# Enumeration of cubic lattice walks by contact class

Gordon M. Crippen
*College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065*

(Received 28 December 1999; accepted 30 March 2000)

Self-avoiding walks on a three-dimensional (3D) simple cubic lattice are often used to model polymers, especially proteins. The Hamiltonian is generally taken to be a function of contacts between sequentially nonadjacent residues. The set of all conformations having a particular set of contacts occupies the same energy level, and one would like to estimate the degeneracy or chain entropy of the level. Degeneracies observed in an exhaustive enumeration of short chain configurations are fitted to simple empirical formulas depending on the length of the chain, the number of contacts, and statistics related to the particular set of contacts. © *2000 American Institute of Physics.* [S0021-9606(00)51524-4]

## INTRODUCTION

A very simple way to model polymers, particularly proteins, is as a self-avoiding walk on a cubic lattice, where each of the $k$ residues is represented as a point, and sequentially adjacent residues are on adjacent lattice points. The number, $N$, of such walks is used to calculate the configurational entropy of the chain. To model various physical systems, different constraints have been imposed on the lattice walks, such as unconstrained,[1] closed loops,[2] or confined to restricted volumes.[3] For short chains, exhaustive enumeration[1] can be used, but for long chains, Monte Carlo methods are often employed.[4] For minimalist statistical mechanical models of protein folding, the Hamiltonian is usually defined to be a sum over residue–residue contacts, where residues $i$ and $j$ are in contact if they occupy adjacent lattice points but their sequence separation $|i-j|\geqslant 3$. Thus the set of chain conformations having exactly certain contacts and no others, all have the same energy. To calculate the free energy of some macroscopic state consisting of one or more of such contact classes, one need only evaluate the energy of a single representative of each class and estimate the number of self-avoiding walks in the classes.

In this study we exhaustively enumerate all self-avoiding cubic lattice walks for short chains, and sort them into classes according to the number of residues $k$ and the list of contacts. In order to extrapolate to longer chains, these results are fitted to simple empirical expressions.

## RESULTS

Lattice walks are enumerated on an infinite cubic lattice, subject only to the self-avoiding constraint and that they are unique up to a rigid translation, rotation, and mirror reflection. Thus, in Table I the total number of self-avoiding walks, $N_w$, is 2 for $k=3$, corresponding to the straight and bent conformations, rather than the 30 walks enumerated by Sykes *et al.*[1] that includes six different positions for the second residue and four different positions for the third residue in the bent conformation. For $k=4$, the six walks include only one nonplanar conformation and not its mirror image. Figure 1 shows that the trend is very smooth and linear for $k>4$, fitting

$$\ln N_w = -4.48 + 1.514k. \tag{1}$$

Of course, the number of self-avoiding walks having $k$ steps and no contacts, $N_0$, has a somewhat smaller scaling exponent. Taking into account the curvature in Fig. 1 for small $k$, the empirical equation

$$\ln N_0 = -5.808\,49 + 1.462\,66k + 7.1365/k - 2.787\,42/k^2 \tag{2}$$

fits the data for $1\leqslant k\leqslant 16$ with a standard deviation of 0.03 log units.

The next level of detail is to take into account the number of contacts, $c$, without regard to their arrangement along the chain, as summarized in Table I, columns 4–11. (The $N_8$ column is missing because there are no self-avoiding cubic lattice walks having eight contacts for $1\leqslant k\leqslant 12$.) Since Eq. (2) is a good fit for $\ln N_0$ that can be confidently extrapolated to somewhat greater $k$, we fit the difference, $\ln N(k,c) - \ln N(k,0)$, for the 49 nonzero entries $1\leqslant k\leqslant 12$ and $0\leqslant c \leqslant 9$ from Table I to a functional form that reduces to Eq. (2) when $c=0$,

$$\ln N(k,c) - \ln N(k,0) = c(0.143\,726 - 14.2448/k^2 - 15.2842c/k^2). \tag{3}$$

The standard deviation of the fit is 0.23 log units. One should not attach a lot of significance to the magnitudes of the coefficients or the particular types of terms because they were selected automatically from ten simple functional forms by forward stepwise linear regression.[5] Figure 2 shows the fit for three values of $c$. When Eq. (3) predicts $\ln N(k,c)<0$, it is equivalent to predicting $N(k,c)=0$. Most of the errors in the fit arise for many contacts relative to the chain length, whereas the behavior for longer chains having few contacts is smoother. This still falls short of addressing the original motivation for the study, because when modeling het-

 

TABLE I. For chain length $k$, the number of self-avoiding lattice walks total, $N_w$, and numbers having $c$ contacts, $N_c$.

| $k$ | $N_w$ | $N_0$ | $N_1$ | $N_2$ | $N_3$ | $N_4$ | $N_5$ | $N_6$ | $N_7$ | $N_9$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 6 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 22 | 16 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 92 | 57 | 27 | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 402 | 218 | 128 | 52 | 4 | 0 | 0 | 0 | 0 | 0 |
| 8 | 1832 | 854 | 602 | 270 | 103 | 0 | 3 | 0 | 0 | 0 |
| 9 | 8453 | 3432 | 2812 | 1446 | 646 | 99 | 18 | 0 | 0 | 0 |
| 10 | 39 640 | 13 856 | 12 954 | 7578 | 3597 | 1413 | 160 | 82 | 0 | 0 |
| 11 | 186 296 | 56 522 | 59 276 | 38 473 | 20 222 | 8741 | 2460 | 506 | 96 | 0 |
| 12 | 881 147 | 230 340 | 268 043 | 191 154 | 110 762 | 51 136 | 22 287 | 5388 | 1964 | 73 |
| 13 | 4 162 866 | 943 077 | | | | | | | | |
| 14 | 19 721 230 | 3 852 153 | | | | | | | | |
| 15 | 93 250 730 | 15 773 323 | | | | | | | | |
| 16 | 441 549 914 | 64 430 202 | | | | | | | | |

eropolymers with specified sequences, the energy of a conformation depends on which contacts are formed, not just how many.

More detailed classification of walks according to contact patterns is much harder to fit. For example for $k=6$, there are nine walks having only the single contact 1-4, but only one walk having the single contact 1-6. The discrete combinatorics of the cubic lattice gives rise to much more complicated restrictions than just requiring that $|i-j|$must be odd for any contact. Defining classes of conformations in terms of an exact set of contacts implies that avoiding other contacts is sometimes a significant constraint. For example, there are 844 walks of 12 residues having exactly the three contacts 1-4, 1-6, and 9-12, but there is only one walk of the same chain length having the three contacts 1-6, 1-10, and 5-12 (see Fig. 3). Because the number of contact classes increases so rapidly with chain length, exact enumerations were restricted to $1 \leqslant k \leqslant 12$, as shown in Table I. Various schemes for fitting the data were tried, such as exploiting the general trend that contacts closing small loops are less restrictive than ones with large sequence separations. Thus, for the number of walks of $k$ steps and a particular set $\{c\}$ of $c$ contacts,

$$\ln N(k,c,\{c\}) = 0.3612 + 0.1005k - 0.044\,62c + 2.422k_0 + 0.2394k_1 + 0.1162k_2 - 3.035k_0/k + 0.4384k_3/k - 1.181f \quad (4)$$

fits the data for the 11 908 contact patterns in the range $1 \leqslant k \leqslant 12$ with a standard deviation of 0.81 log units. The terms $k_0, k_1, k_2, k_3$ refer to the numbers of residues in the chain falling within the ranges of 0–3 contacts, respectively. In addition, $f$ is the total number of residues on either end of the chain that are unconstrained by contacts. For example, if $k=12$, $c=2$, and $\{c\}=\{2-5,7-10\}$, then $k_0=4$, $k_1=8$, $k_2=k_3=0$, and $f=3$.

Figure 4 shows the fit of Eq. (4) to the observed numbers of walks for the different contact patterns. The scatter in the plot shows a better fit would require a more detailed description of the contact set than the parameters used in Eq. (4). The improved description would not simply involve more

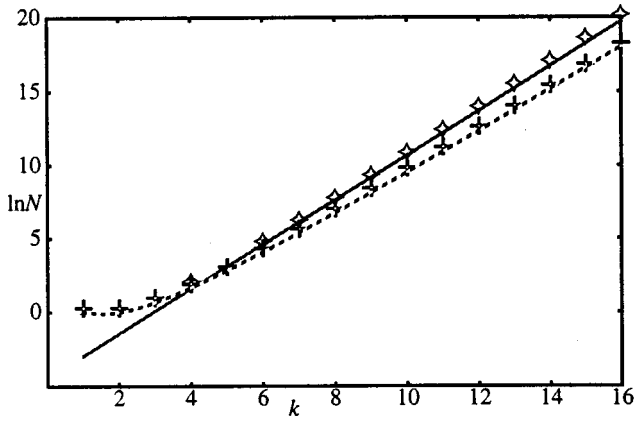


FIG. 1. Total number of self-avoiding walks, $N_w$ (diamonds), the fit to $N_w$ by Eq. (1) (solid curve), the number of walks having no contacts, $N_0$ (crosses), and the fit to $N_0$ by Eq. (2) (dotted curve).

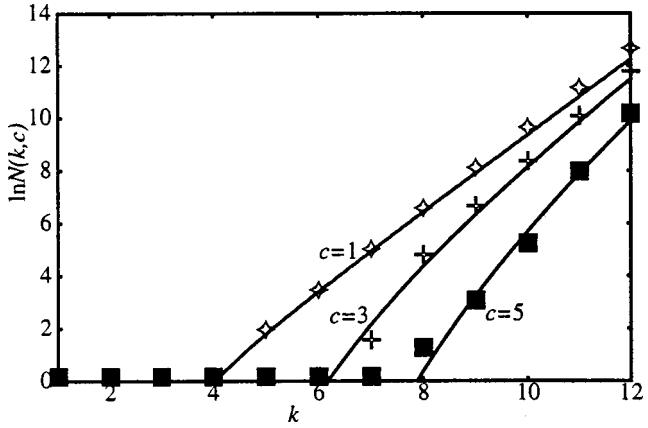


FIG. 2. Use of Eq. (3) for calculating $N(k,c)$, the total number of self-avoiding walks of length $k$ having $c$ contacts (Table I) for $c=1,3$, and 5.

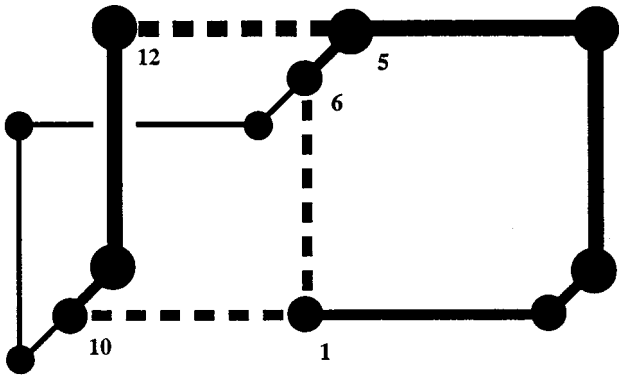


FIG. 3. An example for $k=12$ residues of a class of self-avoiding lattice walks having only three specified contacts (dashed lines) and yet only a single configuration.

terms like $k_n$ for $n>3$, because the contact set for Fig. 3 is poorly fit, yet it involves no higher order features of this type. It is clear that Eq. (4) provides only a convenient way to estimate the much more complicated function $\ln N(k,c,\{c\})$ for $1\leqslant k\leqslant 12$. In order to test whether Eq. (4) provides any useful extrapolation to longer chain lengths, $\ln N_0$ was calculated by Monte Carlo for $k=25, 50, 100,$ and 200. The Rosenbluth and Rosenbluth[6] Monte Carlo calculation of chain entropy agrees very well with the values obtained from exhaustive enumeration, and the error bars are scarcely visible in log plots such as Figs. 5 and 6. Note that for $c=0$, Eq. (4) reduces to

$$\ln N_0 \approx -2.6738+1.3415k, \tag{5}$$

which turns out to be the slight underestimate shown in Fig. 5 (standard error 5.6 log units). For long chains covering the sizes of small to moderate proteins, a better fit is $\ln N_0 \approx -4.41+1.40k$. Any least squares fit is the consequence of its training set, so although Eq. (4) was trained on all $\ln N(k,c,\{c\})$ for $1\leqslant k\leqslant 12$, it is surprising that it comes anywhere near fitting $\ln N_0$ for longer chains.

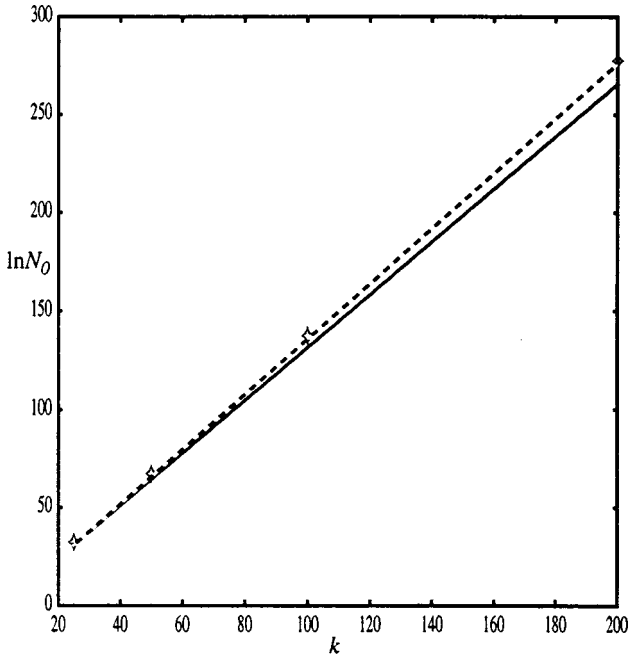


FIG. 5. $\ln N_0$ for $k=25$, 50, 100, and 200 as determined by Monte Carlo. The solid line is the extrapolation from Eq. (4), and the dotted line is a least squares fit to the four points.

In order to test Eq. (4) for longer chains and $c=1$, 2, and 3, Monte Carlo determinations of $\ln N$ were made for $k=50$ and some 20 different choices of $\{c\}$. Clearly it is impractical to evaluate the number of walks for all possible choices of contacts for such a chain length, so the sampling shown in Fig. 6 is by no means exhaustive, nor is it supposed to be representative. The main point is that even at four times

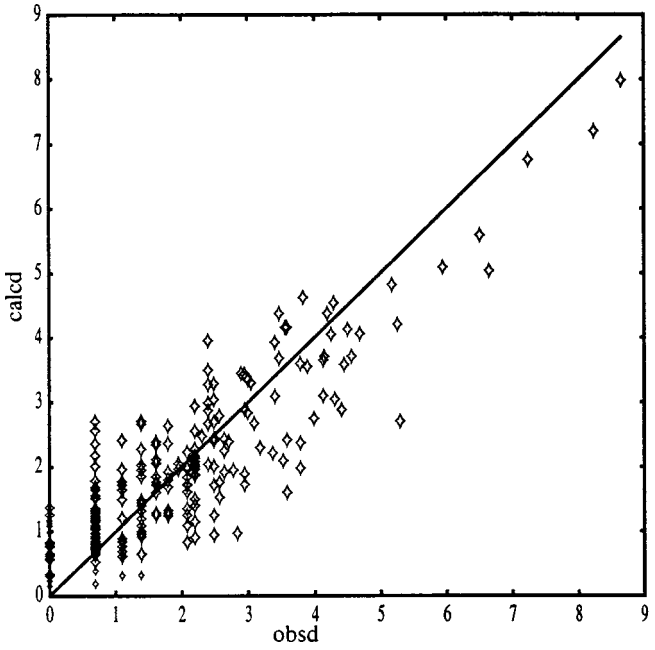


FIG. 4. A log-log plot of the number of walks observed in an exhaustive enumeration vs the calculated number [Eq. (4)]. Points are shown for every 50th set of contacts out of the full list of 11 908 patterns.

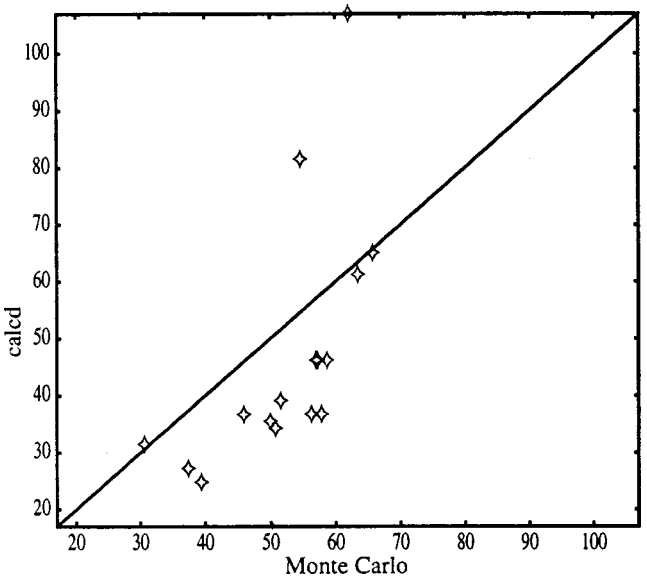


FIG. 6. A log-log plot of the number of walks for $k=50$ and various choices of contacts $c=1$, 2, 3, as determined by Monte Carlo vs the calculated number [Eq. (4)].

the chain length of the cases used to determine Eq. (4), it provides overestimates and underestimates and a few close approximations. The underestimates arise from a few contacts spanning 10 or 20 chain points but leaving free ends ($f \gg 0$), and the one substantial overestimate was from $\{c\} = \{1-4,47-49\}$.

In summary, it is possible to develop a simple regression expression that approximates the number of self-avoiding walks on a cubic lattice having a given chain length and exactly a given set of close contacts. Although this provides only an approximate fit for short chains, it can be used as a convenient estimate for longer chains at similar error levels. In the process, some interesting examples have been found of large deviations from the regression fit. These may provide clues for a more insightful expression and a better fit.

## ACKNOWLEDGMENTS

This work was supported by NSF Grant No. DBI-9614074, NIH Grant No. GM-59097, and the Vahlteich Research Award Fund.

[1] M. F. Sykes, A. J. Guttmann, and P. D. Roberts, J. Phys. A **5**, 653 (1972).
[2] M. F. Sykes, D. S. McKenzie, M. G. Watts, and J. L. Martin, J. Phys. A **5**, 661 (1972).
[3] A. Kloczkowski and R. L. Jernigan, J. Chem. Phys. **109**, 5147 (1998).
[4] D. L. Zhao, Y. Huang, Z. R. He, and R. Y. Qian, J. Chem. Phys. **104**, 1672 (1996).
[5] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear Regression Models* (Irwin, Chicago, 1996).
[6] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).