Efficient Method To Count and Generate Compact Protein Lattice Conformations

A. Kloczkowski* and R. L. Jernigan

Molecular Structure Section, Laboratory of Experimental and Computational Biology NCI, NIH, Building 12B, Room B116, Bethesda, Maryland 20892-5677

Received May 12, 1997

Revised Manuscript Received August 14, 1997

The generation and enumeration of all possible conformations of macromolecules is a long standing problem. The most common requirement has been to obtain a representative set of conformations of a random coil, which usually are placed on a lattice of one type or another. For these random coil cases, the usual practice has been to enumerate completely the conformations of short chains.[1] For longer chains, where a complete enumeration of conformations is impossible, the common practice for chain generations has been to utilize various Monte Carlo approaches.[2]

Globular proteins differ from random coils in having dense, compact cores partly as a result of the segregation between hydrophobic and polar residues. Because of their dense cores, compact self-avoiding walks (chains) on lattices provide an excellent model for globular proteins.[3−9] A compact self-avoiding walk is defined here as a self-avoiding walk with a compact shape, such that all sites within the shape are occupied; there can be no voids. With the self-avoiding walk method it is theoretically possible to generate and enumerate all possible compact conformations within a given volume, but the time required for computations grows geometrically with the length of the chain.[10]

The native conformations of proteins are compact and unique. The essence of the protein folding problem is to find, for a given sequence of amino acids, the most favorable conformation. This search for a unique form means that random search methods will often fail; complete enumerations, whenever feasible, are preferable.

The method presented here is an extension and modification of the method for phenomenological renormalization of the self-avoiding walk on the square lattice.[11,12] The method was used later by Schmalz, Hite, and Klein for enumerations of Hamiltonian circuits in two dimensions on the square and honeycomb lattices.[13] The Hamiltonian circuit 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 to the starting point. The regular Hamiltonian path does not need to satisfy the second condition, and the walk (chain) has two ends.

Here we extend this method to Hamiltonian circuits in three dimensions on the cubic lattice and to Hamiltonian paths (chains) both in two dimensions on the square lattice and in three dimensions on the cubic lattice. First we describe briefly the method used by Schmalz, Hite, and Klein for enumerations of Hamiltonian circuits upon rectangles on the square lattice.

The main simplifying idea in this method is to take individually each column of sites on the square lattice and define as "states" the connectivities (on one side) of these sites. With such a definition of a "state" there are relatively few allowed "transitions" from a given state to the states of the neighboring column.

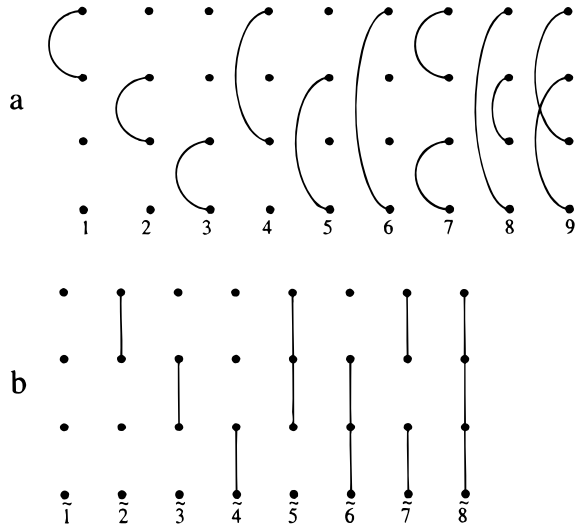

Figure 1. Connectivity states (a) and bond distributions (b) for the generation of Hamiltonian circuits within $4 \times n$ rectangles on the square lattice.

To illustrate this method, let us consider, as an example, 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 1a shows all possible external connectivities to one side of the 4 points on a line. Figure 1b shows all possible distributions of bonds between the 4 points on a line, including the case with no bonds (no. 1). We note that intersecting connectivities, such as no. 9 in Figure 1a, are not allowed. For 4 points there are 8 connectivity states (nonintersecting connectivities). However, two of these states (no. 4 and 5) are not possible for circuits on $4 \times m$ rectangles, for parity reasons, so the number of connectivity states is further reduced to 6. The number of all possible distributions of bonds for m points on a line is 2^{m-1}. The transfer matrix **T** is constructed by combining all connectivity states (Figure 1a) with all bond distributions (Figure 1b) and finding the resulting connectivity states formed by such combinations. For example, the combination of the connectivity state no. 1 (from Figure 1a) and the bond distribution no. 4 (from Figure 1b) shown in Figure 2a leads to connectivity state no. 7 from Figure 1a. The combinations that 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} shows the number of different ways to realize this transition. We note that 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 first construct the vector **u** of the starting states with elements u_i, for each connectivity state i (such as in Figure 1a), 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 shows the number of different ways in which this may be realized. As starting states, we use the distributions of bonds (such as in Figure 1b) that do not contain any unoccupied sites (nos. 7 and 8 in Figure 1b) and figure out the connectivity state to which the given distribution of vertical bonds transforms if the horizontal bonds connecting to vertical bonds in the neighboring column on the right side are added. We also construct the vector **v** of the ending states with

S0024-9297(97)00662-1 CCC: $14.00

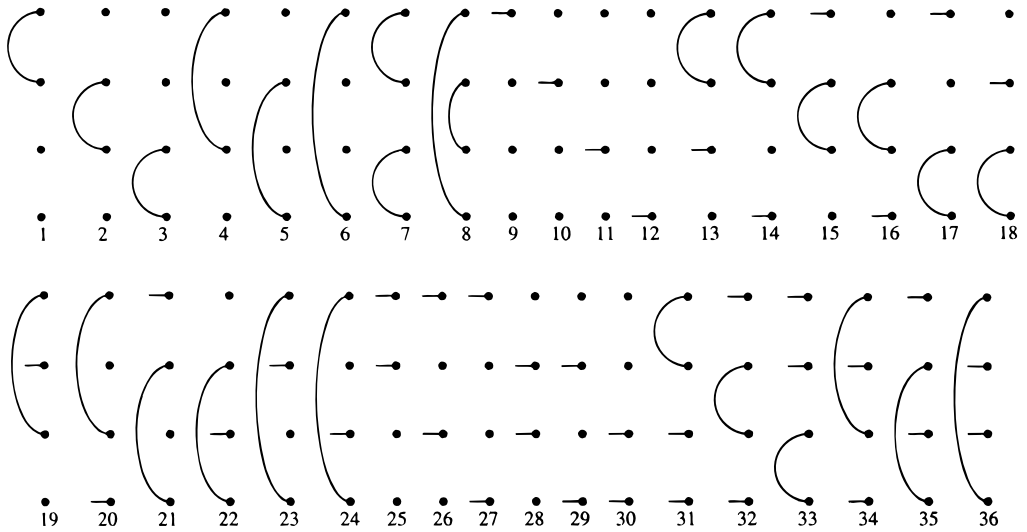

Figure 2. Connectivity states for the enumeration of Hamiltonian chains within 4 × *n* rectangles on the square lattice.

elements v_i, showing if a given connectivity state *i* may form a closed circuit by combining it with the distribution of vertical bonds.

The number N_c of possible Hamiltonian circuits on the rectangle of size $m \times n$ on the square lattice is then given by the simple formula

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

with the superscript T denoting the transpose of vector **u**.

To extend the method to Hamiltonian chains having two ends in two dimensions on the square lattice, we generalize the definition of the connectivity state to include all connectivities with two or fewer ends. Figure 2 shows all possible *nonintersecting* connectivities for 4 points on a line. The connectivities to the ends are represented on the connectivity diagrams as single connected left-sided lines. Similarly, we generalize the definition of the distribution of bonds by also including the ending points of the chain. Distributions at intermediate stages containing the two ends attached to the same single bond (or to a sequence of connected bonds) are not allowed. Also the end points cannot be placed in the middle of a sequence of connected bonds. For 4 points on the line there are 65 different distributions.

The transfer matrix **T** is constructed in the same way as for Hamiltonian circuits by combining connectivity states with bonds and end distributions and finding out the feasible connectivity states of the next neighboring column on the right for this combination. For example the combination of connectivity state 23 from Figure 2 with the bond distribution no. 2 shown in Figure 1b leads to connectivity state 12. Similarly, there are disallowed cases: combinations with unnoccupied sites, triple connections, double connections of the ends, and the formation of loops. Additionally, these combinations cannot lead to the formation of more than two ends or to the disintegration of the chain. The vectors of the starting states **u** and end states **v** are constructed in the same way as for Hamiltonian circuits.

The enumeration of the number of possible Hamiltonian chains is carried out by employing the same equation (eq 1) as for Hamiltonian circuits. The first two columns in Table 1 show examples of enumerations of Hamiltonian chains on a square lattice in a rectangle of size $m \times n$ for $m = 5$ and $m = 8$ and varying *n*.

The method can be generalized readily to Hamiltonian circuits and Hamiltonian chains in three dimensions. For simplicity, we consider compact self-avoiding cubic lattice walks within the shape of a parallelepiped of size $l \times m \times n$. In three dimensions, states are defined within planes instead of lines as for two dimensions. We define the connectivity states as all possible connectivities above the plane of points within the rectangle of size $l \times m$. In this three-dimensional case, the condition that connectivites cannot intersect (as in 2D for the square lattice) no longer holds and the number of possible states becomes larger than in two dimensions. Similarly to the square lattice, we construct all possible distributions of bonds but now within the cross-section $l \times m$ of the rectangle, and subject to the constraints that loops and sites directly connected to more than two bonds not be allowed. To illustrate the problem, let us consider Hamiltonian circuits for the simplest case of parallelepipeds of size $2 \times 2 \times n$. Figure 3a shows all possible connectivity states for the 2×2 cross-section, and Figure 3b all possible distributions of bonds. Only the first 6 connectivity states in Figure 3a contribute to the formation of closed circuits. States 7 and 8 are not feasible in $2 \times 2 \times n$ circuits because of parity, nor is state 9 because there is not enough room for such a crossover. The transfer matrix **T** is constructed in a way similar to that for circuits in two dimensions on the square lattice by combining all compatible connectivity states and bond distributions and specifying the resulting connectivity states. Similarly to the 2D case, the combinations leading to unoccupied sites, triple connections, and the formation of small loops are not allowed. The vectors of starting states **u** and ending states **v** are constructed in a way similar to that above. The number of possible Hamiltonian circuits N_c is again given by eq 1.

The transfer matrix method can also be readily generalized for Hamiltonian chains in three dimensions on the cubic lattice. Similarly, we generalize the connectivity states and the bond distributions within the $l \times m$ rectangle by allowing up to two ends. For example, within the $2 \times 2 \times n$ parallelepipeds, for Hamiltonian chains there are 37 possible connectivity states and 105 bond distributions. The transfer matrix **T** and the vectors of starting and ending states are constructed using the same rules as for the square lattice. The number of possible Hamiltonian chains

Table 1. Numbers of Hamiltonian Chains N_c on the Square Lattice within Rectangles of Size $m \times n$ for Variable Length n and $m = 5$ (Column 1) or $m = 8$ (Column 2)[a]

n	m=5	m = 8	2×3 circuit	2×3 chain
2	22	58	22	584
3	132	1,578	324	16,880
4	1,006	38,984	4,580	413,484
5	4,324	602,804	64,558	9,044,404
6	26,996	12,071,462	908,452	183,901,520
7	109,722	175,905,310	12,788,368	3,554,023,752
8	602,804	3,023,313,284	180,011,762	66,143,917,612
9	2,434,670	43,551,685,370	2,533,935,102	1,195,800,854,580
10	12,287,118	682,958,971,778	35,668,766,942	21,126,320,132,860
11	49,852,352	9,735,477,214,522	502,089,257,086	366,320,570,734,408
12	237,425,498	144,397,808,917,246	7,067,628,303,570	6,254,157,199,388,224
13	969,300,694	2,033,155,413,979,838	99,487,032,472,186	105,394,020,380,350,960
14	4,434,629,912	29,105,375,742,858,518	1,400,423,058,519,806	1,756,465,401,885,524,148
15	18,203,944,458	404,654,754,079,984,324	19,712,968,529,886,538	28,993,901,551,356,778,604
16	80,978,858,522	5,656,098,437,704,094,140	277,488,381,860,243,258	474,632,414,080,513,778,980
17	333,840,165,288	77,710,312,229,803,403,554	3,906,048,038,321,346,590	7,713,216,553,000,476,231,108
18	1,456,084,764,388	1,067,886,114,091,399,967,842	54,983,243,535,983,979,930	124,540,316,930,590,423,294,540
19	6,021,921,661,718	14,517,649,840,508,475,301,004	773,968,225,697,327,416,538	1,999,358,896,478,780,039,051,264
20	25,904,211,802,080	196,974,144,293,101,997,656,968	10,894,715,841,895,810,724,334	31,933,091,746,629,675,832,535,900

[a] Columns 3 and 4 show the numbers of Hamiltonian circuits and Hamiltonian chains N_c, respectively, on the cubic lattice within parallelepipeds of size $2 \times 3 \times n$ for variable height n.

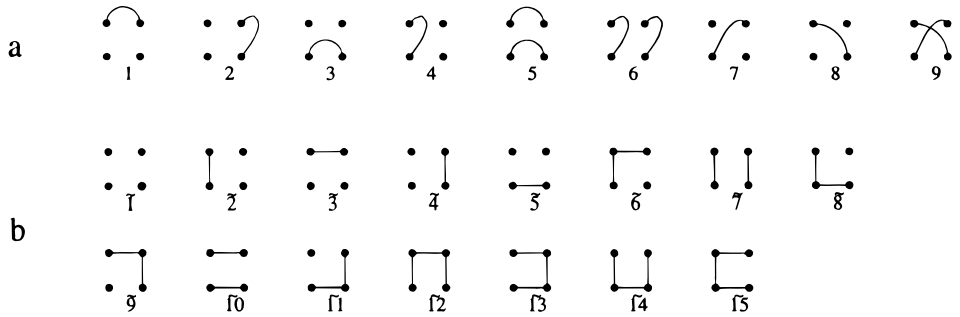

Figure 3. Connectivity states (a) and bond distributions (b) for the generation of Hamiltonian circuits within $2 \times 2 \times n$ parallelepipeds on the cubic lattice.

within the parallelepiped is again calculated by using eq 1.

As an illustration of the method, we show in columns 3 and 4 in Table 1 the counts of the Hamiltonian circuits and Hamiltonian chains for parallelepipeds of size $2 \times 3 \times n$ with variable height n.

The construction of the transfer matrices is being done automatically, so it is easy to generalize the method to allow voids or to extend it to irregular shapes. The application of the transfer matrix method to other types of lattices is also possible.

The exact enumeration of the number of possible conformations within simple geometries is a simple strict criterion for determining the correctness of the method. We have first enumerated all possible Hamiltonian circuits and Hamiltonian walks within various small-sized rectangles and parallelepipeds by using the traditional self-avoiding walk method. Then we compared these enumerations with results of the transfer matrix method, and the results were always identical.

A major advantage of the transfer matrix method is that, once the transfer matrix is calculated, enumerations can be carried out easily for any length n. The main difficulty of the present method is the rapidly growing memory requirements. Consequently, complete enumerations, especially for large numbers of points in 3D on the cubic lattice, become difficult. We can, however, use the transfer matrix method for random sampling of the conformational space. By randomly choosing the connectivity states and the bond distributions, we can generate a sample of compact conformations without attrition. Of course, the simple random choice of the connectivity states and the bond distributions in the Monte Carlo method may produce biased sampling, and probably a more refined approach is needed to avoid biased sampling in the generation of chains. This problem requires further studies.

The proposed Monte Carlo method may also be used for the generation of noncompact random coil chains (random self-avoiding walks). The method may allow

for the very fast generation of random coil chains by growing them column by column in 2D (or plane by plane in 3D) without attrition. The traditional methods use the linear growth of the chain, and the attrition is a main obstacle to the efficient generation of long chains. The extension of the method to off-lattice chains is also possible, because the approach is based on the most fundamental mathematical properties of the connectivity of the chain.

Another major advantage of this method is that we can easily reduce the number of conformations by imposing constraints on the generated chains. For example, we may fix one or both ends of the chain, or fix positions of some secondary structure elements of proteins, such as α helices or β sheets. Such externally imposed constraints significantly reduce the conformational space and enable a complete generation of conformations for longer chains. The transfer method is powerful in this regard, since we can directly fix various structural elements.

We can easily generalize the transfer matrix method used for counting compact Hamiltonian walks to the generation of compact conformations themselves. The uniqueness of a conformation is specified by an alternating sequence of bond distributions and connectivity states.

The transfer matrix method can be easily generalized to irregular protein shapes. Equation 1 is then replaced by

$$N_c = \mathbf{u}^T(\mathbf{T}_2\mathbf{T}_3\mathbf{T}_4...\mathbf{T}_{n-1})\mathbf{v} \qquad (2)$$

where the individual transfer matrices $\mathbf{T}_i$ for transitions between the ith and the $(i + 1)$th columns are specified to conform to the shape.

The transfer matrix method can be a powerful tool for studying protein folding. It is extremely rapid since, instead of linearly growing the conformation, we build the whole conformation column by column (or plane by plane in 3D), completely avoiding attrition. If each site within a column (plane) were assigned a "letter" of the amino acid alphabet, then, assuming that we have only contact interactions between nonbonded neighbors, all calculations of energy with the addition of a new column (plane) can be performed immediately and some blocks of columns of higher energy might be discarded on the fly.

The new approach presented in this paper ought to become a standard useful method for studying globular proteins because of its large advantages over previous methods, including a recently published combinatorial algorithm based on two-matching and patching of bipartite graphs.[14]

One important application would be to obtain an estimate of the chain entropies of proteins. An interesting finding is that entropies per site, for infinitely long strips ($n \rightarrow \infty$) defined as[13]

$$\ln \kappa_w = \lim_{n \rightarrow \infty} \frac{\ln N_{n \times w}}{n \times w} \qquad (3)$$

where $N_{n \times w}$ is the number of compact conformations within the strip of size $n \times w$, and w is the number of sites in the cross-section (i.e., $w = m$ for rectangles in 2D, and $w = l \times m$ for parallelepipeds in 3D), are different (for the same cross-section w) for chains and circuits in 3D, while in 2D they are the same. For example, in the case of parallelepipeds of the size $2 \times 3 \times n$ shown in Table 1, the calculated value of $\kappa_{2\times3}$ is 1.570 051 for chains and 1.553 873 for circuits. In the case of rectangles on the square lattice of size $6 \times n$, κ_6 equals 1.334 651 for both chains and circuits.

Other applications of the transfer matrix method to optimization problems such as the traveling salesman problem are also possible.

Due to the space limit of the Communication, the more detailed explanation of the method and application to many other problems will be presented in forthcoming papers.

Acknowledgment. We are grateful to the National Research Council for the Senior Research Fellowship awarded to A. Kloczkowski.

References and Notes

(1) Madras, N.; Slade, G. *The Self-Avoiding Walk*; Birkhauser: Boston, 1993; pp 393−394.
(2) Sokal, A. D. In *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*; Binder, K., Ed.; Oxford University Press: New York, 1995; pp 47−124.
(3) Chan, H. S.; Dill, K. A. *J. Chem. Phys.* **1989**, *90*, 492; **1990**, *92*, 3118; **1991**, *95*, 3775.
(4) Skolnick, J.; Kolinski, A. *Science* **1990**, *250*, 1121.
(5) Shakhnovich, E. I.; Gutin, A. M. *J. Chem. Phys.* **1990**, *93*, 5967.
(6) Shakhnovich, E. I.; Gutin, A. M. *Nature* **1990**, *346*, 773.
(7) Hinds, D. A.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 2536.
(8) Covell, D. G.; Jernigan, R. L. *Biochemistry* **1990**, *29*, 3287.
(9) Bahar, I.; Jernigan, R. L. *Biophys. J.* **1994**, *66*, 454, 467.
(10) Pande, V. S.; Joerg, C.; Grosberg, A. Y.; Tanaka, T. *J. Phys. A* **1994**, *27*, 6231.
(11) Klein, D. J. *J. Stat. Phys.* **1980**, *23*, 561.
(12) Derrida, B. *J. Phys. A* **1981** *14*, L5.
(13) Schmalz, T. G.; Hite, G. E.; Klein, D. J. *J. Phys. A* **1984**, *17*, 445.
(14) Ramakrishnan, R.; Pekny, J. F.; Caruthers, J. M. *J. Chem. Phys.* **1995**, *103*, 7592.

MA970662H