Available online at www.sciencedirect.com
SCIENCE DIRECT®

Polymer 45 (2004) 707–716

polymer

www.elsevier.com/locate/polymer

The transfer matrix method for lattice proteins—an application with cooperative interactions

Andrzej Kloczkowski*, Taner Z. Sen, Robert L. Jernigan

L.H. Baker Center for Bioinformatics and Biological Statistics, 112 Office and Laboratory Bldg., Iowa State University, Ames, IA 50011-3020, USA

Received 17 October 2003; received in revised form 22 October 2003; accepted 22 October 2003

Abstract

The transfer matrix method for generating lattice conformations of proteins is explained and applied to lattice proteins having high-level cooperativity to represent hydrophobic interactions. The main advantage of the method is the extremely efficient attrition-free generation and enumeration of compact conformations. We review the application of the method for the generation and complete, exact enumeration of all conformation for linear and cyclic chains in 2D on the square lattice and in 3D on the cubic lattice. We show for compact conformations that the growth of the chain in a piecewise way, cross-section by cross-section, is much more efficient than the traditional linear chain growth. We discuss an extension of the method by including information about the amino acid sequence. We develop a Zimm–Bragg [J Chem Phys 31 (1959) 476–85]-like theory of hydrophobic cluster formation by using the transfer matrix method. We show that the transfer matrix approach to the generation and averaging over chain conformations can be formulated as an algebraic problem. We show also how the transfer matrix method can be extended to off-lattice proteins.

Keywords: Transfer matrix method; Helix–coil transition; Zimm–Bragg theory

1. Introduction

Understanding protein folding is extremely important for scientific progress in biology, but it is a problem of high complexity to find the structure that corresponds to the global minimum in energy. Globular proteins have dense, compact cores composed of mostly hydrophobic residues, hence compact self-avoiding walks on lattices represent the simplest model capturing basic features of globular proteins [2–18]. A compact self-avoiding walk is defined as a walk such that all sites within the protein's shape have been visited once and only once, voids (unvisited sites) are not allowed. In mathematics, such walks are called Hamiltonian paths, or in the case when the starting and the ending points of the walk coincide to form closed loops: Hamiltonian cycles. The imposition of a lattice greatly simplifies the problem of the generation of protein conformations, enabling (at least theoretically) the generation and enumeration of all possible compact conformations within a given volume. Unfortunately, the computer time required for computations grows geometrically with the length of the chain, imposing practical limitations on the length of the protein chain for which the complete enumerations are possible. Because of these limitations usually, instead of a complete enumeration of the conformational space, a random sampling of the space is performed. This is a process which is inconsistent with the requirements, since the protein native conformation is unique and any random search method would typically fail to locate the native structure. Therefore complete enumerations, whenever feasible for protein folding, are to be strongly preferred.

One major problem encountered in the generation of compact conformations in a standard way by growing the chain linearly is attrition. Attrition is the result of the excluded volume in a dense system—the site visited once by the self-avoiding walk cannot be visited again. Because of this, during the linear generation of the chain (walk) we encounter growing numbers of dead ends, after which further generation of the chain is impossible (but there remain unoccupied sites)-it becomes necessary to back off one or more steps and try other possibilities to complete the walk. The attrition grows geometrically with the length of the chain and becomes a major obstacle in generation of

* Corresponding author. Tel.: +1-515-294-9598; fax: +1-515-294-3841.
E-mail address: kloczkow@iastate.edu (A. Kloczkowski).

0032-3861/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2003.10.072

compact conformations. Likewise attrition becomes a problem when unoccupied sites become inaccessible. These problems of attrition can be avoided completely by a novel method of generation of compact self-avoiding walks. This method is based on the algebraic transfer matrix formalism. The main idea of this method is a different approach to the chain connectivity in compact conformations. Instead of the traditional linear growth of the chain (that leads to the attrition) the chain is grown in a piecewise way, cross-section by cross-section, until the Hamiltonian path (or Hamiltonian cycle) is completed.

2. The transfer matrix method

The transfer matrix method was first applied to phenomenological renormalization of the self-avoiding walk on the square lattice [19,20]. The method was used later by Schmaltz, Hite and Klein for enumerations of Hamiltonian circuits in 2D on the square and honeycomb lattices [21]. Kloczkowski and Jernigan have extended this method to Hamiltonian circuits in 3D on the cubic lattice, and to Hamiltonian paths (chains) both in 2D on the square lattice and in 3D on the cubic lattice [22–25].

Because chains with no ends (circuits) are simpler to treat, we first explain briefly the idea of the transfer matrix method for enumerations of Hamiltonian circuits in 2D upon rectangles with the square lattice. The Hamiltonian circuit (Fig. 1a) 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. 1b) does not need to satisfy the second condition, and the walk (chain) has two ends.

The main simplifying idea in this method is to take individually each column of sites on the square lattice and define the connectivities (on one side) of these sites as 'states'. With such a definition of a 'state' there are relatively few allowed 'transitions' from a given state to the states of the neighboring column. To 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. Fig. 2a shows all possible external connectivities to one side of the 4 points on a line. Fig. 2b shows all possible distributions of bonds between the 4 points on a line, including the case with no bonds (#1 where all bonds would be to the neighboring lines). We note that intersecting connectivities such as #9 in Fig. 2a are not allowed. Additionally, connectivities #4 and #5 in Fig. 2a are not allowed due to the parity reasons, so the total number of the possible connectivity states is only six in this simple example.

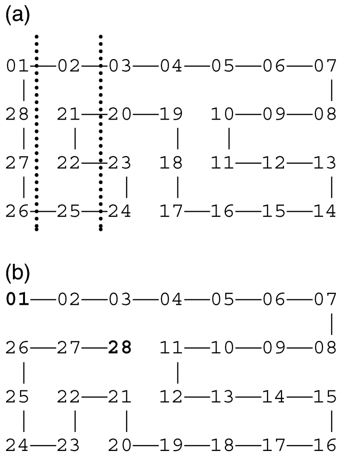


Fig. 1. Hamiltonian circuit (a) and Hamiltonian path (b) on the square 4 × 7 lattice. Cross-sections defining the first two connectivity states from the left for the Hamiltonian circuit are shown.

The transfer matrix **T** is constructed by combining all connectivity states (Fig. 2a) with all bond distributions (Fig. 2b) and finding the resulting connectivity states formed by their combinations. The combinations, which lead to unoccupied sites, triple connections, or to the formation of small loops, are not allowed. To better understand this approach, consider the Hamiltonian circuit illustrated in Fig. 1a. The first connectivity state (on the left, #6 in Fig. 2a) is obtained by making the first vertical cross-section, shown in Fig. 1a as a dashed line. The next connectivity state (#8 in Fig. 2a) corresponds to the second vertical cross-section (dashed line) and is obtained by superimposing the previous connectivity state on the bond distribution (#3 in Fig. 2b) in the second column of sites in Fig. 2a. 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 **T** are either 0 or 1, but generally T_{ij} can be larger than 1.) We

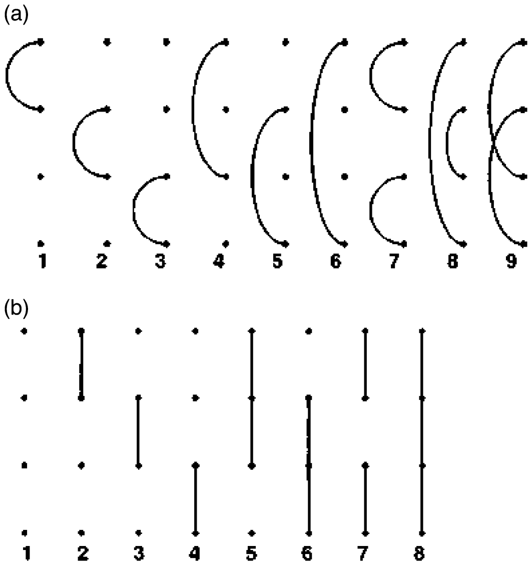


Fig. 2. All possible connectivity states (a) and bond distributions (b) for generation of Hamiltonian circuits within the rectangles of size $4 \times n$.

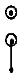
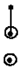
construct the vector **u** of the starting states with elements u_i, for each connectivity state i (such as in Fig. 2a) as the first state on the left in the process of building a circuit (we use a left to right convention). The number u_i identifies the number of different ways in which this may be realized. As starting states, we use the distributions of bonds (such as in Fig. 2b) that do not contain any unoccupied sites (#7 and 8 in Fig. 2b). 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 (the first dashed line in Fig. 1a) is taken. (The distribution of bonds #7 in Fig. 2b leads to the connectivity state #7 in Fig. 2a, while the distribution #8 leads to the connectivity state #6.) We also construct the vector **v** of the ending states with elements v_i, determining whether 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

$$N_c = \mathbf{u}^{\mathrm{T}}(\mathbf{T})^{n-2}\mathbf{v} \qquad (1)$$

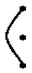
with the superscript T denoting the transpose of vector **u**. For the purpose of this example, if we omit the impossible states 4, 5 and 9 in Fig. 2a and renumber the remaining states from 1 to 6 then the transfer matrix **T**, the vectors of the starting states **u** and the ending states **v** are:

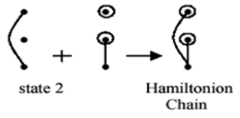
$$\mathbf{T}=\begin{pmatrix}0&0&0&1&1&0\\0&0&0&1&0&0\\0&0&0&1&1&0\\1&1&1&0&0&1\\0&0&0&1&1&0\\1&0&1&0&0&1\end{pmatrix},\quad \mathbf{u}=\begin{pmatrix}0\\0\\0\\1\\1\\0\end{pmatrix},\quad \mathbf{v}=\begin{pmatrix}0\\0\\0\\1\\0\\1\end{pmatrix} \qquad (2)$$

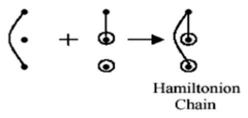
We have extended the method to Hamiltonian chains with two ends (Hamiltonian paths) in 2D on the square lattice by generalizing the definition of the connectivity state to include the connectivities with up to two ends, and by generalizing bond distributions by including up to two ends [25]. Fig. 3 shows all possible connectivity states for the generation of Hamiltonian paths on the rectangle of size $3\times n$ on the square lattice, and Fig. 4 shows all possible distributions of bonds.

The vector of starting states **u** for this case is given by Eq. (3a) and the vector of ending states **v** is given by Eq. 3b. For example, the connectivity state 11 has value 2 as the starting state in Eq. (3a) because it can be formed both from  and  in the first column of the rectangle.

$$\begin{array}{c}1\\2\\3\\4\\5\\6\\7\\8\\9\\10\\11\\12\end{array}\begin{bmatrix}0\\1\\0\\1\\0\\1\\1\\0\\1\\1\\2\\1\end{bmatrix} \qquad (3a) \qquad\qquad \begin{array}{c}1\\2\\3\\4\\5\\6\\7\\8\\9\\10\\11\\12\end{array}\begin{bmatrix}1\\2\\1\\1\\2\\1\\1\\0\\1\\0\\1\\0\end{bmatrix} \qquad (3b)$$

Similarly the connectivity state 2  has value 2 as the ending state because two different distributions of bonds lead to the formation of the Hamiltonian chain (path), namely:





The transfer matrix is built in a similar way as for Hamiltonian circuits by superimposing the connectivity state upon the bond distribution and finding out what will be the connectivity state in the next column (cross-section) resulting from this superimposition. The superimpositions leading to non-physical cases such as: (1) unoccupied sites, (2) triple connections, (3) double connection of chain ends, (4) formation of loops, (5) creation of more than two ends, and finally (6) to chain disintegration, or breaking into separate pieces, are not allowed. The transfer matrix for the

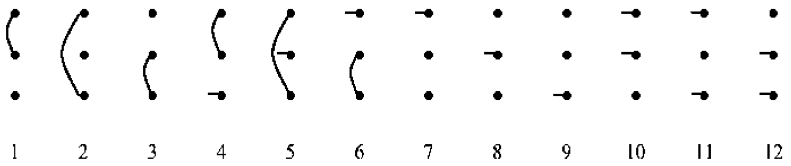
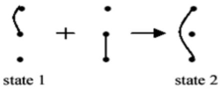


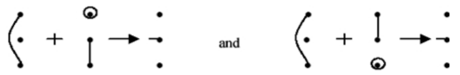
Fig. 3. Connectivity states for generation of Hamiltonian paths within the rectangles of size $3\times n$.

case of $3 \times n$ square lattice is given below:

$$\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 2 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad (4)$$

For example $T_{12} = 1$ because there is a unique combination of the connectivity state 1 with bond distributions leading to the connectivity state 2: 

On the other hand $T_{28} = 2$ because there are two different bond distributions that can combine with the state 2 leading to the connectivity state 8, namely:



Once (for a given cross-section, in this example $m = 3$) the transfer matrix is calculated the number of Hamiltonian paths for any length n of the rectangle is calculated from Eq. (1).

We have also generalized the transfer matrix method to 3D on the cubic lattice both for Hamiltonian circuits and Hamiltonian paths [24]. In 3D, we consider Hamiltonian circuits and Hamiltonian paths on the parallelepipeds of size $l \times m \times n$.

Fig. 5 shows examples of a Hamiltonian circuit and a Hamiltonian path. In this case, both the connectivity states and bond distributions are defined for the planar cross-section 2×2. As an example let us consider the possible connectivity states and bond distributions for Hamiltonian circuits. Fig. 6 shows all possible connectivity states for the generation of Hamiltonian circuits on $2 \times 2 \times n$ cubic lattice.

It should be noted that connectivity states 7, 8 and 9 in Fig. 6 are unphysical and the number of possible connectivity states is reduced to 6. Fig. 7 shows all possible distributions of bonds within the 2×2 cross-sections.

The transfer matrix is again created by superimposing connectivity states upon bond distributions and finding out the connectivity states in the next cross-section. For example the superimposition of the connectivity state #3 in Fig. 6 and the bond distribution #6 in Fig. 7 leads to the connectivity state # 2 in the next cross-section. Similarly as in two-dimensional superimpositions leading to unoccupied sites, triple connections or the formations of small loops are not allowed. Once (for a given cross-section $l \times m$) the transfer matrix is obtained the enumeration of Hamiltonian circuits is reduced to matrix multiplication (Eq. (1)).

Similarly, as in 2D the transfer matrix method is generalized for Hamiltonian paths in 3D by considering up to two ends in the definition of connectivity states and bond distributions. The transfer matrix is formed by the superimpositions of connectivity states and bond distributions subject to similar rules as for two dimension (unoccupied sites, triple connections, double connection of chain ends, formation of loops, creation of more than two ends and chain disintegration are not allowed).

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 size of the transfer matrix equals to the number of all possible combinations of connectivities (including the connectivity to chain ends for Hamiltonian paths) within the cross-section and therefore grows with the size of the cross-section. Additionally, in 3D, the number of connectivity states grows much faster than in 2D, because in 3D the requirement that connectivities cannot intersect (such as state #9 in Fig. 2a) does not hold anymore. (The state #9 in Fig. 6 is an exception due to the small size of the cross-section.)

The complete enumerations of Hamiltonian circuits and Hamiltonian paths in 2D on the square lattice and in 3D on the cubic lattice were published previously [24,25]. For example, the number of Hamiltonian paths (chains with ends) within the 8×12 rectangle on the square lattice is 144,397,808,917,246 and the number of Hamiltonian circuits (no ends) within the $3 \times 3 \times 8$ parallelepiped on the cubic lattice is 468,855,089,493,448. The computer program was used to calculate transfer matrices as large as 3104×3104. Now, because of significant increases in computer memory, and if sparse matrix efficiencies were utilized, these calculations could be extended to significantly larger matrices [26].

3. Extension of the method by adding potentials

The transfer matrix method of 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 be easily performed even for

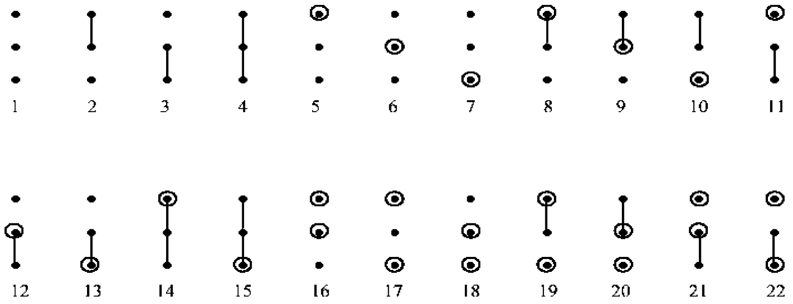


Fig. 4. Bond distributions for generation of Hamiltonian paths within the rectangles of size $3 \times n$. Small circles indicate chain ends.

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 has been automated in order to access larger structures. Because calculations of transfer matrices are generated with a computer program, we are only limited by the storage requirements of large matrices. The conversion into algebra of the highly complex compact self-avoiding walk problem is the origin of the beauty and power of this method. The method has only been used to generate and enumerate compact lattice conformations. However, the protein folding problem is a combination of two problems: the problem of generating all possible conformations (structures), and the problem of imposing the sequence of amino-acids on the structure and evaluating structures to identify low energy forms when the inter-residue potentials are defined. We discuss next, possible future modifications of the method to deal with more complicated problems involving energy calculations, irregularities (including cavities - needed because protein packing is not perfect) in the protein structure and off-lattice applications of the method. However the study of protein folding includes, in addition to the counting and enumerations of all possible conformations (structures), also how to include effects from the sequence of amino acids, calculation of the total energy of the structure and the probabilities of the possible states of the protein.

In the future the transfer matrix method may find a possible application for finding the structures having the lowest energy. We may extend the transfer matrix method by considering the simplest hydrophobic–polar (HP) model of lattice proteins. The HP lattice model has been studied extensively. These studies led to the discovery of many interesting properties of proteins folds. One particularly important discovery was the concept of designability of protein structures as developed by Li, Tang and Wingreen [27–33]. It has been proven that some protein structures are highly designable, i.e. a large number of different sequences of HP residues have the lowest energy associated with the same structure. This explains, in principle, why so many different and diverse protein sequences have similar folds.

We may extend the transfer matrix method by using the two-letter (hydrophobic–polar) labeling of residues. We may use two approaches: in the first, simplest case, the hydrophobic residues are only allowed inside the core of the lattice protein; in the second case, all possible combinations of H and P residues would be permitted within the cross-section are allowed.

The introduction of actual H and P labeling of lattice sites rather than bonds in the sequence corresponds to considering a distribution over different possible distributions of bonds or sequences. This is because we generalize the definition of the distribution of bonds by adding 'colors' associated with the types of the residues, but at lattice sites, not in the sequence.

By using the case of Hamiltonian circuits placed within the rectangle of size $4 \times n$, when hydrophobic residues are only allowed inside the core, the number of possible distributions of bonds is only 10. (This is because we have 8 distribution of bonds shown in Fig. 2b where for four points on the line the ending two points 1 and 4 are given the color 'P' and the two points 2 and 3 in the center are given the color 'H', and we have two distributions of bond patterns [#7 and #8 in Fig. 2b] for ends of the rectangle where color P is assigned to all four points on the line. We need these latter two distributions of bonds for the start and termination of the Hamiltonian space.) In the second model, when H and P colors can be distributed within the cross-section without

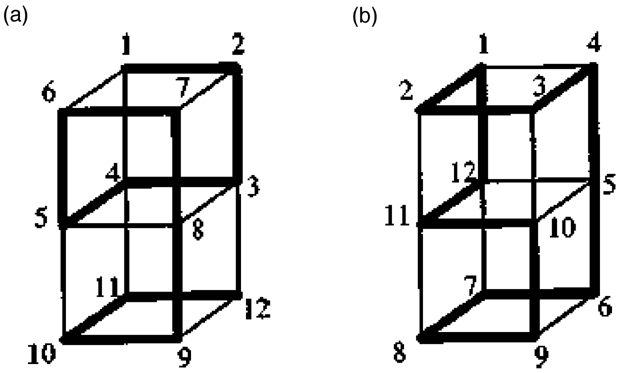


Fig. 5. Hamiltonian path (a) and Hamiltonian circuit (b) for $2 \times 2 \times 3$ cubic lattice.

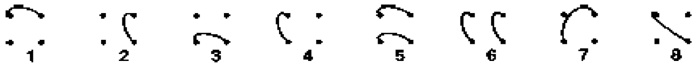


Fig. 6. All possible connectivity states for generation of Hamiltonian circuits within the parallelepipeds of size $2\times2\times n$.

restriction, the number of possible bond distributions is 128, because for each of the 8 distributions of bonds in Fig. 2b there are 2^4 possibilities of assigning H and P colors.

To illustrate this method let us consider the Zimm–Bragg [1,34] type of the model of proteins using the transfer matrix combined with HP potentials. For simplicity we will consider Hamiltonian circuits (no ends) on the square lattice within rectangles of size $4\times n$. (Here the numerical calculation were performed for the length of the rectangle $n=10$, by example.) It is assumed that all $2\times(n-2)$ residues inside the rectangle are hydrophobic, while all $2n+4$ residues on the 'surface' of the rectangle are polar. The transfer matrix method enables the generation of all conformations within the rectangle of size $4\times n$. To keep track of all the nearest neighbor interactions between the neighboring columns in the rectangle it is useful to define fragments (tiles) of size 4×2 consisting of two columns. Generation of Hamiltonian circuits imposes certain rules for fragment candidates. First, every node on the lattice should be connected with exactly two other nodes. Additionally, the connections of the nodes cannot lead to the disintegration of the chain (smaller circuits), and the chessboard-like parity of the square lattice eliminates some possibilities for fragments. To achieve this aim, three types of fragments are defined: starting, middle, and ending fragments. Note that because of the symmetry the ending fragments will be mirror images of starting fragments. The number of starting, middle, and ending fragments are 6, 36, and 6, respectively. Fig. 8 shows all six possible starting fragments for the generation of these Hamiltonian circuits.

A Hamiltonian circuit is created when a starting fragment is followed by middle fragments (or directly by an ending fragment), which can be used to increase the size of Hamiltonian circuits, and the circuit is finished when an ending fragment is matched with a preceding middle fragment. Assembling the fragments to create a circuit requires fitting the second column of a given fragment with the first column of a matching fragment according to a connection table (transfer matrix) created for specifying possible matches between fragments. The first two columns of the Hamiltonian circuit are determined by a starting tile, and every matching middle or ending fragment (tile) adds one more column to the forming circuit. After identifying each possible fragment, statistical weights are assigned depending on the nature of non-bonded interactions a fragment adds. In our lattice model, the outside nodes are defined as polar, and the inside (core) nodes are defined as hydrophobic. Three distinct non-bonded interactions are defined depending on the types of nodes that are involved in the interaction: hydrophobic, polar, and mixed with contact, non-bonded energies E_{HH}, E_{PP}, and E_{HP}, respectively. The statistical weights for each interaction are defined as h, p, and m, in Eq. (5)

$$h=\mathrm{e}^{-E_{\mathrm{HH}}/RT},\qquad p=\mathrm{e}^{-E_{\mathrm{PP}}/RT},\qquad m=\mathrm{e}^{-E_{\mathrm{HP}}/RT} \tag{5}$$

In our calculations we have set the E_{PP} to zero to define the energy scale reference point so that $p=1$.

Following this methodology, the non-bonded interactions are counted for each fragment and a statistical weight is assigned. For example, if two hydrophobic and one mixed non-bonded interactions are formed in any given fragment, then the statistical weight of this added fragment will be h^2m. If three hydrophobic non-bonded interactions are formed in a fragment, than an additional statistical weight, s, is assigned and the resulting statistical weight becomes sh^3. This extra weight represents a cooperativity cost for the formation of three clustered hydrophobic interactions inside the protein core.

Since the circuit generation involves a starting fragment, middle fragment(s), and an ending fragment, the statistical weights are defined according to the non-bonded interactions created at each step. As a result, the calculation of statistical weights differs for starting and other types of fragments. For a starting fragment, the calculation of non-bonded interactions consists of searching two columns of nodes in both horizontal and vertical directions. In contrast, for all middle and ending fragments, the non-bonded interactions along the vertical direction in the first columns are not taken into account in order to avoid double counting. The exception to this, is the first starting tiles, where these interactions must be included. For example the statistical weight for the 6 starting tiles in Fig. 8 are phm^2, $pm^2\ m^4$, m^3, m^2h, and m^3, respectively. Since the total number 36 of possible fragments (tiles) is too large for presentation, Fig. 9

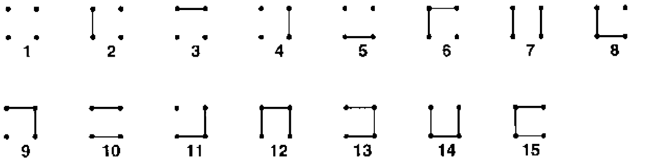


Fig. 7. All possible distributions of bonds for generation of Hamiltonian circuits within the parallelepipeds of size $2\times2\times n$.

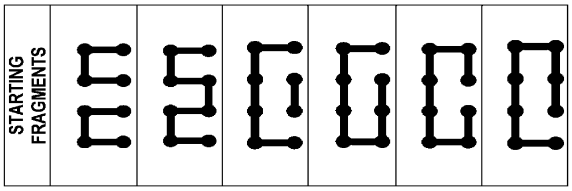


Fig. 8. Starting fragments for generation of Hamiltonian circuits within the rectangles of size $4\times n$.

show only representative fragments for each of the possible statistical weights. The frequencies listed in Fig. 9 are the counts of conformations of different fragments having the same statistical weight.

The statistical weights for each fragment are stored in three matrices created according to the transfer matrix (connection table) rules: $\mathbf{S}_{6\times36}$, $\mathbf{M}_{36\times36}$, and $\mathbf{E}_{36\times6}$ for starting, middle, and ending states.

By an analogy with the Zimm–Bragg [1] theory we can define the partition function for all Hamiltonian paths given by the matrix multiplication. For example the partition function Z for Hamiltonian paths placed on a 6×4 lattice is:

$$Z = \mathbf{e}^{\mathrm{T}}\mathbf{S}\mathbf{M}_1\mathbf{M}_2\mathbf{M}_3\mathbf{E}\mathbf{e} \tag{6}$$

where $\mathbf{e}$ is the vector of 1's and T denotes the transpose.

It is possible to calculate the statistical averages of non-bonded pair types using the well-known results of Zimm–Bragg theory [1]. For example, the mean number of hydrophobic non-bonded interactions can be calculated using

$$\langle N_h \rangle = \frac{\partial(\ln Z)}{\partial(\ln Z)} = \frac{\partial(Z)}{Z\partial(\ln h)} \tag{7}$$

Jernigan in 1966 [34,35] proposed a method for handling this equation through the decomposition of Z and the direct matrix product involving only one derivative

$$\frac{\partial(Z)}{\partial(\ln h)} = \sum_{m=0}^{N-4} \mathbf{e}^{\mathrm{T}}\mathbf{S}\mathbf{M}^{m}\mathbf{M}'\mathbf{M}^{N-m-4}\mathbf{E}\mathbf{e} \tag{8}$$

where $\mathbf{e}$ is the vector of 1's and

$$\mathbf{M}' = \frac{\partial \mathbf{M}}{\partial(\ln h)} \tag{9}$$

Figs. 10 and 11 show the results of calculations for a rectangle of length $n = 10$, with $p = 1$ for varying m (with $s = 10^{-10}$ in Fig. 10) and for various s values (with $m = 60$ in Fig. 11). Both figures show characteristic sigmoidal shapes similar to the Zimm–Bragg helix–coil transition [1,34]. The average value $\langle N_h \rangle$ (calculated with the use of statistical weights) is a measure of the extent of hydrophobic interactions. Fig. 10 show that this hydrophobic affinity is increasing for smaller values of m. (For m close to one the curves are similar to a step function). The extra cooperativity of the HH non-bonded contacts afforded by the statistical weight s leads to sharp increase in the hydrophobic affinity and the step-function-like shape of the plots in Fig. 11 for small s. It is interesting that the one-dimensional helix–coil theory has helix formation values of σ typically near 10^{-4}, and here in 2D we find values leading to sharp transitions to be near σ^2 (although these are not really comparable in any specific ways).

The calculations in Figs.10 and 11 were performed for quite arbitrary values of parameters h, m, and s. It is worth mentioning that the sigmoidal character of the plots is nearly universal and does not occur only for a few unique sets of values of these parameters. However, some choices of these parameters lead to a characteristically different behavior manifested of the plots. Fig. 12 shows the results ($\langle N_h \rangle$ as a function of s) obtained for two frequently used sets of values for the energies of hydrophobic–polar interactions. The dashed line in Fig. 12 corresponds to the simplest HP model. In this case the plot of $\langle N_h \rangle$ as a function of s is almost a straight line. The solid line in Fig.12 corresponds to the values of energies of HH, HP and PP interactions the same as in the paper by Li, Helling, Tang and Wingreen [33]. The plot in this case has the sigmoidal character.

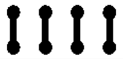
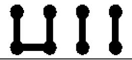
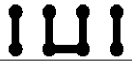
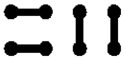
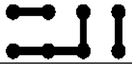
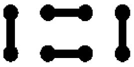
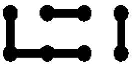
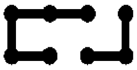
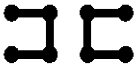
	FREQUENCY	REPRESENTATIVE
hm^2	4	
hm	4	
m^2	3	
ph^2m	4	
phm	4	
h^2m^2	4	
h^2m	8	
sh^3	4	
p^2h	1	

Fig. 9. All possible statistical weights of middle fragments with representative fragments corresponding to the given statistical weight. To save the space the fragments have been rotated by 90° clockwise. The interactions in the upper row (corresponding to the left column before the 90° rotation) are not counted in the calculation of the statistical weights, because these were included in the previous stage.

We should note that results shown in Figs. 10–12 were obtained by averaging over all possible conformations (those completely enumerated by the original transfer matrix method). The present method is an extension of the transfer matrix method, that includes the complete enumeration of conformations, calculates the partition function associated with non-bonded interactions in conformations of lattice proteins.

Similar sigmoidal curves (not shown here because of the space limit) we obtain by plotting $\langle N_m \rangle$ as a function of m. Calculations performed for different sizes of the $4 \times n$ rectangle show similar sigmoidal Zimm–Bragg helix–coil-like cooperative transitions. The numerical calculations are

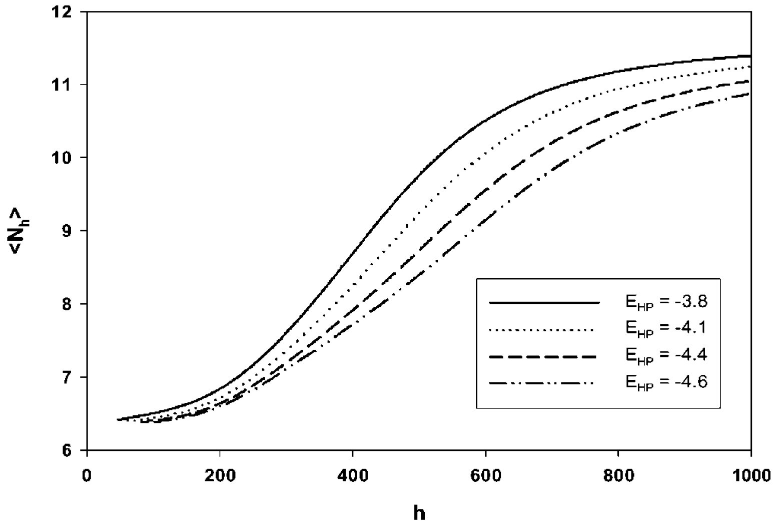


Fig. 10. The statistical average $\langle N_h \rangle$ as a function of h for several different values of E_{HP} (in units RT).

extremely fast and take only about 10 s on a desktop PC. By contrast, for the same rectangle size 4×10 the calculations based on traditional method of generation of all conformations take few minutes on a PC. However, because the computer time required for the traditional method grows geometrically with the size of the chains, for longer chains the transfer matrix approach shows enormous computational advantages over conventional methods. The calculations can be also performed for chains with ends (Hamiltonian paths) in 2D and for Hamiltonian paths and Hamiltonian circuits in 3D. We plan to study this problem in the near future. One principal problem with lattice proteins in 3D is that the hydrophobic cores of these model proteins are too small when the total number of points is small. For example in the case of the cube $3 \times 3 \times 3$ the hydrophobic core consists of a single residue. Another difficulty in 3D is the significant increase of the size of the transfer matrix.

4. Discussion

The main advantage of the transfer matrix method is that, by generating the conformations, cross-section by cross-section, all energies are instantly calculated; the energy is only associated with the nearest neighbor interactions, i.e. only the contacts inside the distribution of points on the line and the contacts with the next cross-section (given by the transfer matrix method) contribute to the total energy of the protein. Since we are looking for a structure with the lowest energy, we may reduce the space of all sequence-conformation combinations by storing only data corresponding to the set of lower energies, and we may discard unnecessary high-energy cases. Because the whole problem of protein folding is reduced to matrix algebra, the transfer matrix method has significant advantages over the traditional method of growing the chains and calculating their energies. One particular problem worth further study is the

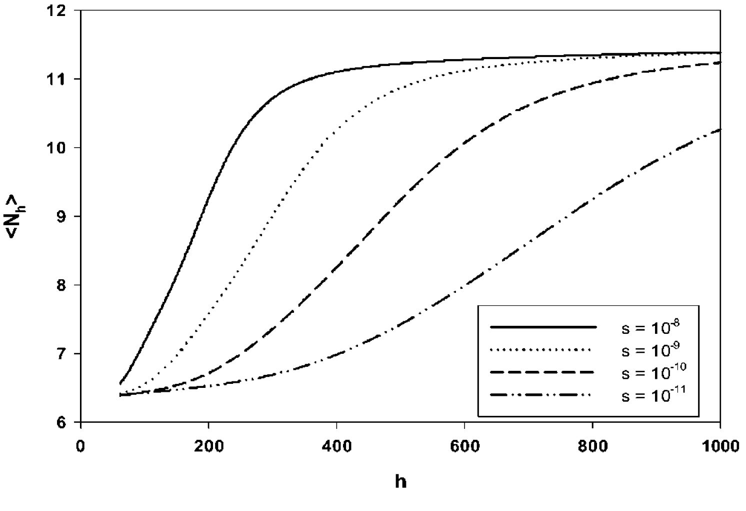


Fig. 11. The statistical average $\langle N_h \rangle$, as a function of h for several different values of s.

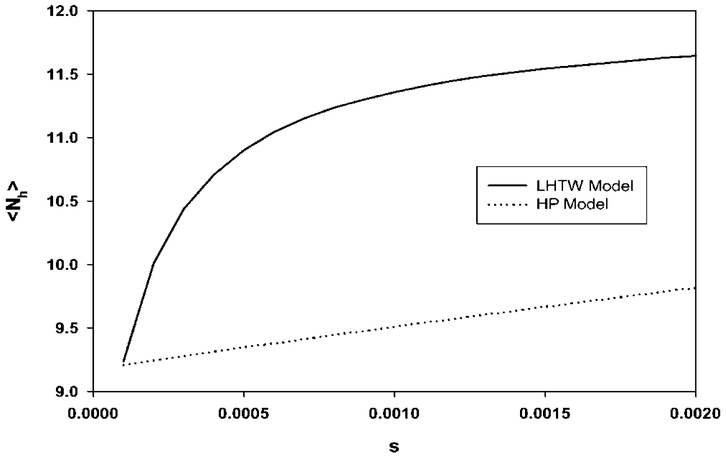


Fig. 12. The statistical average $\langle N_h \rangle$, as a function of s for the energy parameters corresponding to the simplest HP model (dashed line) and for corresponding to the Li, Helling, Tang and Wingreen model [33] (solid line).

possibility of defining the 'designability' of structures within the matrix formalism.

Another possibility is the redesign the transfer matrix method by using the formalism of direct products of matrices (developed by Jernigan and Flory for polymer chains) [36]. This could be done by defining new 'combined-states' by combining connectivity states with distributions of bonds and also with residue type variations. This new definition could be especially important for Hamiltonian paths in 2D and for both paths and circuits in 3D, because some of the elements of the transfer matrix method in the present form are integers larger than one (degenerate), since several different distributions of bonds superimposed on a given connectivity state lead to the same connectivity state in the next cross-section. The introduction of 'combined-states' will eliminate the degeneracy, which is important for energy calculations. The use of that formalism of direct products of matrices would allow us to apply many known methods from polymer computations to the lattice protein folding problem.

The inclusion of 'vacancies': unoccupied sites within the protein structure and irregularities [37–42] of the protein surface is another possible extension of the method. Unfortunately this increases the number of possible connectivity states in the method, so it would be easier to implement it in 2D.

Another possibility of the method is its generalization to off-lattice structures. It is much easier to generate compact conformations on lattices, because 2D or 3D lattices, especially square and cubic lattices, enable the formulation of the transfer matrix method in a highly rigorous way, such that all possible conformations can be enumerated and generated. However, this method can also be generalized to off-lattice models, although, completely rigorous definitions of the model are not possible. The simplest off-lattice models are those where the protein chain is densely packed into a regular three (or two) dimensional shape. An ellipsoidal model of dense protein packing provides a good starting point for the development of an off-lattice transfer matrix method. We can slice the ellipsoid into several uniformly spaced pieces of equal width by using several parallel equidistant planes in the same way an egg is cut uniformly by a slicing machine (Fig.13).

We may consider all conformations generated within the ellipsoidal shape and determine all possible connectivity states for a given slice (cross-section of ellipsoid). For the ellipsoid of revolution, the cross-section is a circle whose radius changes from $R = 0$ at two ends of the ellipsoid to $R = R_{\max}$ in the center of the ellipsoid. The connectivity state for a given cross-section illustrates how all pieces of a chain coming from the left side of the slicing plane enter the cross-section and reveals the topological connectivity of these points. This is a direct generalization of the connectivity states from the lattice models. Similarly to the lattice model, we can define a distribution of bonds within a given slice of the ellipsoid that when combined with the connectivity states at the given cross-section produces the connectivity state at the subsequent cross-section. This piece-wise approach to protein configurations (instead of the traditional picture of the protein as a linear chain) formulated in terms of connectivities may permit larger calculations of conformations that can shed a completely new light on protein structures.

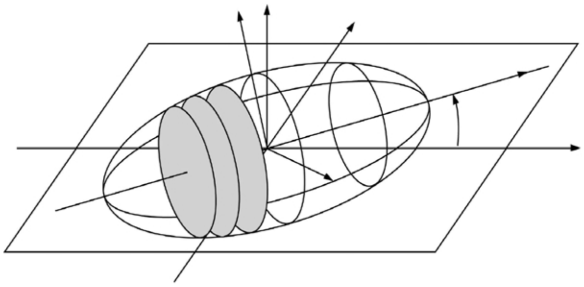


Fig. 13. An ellipsoid constricting the size of the generated protein chains. Cross-sections of the ellipsoid provide parallel planes to be used for the off-lattice transfer method.

References

[1] Zimm BH, Bragg JK. Theory of the phase transition between helix and random coil in polypeptide chains. J Chem Phys 1959;31:476–85.

[2] Pande VS, Joerg C, Grosberg AY, Tanaka T. Enumeration of Hamiltonian walks on a cubic lattice. J Phys A 2003;27:6231–6.

[3] Chan HS, Dill KA. The effects of internal constraints on the configurations of chain molecules. J Chem Phys 1990;92:3118–35.

[4] Chan HS, Dill KA. Polymer principles in protein structure and stability. Annu Rev Biophys Biophys Chem 1991;20:447–90.

[5] Chan HS, Dill KA. Sequence space soup of proteins and copolymers. J Chem Phys 1991;95:3775–87.

[6] Chan HS, Dill KA. The protein folding problem. Phys Today 1993;46: 24–32.

[7] Madras N, Slade G. The self-avoiding walk. Boston: Birkhauser; 1993.

[8] Chan HS, Dill KA. Interchain loops in polymers—effects of excluded volume. J Chem Phys 1989;90:492–509.

[9] Hinds DA, Levitt M. A lattice model for protein-structure predicton at low resolution. Proc Natl Acad Sci USA 1992;89:2536–40.

[10] Shakhnovich EI. Modeling protein folding: the beauty and power of simplicity. Fold Des 1996;1:R50–4.

[11] Kolinski A, Skolnick J. Lattice models of protein folding, dynamics and thermodynamics. New York: Chapman & Hall; 1996.

[12] Kolinski A, Skolnick J. High coordination lattice models of protein structure, dynamics and thermodynamics. Acta Biochim Pol 1997;44: 389–422.

[13] Sokal AD. In: Binder K, editor. Monte Carlo and molecular dynamics simulations in polymer science. Oxford: Oxford University Press; 1995. p. 47.

[14] Ramakrishnan R, Pekny JF, Caruthers JM. A combinatorial algorithm for effective generation of long maximally compact lattice chains. J Chem Phys 1995;103:7592–604.

[15] Bahar I, Jernigan RL. Stabilization of intermediate density states in globular-proteins by homogenous intramolecular attractive interactions. Biophys J 1994;66:454–66.

[16] Bahar I, Jernigan RL. Cooperative structural transitions induced by nonhomogenous intramolecular interactions in compact globular proteins. Biophys J 1994;66:467–81.

[17] Covell DG, Jernigan RL. Conformations of folded proteins in restricted spaces. Biochemistry 1990;29:3287–94.

[18] Shakhnovich E, Gutin A. Enumeration of all compact conformations of copolymers with random sequence of links. J Chem Phys 1990;93: 5967–71.

[19] Derrida B. Phenomenological renormalization of the self avoiding walk in 2 dimensions. J Phys A 1981;14:L5–L9.

[20] Klein DJ. Asymptotic distribution for self-avoiding walks constrained to stips, cylinders, and tubes. J Stat Phys 1980;23:561–86.

[21] Schmalz TG, Hite GE, Klein DJ. Compact self-avoiding circuits on two dimensional lattices. J Phys A 1984;17:445–53.

[22] Kloczkowski A, Jernigan RL. Efficient method to count and generate compact protein lattice conformations. Macromolecules 1997;30: 6691–4.

[23] Kloczkowski A, Jernigan RL. Computer generation and enumeration of compact self-avoiding walks within simple geometries on lattices. Comput Theor Polym Sci 1997;7:163–73.

[24] Kloczkowski A, Jernigan RL. Transfer matrix method for enumeration and generation of compact self-avoiding walks. II. Cubic lattice. J Chem Phys 1998;109:5147–59.

[25] Kloczkowski A, Jernigan RL. Transfer matrix method for enumeration ansi generation of compact self-avoiding walks. 1. Square lattices. J Chem Phys 1998;109:5134–46.

[26] Jensen I. Enumeration of compact self-avoiding walks. Comput Phys Commun 2001;142:109–13.

[27] Emberly EG, Wingreen NS, Tang C. Designability of alpha-helical proteins. Proc Natl Acad Sci USA 2002;99:11163–8.

[28] Li H, Tang C, Wingreen NS. Are protein folds atypical? Proc Natl Acad Sci USA 1998;95:4987–90.

[29] Li H, Tang C, Wingreen NS. Designability of protein structures: a lattice-model study using the Miyazawa–Jernigan matrix. Prot: Struct Funct Genet 2002;49:403–12.

[30] Melin R, Li H, Wingreen NS, Tang C. Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. J Chem Phys 1999;110:1252–62.

[31] Miller J, Zeng C, Wingreen NS, Tang C. Emergence of highly designable protein-backbone conformations in an off-lattice model. Prot: Struct Funct Genet 2002;47:506–12.

[32] Wang TR, Miller J, Wingreen NS, Tang C, Dill KA. Symmetry and designability for lattice protein models. J Chem Phys 2000;113: 8329–36.

[33] Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science 1996;273: 666–9.

[34] Poland D, Scheraga HA. Theory of helix–coil transitions in biopolymers. New York: Academic Press; 1970.

[35] Flory PJ, Miller WG. A general treatment of helix–coil equilibria in macromolecular systems. J Mol Biol 1966;15:284–97.

[36] Flory PJ. Statistical mechanics of chain molecules. New York: Wiley Interscience; 1969.

[37] Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. Analytical shape computing of macromolecules I: molecular area and volume through alpha shape. Proteins 1998;33:1–17.

[38] Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 1998;7:1884–97.

[39] Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. Analytical shape computing of macromolecules II: identification and computation of inaccessible cavities inside proteins. Proteins 1998;33: 18–29.

[40] Liang J, Zhang JF, Chen R. Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential Monte Carlo method. J Chem Phys 2002;117:3511–21.

[41] Liang J, Dill KA. Are proteins well-packed? Biophys J 2001;81: 751–66.

[42] Edelsbrunner H, Facello M, Liang J. On the definition and the construction of pockets in macromolecules. Dis Appl Math 1998;88: 83–102.