for the square lattice. Additionally, for the square lattice, the parity effect associated with its chessboardlike nature (two sites that form a contact must be separated by an even number of other sites in the path) substantially reduces the number of possible connectivity states. The triangular lattice does not have this feature and all possible connectivities must be included.

We will consider Hamiltonian circuits and Hamiltonian walks within parallelograms of various sizes on the triangular lattice, such as the 5×5 parallelogram shown in Fig. 2. We will concentrate in detail on the simplest case of Hamiltonian circuits within the parallelogram of the size $3\times n$ that will enable us to better understand the proposed transfer matrix method.

The connectivity states for the triangular lattice are defined similarly as for the square lattice by taking the cross section along the skewed column and figuring out how various pieces of the chain are connected on the left side of this cross section. The only generalization of this approach relative to the square lattice is that additionally to regular connectivity states similar to these shown in Fig. 6(a) for the square lattice we need to consider situations such as that shown in Fig. 7 where in the second skewed column the upper and the lower sites are connected but, additionally, the site in the middle of the second column has already been occupied and therefore must be excluded in the process of choosing the transition to the connectivity states in the next cross section. We use the symbol of an x to denote these excluded sites in the generalized connectivity states. Figure 8 shows all possible connectivity states for the $3\times n$ parallelograms on the triangular lattice. We note that the connectivity state containing the excluded site at the top of the skewed column is not possible.

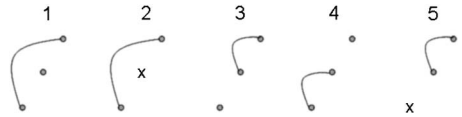

FIG. 8. All possible connectivity states for Hamiltonian circuits on $3\times n$ parallelograms on the triangular lattice. The cross symbol denotes connectivity states containing excluded nodes, such as the central node in Fig. 7.

The idea of a bond distribution defined on a line cross section for the 2D square lattice must now be generalized to include all bonds within a cross section of length 2. A bond distribution may contain bonds in the first (counting from the left) skewed vertical column of such cross section, and bonds between lattice nodes in the first and the second skewed vertical column, but cannot contain bonds in the second skewed vertical column. Additionally bond distributions that have nodes connected with more than two other nodes (triple or higher order connections) or have less than two bonds joining two skewed vertical columns are not allowed. The bond distributions cannot have an isolated (nonbonded) node at the top of the first skewed column. (Such isolated nodes are, however, allowed for other positions in this column since they can be matched with connectivity states containing excluded nodes.) Additionally the first skewed vertical column should have at least two nodes linkable from the left and closed loops formed by bonds in the two-column cross section are forbidden. Finally, bond distributions forming isolated excluded nodes in the second skewed column are not allowed. Figure 9(a) shows all possible bond distributions defined on a cross section of length 2 that satisfy these conditions. The bond distributions in Fig. 9(a) are grouped according to connectivity states (indicated by Nos. 1–5 and defined in Fig. 8) obtained by superposition of a given bond

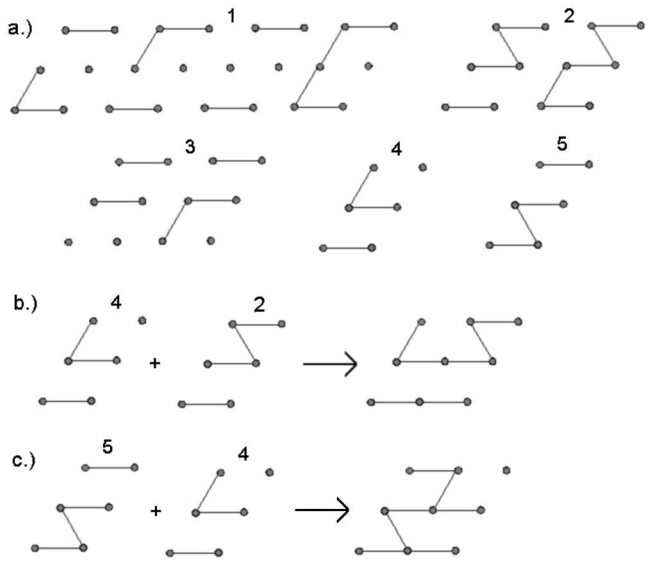

FIG. 9. Bond distributions (a) for each of the five connectivity states. All other distributions will not lead to valid conformations. (b) shows an example of a valid transfer from one state to another while (c) shows an invalid transfer from state 5 to state 4 because of a triple-bonded node.

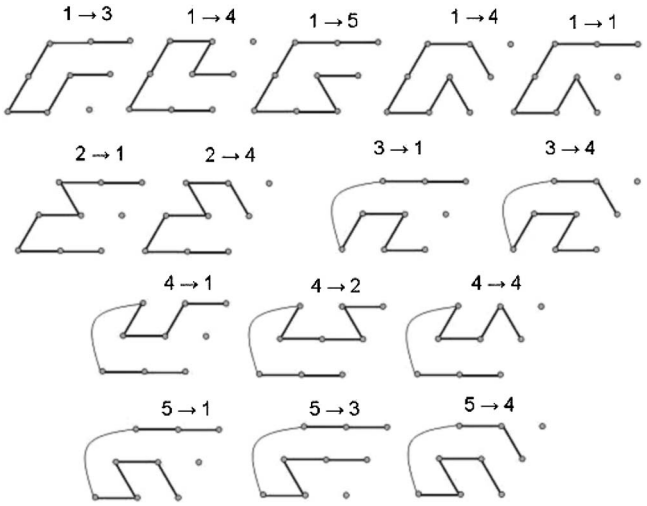

FIG. 10. All possible transitions between various connectivity states for Hamiltonian circuits on $3\times n$ parallelograms on the triangular lattice. The notation $1\rightarrow 3$ means the transition from connectivity state 1 (in Fig. 8) to connectivity state 3.

distribution with a preceding connectivity state. Figures 9(b) and 9(c) show examples of valid and invalid transfers between two states, respectively.

Figure 10 illustrates all possible transitions between connectivity states for the Hamiltonian circuits on $3\times n$ parallelograms on the triangular lattice.

Because of this the transfer matrix has the following form:

$$\boldsymbol{T}=\begin{bmatrix}1&0&1&2&1\\1&0&0&1&0\\1&0&0&1&0\\1&1&0&1&0\\1&0&1&1&0\end{bmatrix}.$$

The element T_{14} of the matrix has the value 2 because there are two different ways to transfer from the connectivity state 1 to the connectivity state 4 in the next cross section by using two completely different bond distributions, as shown in Fig. 9.

The vector of starting states is obtained by considering all possible distributions of bonds in the first skewed column on the left and horizontal bonds joining the first column with the second one that leaves no voids in the first column and figuring out the resulting connectivity state in the second column. Figure 11(a) illustrates all these possibilities for Hamiltonian circuits on $3\times n$ parallelograms on the triangular lattice.

The vector $\mathbf{u}$ of the starting states is therefore

$$\boldsymbol{u}=\begin{bmatrix}1\\1\\0\\1\\0\end{bmatrix}.$$

The ending connectivity states are those that lead to the closing of the circuit. Figure 11(b) illustrates all these ending connectivity states for $3\times n$ parallelograms.

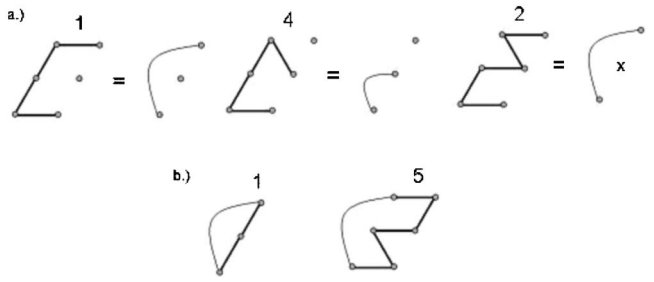

FIG. 11. Connectivity states that are the starting states (a) or the ending states (b) for Hamiltonian circuits on the $3\times n$ parallelograms on the triangular lattice.

The vector $\mathbf{v}$ of ending states that follows from Fig. 11(b) is

$$\boldsymbol{v}=\begin{bmatrix}1\\0\\0\\0\\1\end{bmatrix}.$$

The number of Hamiltonian circuits for the parallelogram of length n is then obtained from the equation $N_c=\mathbf{u}^T(T)^{n-2}\mathbf{v}$. Table II shows the computed numbers of Hamiltonian circuits (N_n) within parallelograms of size $3\times n$ on the triangular lattice.

The last column in Table II shows the ratio of the number of Hamiltonian circuits N_n/N_{n-1} for parallelograms differing in size by one column. Table II shows that this ratio converges rapidly with increasing size of the system.

TABLE II. Numbers of Hamiltonian circuits (N_n) within parallelograms of size $3\times n$ on the triangular lattice and the ratio N_n/N_{n-1}.

n	N_n	N_n/N_{n-1}
1	1	1.000 000
2	4	4.000 000
3	13	3.250 000
4	44	3.384 615
5	148	3.363 636
6	498	3.364 865
7	1 676	3.365 462
8	5 640	3.365 155
9	18 980	3.365 248
10	63 872	3.365 227
11	214 944	3.365 231
12	723 336	3.365 230
13	2 434 192	3.365 230
14	8 191 616	3.365 230
15	27 566 672	3.365 230
16	92 768 192	3.365 230
17	312 186 304	3.365 230
18	1 050 578 720	3.365 230
19	3 535 439 040	3.365 230

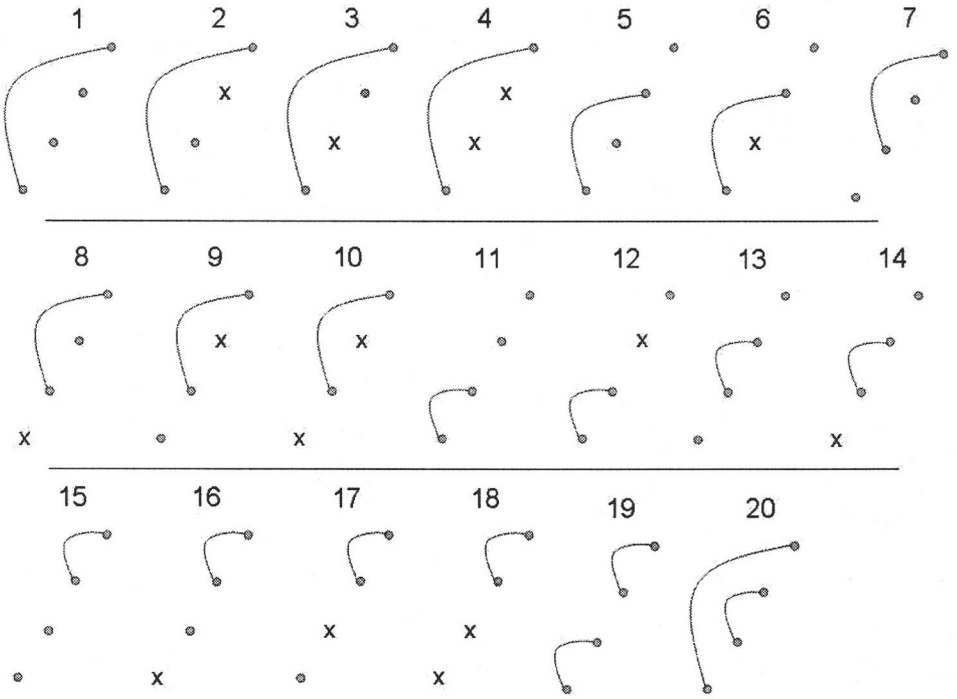

FIG. 12. All possible connectivity states for Hamiltonian circuits on $4\times n$ parallelograms on the triangular lattice.

The number of possible connectivity states for the triangular lattice grows much faster with the width of the parallelogram than for the square lattice. For example, for Hamiltonian circuits on $4\times n$ parallelograms there are 20 possible connectivity states shown in Fig. 12, while for the square lattice of the same cross-section size there are only 6 states.

The starting states for this case are shown in Fig. 13.

Figure 14 shows all ending states.

We are in the process of writing the computer code that will automatically generate transfer matrices and starting and ending vectors of states for varying sizes of parallelograms for both Hamiltonian circuits and Hamiltonian paths on the triangular lattice. In the future we will generalize this method to skewed parallelepipeds on the fcc lattice.

RESULTS

We enumerated the total number of conformations for numerous geometries such as parallelograms, regular (equilateral) triangles, and regular and semiregular (rhomblike) hexagons on the 2D triangular lattice and for several classes of parallelepipeds of the 3D fcc lattice. The total numbers of possible conformations for the various geometries in 2D and 3D are shown in Table III. Note that the computed numbers of Hamiltonian circuits in the case of $3\times n$ parallelograms match exactly with our results obtained with the transfer matrix method.

The first column in each table, n, corresponds to the length of the lattice shape. In the case of the regular triangle

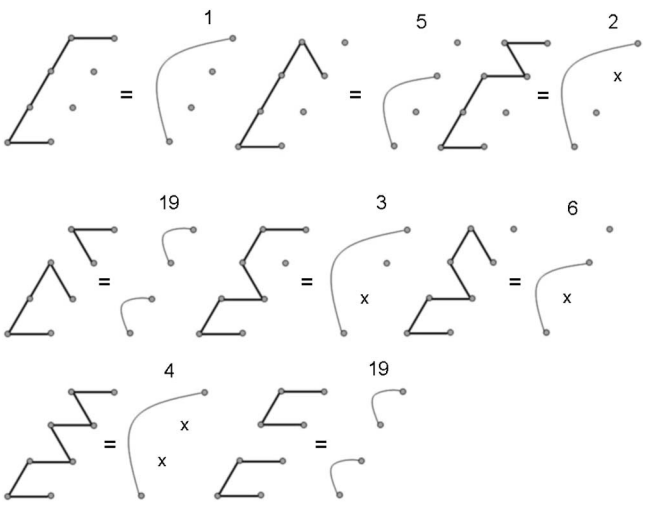

FIG. 13. Connectivity states that are starting states for Hamiltonian circuits on the $4\times n$ parallelograms on the triangular lattice.

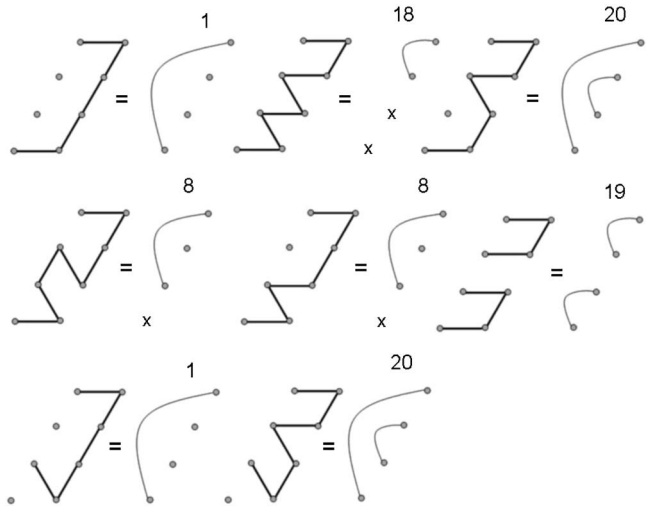

FIG. 14. Connectivity states that are ending states for Hamiltonian circuits on the $4\times n$ parallelograms on the triangular lattice.

TABLE III. Enumerations of all paths and circuits for various geometries for the 2D triangular lattice and the 3D fcc lattice.

(a) $2\times n$ parallelograms

n	N	C_n
2	3	1
3	17	1
4	44	1
5	104	1
6	235	1
7	519	1
8	1131	1
9	2448	1
10	5279	1

(b) $3\times n$ parallelograms

n	N	C_n
2	17	1
3	46	4
4	509	13
5	2 525	44
6	11 731	148
7	52 282	498
8	225 105	1 676
9	943 773	5 640
10	3 873 553	18 980

(c) $4\times n$ parallelograms

n	N	C_n
2	44	1
3	509	13
4	2 984	80
5	63 486	549
6	632 663	3 851
7	6 012 755	26 499
8	55 267 216	183 521
9	494 183 548	2 539 368

(d) $5\times n$ parallelograms

n	N	C_n
2	104	1
3	2 525	44
4	63 486	549
5	704 218	7 104
6	29 534 833	208 200
7	588 668 783	2 950 572

(e) Regular triangle, n,n,n

n	N	C_n
2	1	1
3	4	1
4	38	3
5	656	26
6	22 104	474

(f) Hexagons n,n,m,n,n,m

n,n,m,n,n,m	N	C_n
2, 2, 2, 2, 2, 2	10	6
3, 3, 3, 3, 3, 3	20 843	1 284
2, 2, 3, 2, 2, 3	40	40
2, 2, 4, 2, 2, 4	1 090	132

TABLE III. (*Continued.*)

(g) Skewed parallelepipeds (fcc) $2\times 2\times n$

n	N	C_n
2	203	30
3	8 084	514
4	296 616	10 136
5		

(h) Skewed parallelepipeds $2\times 3\times n$

n	N	C_n
2	11 628	381
3	4 301 512	64 758
4	1 617 258 514	14 000 959

(i) Skewed parallelepipeds 1 tri$\times n$

n	N	C_n
2	62	7
3	618	28
4	5 348	114
5	41 836	468
6	307 764	1 916
7	2 177 928	7 848
8	15 020 794	32 144
9	101 822 828	131 656

(j) Skewed parallelepipeds 2 tri$\times n$

n	N	C_n
2	105	42
3	12 352	726
4	449 942	14 282
5	14 652 475	277 002
6	448 917 888	5 380 484

(k) Skewed parallelepipeds 3 tri$\times n$

n	N	C_n
2	2 188	103
3	173 740	3 722
4	12 656 898	152 922
5	818 944 912	6 188 332

(l) Skewed parallelepipeds 1 hex$\times n$

n	N	C_n
2	137 971	7 588
3	183 278 209	4 542 244

n indicates the length (in lattice units) of the sides of the equilateral triangle. For regular and semiregular (rhomblike) hexagons the set of six numbers (l,n,m,l,n,m) provides lengths in lattice units of all consecutive sides of the hexagon. The second column gives the total number of directed Hamiltonian paths unrelated by symmetry. The third column gives the corresponding number of Hamiltonian circuits. Figure 15 shows the relationship between the number of conformations and the width m of the cross section for various lengths n of the parallelogram.

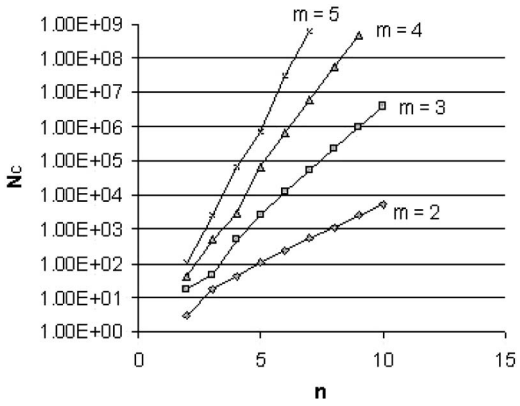

FIG. 15. The plot of the number of possible Hamiltonian chains N_c vs the length n for varying widths m of the $m \times n$ parallelogram.

We also analyzed the relationship between the volume to area ratio and the total number of lattice sites (same as protein length) for a given lattice shape. We used the data fromthe earlier studies on the 3D cubic lattice as a preliminary step. We compared this with data for real protein sequences obtained from the protein data bank (PDB).

We used PISCES (Ref. 19) to cull a data set of protein sequences with the following properties: the maximum percentage identity is less than 25%, the maximum resolution is below 2.0 Å, the maximum R value is below 0.3, the minimum chain length is 40, and the maximum chain length is 60. We used the whole PDB (Ref. 20) entry instead of separate single chains to obtain a nonredundant data set of 26 high-resolution proteins. The minimum chain length is automatically constrained to 40 by PISCES. We limited the maximum sequence length to 60, as our aim here is to analyze how the real proteins compare with a 3D square lattice, instead of observing how volume/area ratio changes with protein length. The molecular surface area and volume calculations were performed with DEEPVIEW.[21]

In the 3D square lattice representation of proteins, an increase in protein length leads to a larger number of occupied lattice nodes and therefore a larger lattice volume. Due to geometrical constraints, with increasing protein length the volume increases at a faster rate than the area, leading to an upward slope of the volume/area ratio. This ratio can therefore be useful for assessing the reliability of any lattice model to represent protein structures. Though other descriptors, such as folding rates and secondary structure formation, can also be used, a simple comparison of the volume/area ratio from simulations and experiments can provide a way of testing the model performance. Figure 16 shows such a comparison to validate the performance of 3D square lattice models for representing real proteins. Here, experimental and simulation data are provided for dissimilar protein lengths. Although computational power limits the availability of results for simulated sequences to a length of 48, the experimental data nicely integrate with the simulation data for proteins up to the length of 60 residues. The presence of noise in the experimental volume/area ratios creates a scattered plot, yet the experimental data are strongly compatible with those obtained in the simulations.

The complementarity between experimental and simulation results in Fig. 16 emphasizes the usefulness of lattices in analyzing conformational properties of proteins constrained by geometrical requirements.

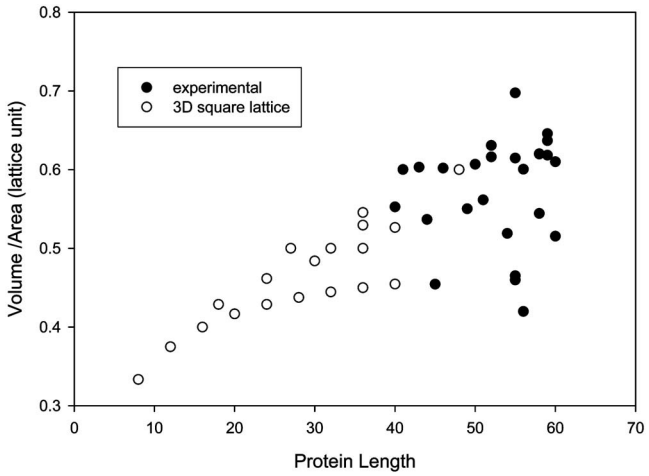

FIG. 16. The volume/area ratio comparison of 3D cubic lattice conformations and real proteins as a function of protein length (length meaning number of residues).

DISCUSSION

The fact that we can enumerate compact lattice conformations and extend the transfer matrix method to the 2D triangular lattice lends support to the idea that this methodology, developed originally in the past for the square and cubic lattices, can be applicable for to other lattices. Since the transfer matrix method reduces highly complicated geometrical packing problem to much simpler algebraic problem the possibility of the reduction of the complexity of the protein folding problem through the transfer matrix approach is extremely appealing. Our approach is especially applicable to a hierarchical method of protein structure prediction which starts with complete enumeration of compact conformations on a lattice. Such approach, originally proposed by Levitt and co-workers,[22,23] is now becoming very popular because of its efficiency.

One specific application for which the transfer matrix method can be applied involves calculation of the average energy of an ensemble of compact lattice conformations for the hydrophobic-polar model.[24] We intend to study this problem in detail for the triangular and fcc lattices in the future.

ACKNOWLEDGMENT

The authors acknowledge the financial support provided by NIH Grant No. R01GM072014 to two of the authors (A.K. and R.L.J.).

[1] H. S. Chan and K. A. Dill, J. Chem. Phys. **92**, 3118 (1990).
[2] H. S. Chan and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 6388 (1990).
[3] H. S. Chan and K. A. Dill, Macromolecules **22**, 4559 (2003).
[4] G. M. Crippen, J. Chem. Phys. **112**, 11065 (2000).
[5] J. G. des Cloizeaux, *Polymers in Solution* (Oxford University Press, Oxford, 1989).
[6] A. J. Guttmann and I. G. Enting, Phys. Rev. Lett. **76**, 344 (1996).
[7] I. Jensen, Comput. Phys. Commun. **142**, 109 (2003).
[8] N. Madras and G. Slade, *The Self-Avoiding Walk* (Birkhauser, Boston, 1993).

[9] E. I. Shakhnovich, Folding Des. **1**, R50 (1996).
[10] M. Peto, A. Kloczkowski, and R. L. Jernigan, J. Phys.: Condens. Matter **19**, 285220 (2007).
[11] D. G. Covell and R. L. Jernigan, Biochemistry **29**, 3287 (1990).
[12] E. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[13] A. Kloczkowski and R. L. Jernigan, Comput. Theor. Polym. Sci. **7**, 163 (1997).
[14] A. Kloczkowski and R. L. Jernigan, Macromolecules **30**, 6691 (1997).
[15] A. Kloczkowski and R. L. Jernigan, J. Chem. Phys. **109**, 5147 (1998).
[16] A. Kloczkowski and R. L. Jernigan, J. Chem. Phys. **109**, 5134 (1998).
[17] T. G. Schmalz, G. E. Hite, and D. J. Klein, J. Phys. A **17**, 445 (1984).
[18] Z. Bagci, A. Kloczkowski, R. L. Jernigan, and I. Bahar, Proteins: Struct., Funct., Genet. **53**, 56 (2003).
[19] G. L. Wang and R. L. Dunbrack, Bioinformatics **19**, 1589 (2003).
[20] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliand, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[21] N. Guex and M. C. Peitsch, Electrophoresis **18**, 2714 (1997).
[22] R. Samudrala, Y. Xia, E. Huang, and M. Levitt, Proteins: Struct., Funct., Genet. Suppl. **3S**, 194 (1999).
[23] Y. Xia, E. S. Huang, M. Levitt, and R. Samudrala, J. Mol. Biol. **300**, 171 (2000).
[24] A. Kloczkowski, T. Z. Sen, and R. L. Jernigan, Polymer **45**, 707 (2004).