DIMACS Technical Report 99-19
April 1999

Statistical Properties of Contact Maps

by

Balakrishna Subramanian [1] and Joel Lebowitz ,[2]
Dept. of Mathematics and Physics.
Rutgers University
New Brunswick, New Jersey 08854

Michele Vendruscolo and Eytan Domany [3]
Department of Physics of Complex Systems
Weizmann Institute of Science
Rehovot 76100, Israel

Ido Kanter
Dept. of Physics
Bar Ilan University
52900 Ramat Gan, Israel

[1]Affliated Graduate Student.
[2]Permanent Member
[3]DIMACS Visitor support

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs, Bellcore and NEC Research Institute.

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

Statistical Properties of Contact Maps

Michele Vendruscolo[(1)], Balakrishna Subramanian[(2)], Ido Kanter[(3)], Eytan Domany[(1)] and Joel Lebowitz[(2)]
[(1)] *Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel*
[(2)] *Department of Mathematics, Rutgers University, New Brunswick NJ 08903*
[(3)] *Department of Physics, Bar Ilan University, 52900 Ramat Gan, Israel*

A contact map is a simple representation of the structure of proteins and other chain-like macromolecules. This representation is quite amenable to numerical studies of folding. We show that the number of contact maps corresponding to the possible configurations of a polypeptide chain of N amino acids, represented by $(N-1)$-step self avoiding walks on a lattice, grows exponentially with N for all dimensions $D > 1$. We carry out exact enumerations in $D = 2$ on the square and triangular lattices for walks of up to 20 steps and investigate various statistical properties of contact maps corresponding to such walks.. We also study the exact statistics of contact maps generated by walks on a ladder.

PACS numbers: 87.15.By, 87.10.+e, 5.50.+q

I. INTRODUCTION

Prediction of a protein's structure from its amino acid sequence is an important and challenging open problem. The first choice one has to make when approaching the problem is that of *structure representation.* One of the most minimalist representations of a protein's structure is in terms of its *contact map* [1,2] which, for a polypeptide chain of length $N-1$, is an $N \times N$ matrix $\mathbf{S}$. Denoting by i, j the position index of the amino acids along the chain, the elements of $\mathbf{S}$ are defined as

$$S_{ij} = \begin{cases} 1 & \text{if amino acids } i \text{ and } j \text{ are in contact} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

"Contact" can be defined in various ways: for example [3], one can set $S_{ij} = 1$ when there exist two heavy (all but hydrogen) atoms, one from amino acid i and one from j, separated by less than some threshold distance. Contact maps are independent of the coordinate frame used and for compact structures, such as the native state of proteins, with many contacts, it is relatively easy to go from a map to a set of possible structures to which it may correspond [2,4,5]. On a lattice, a protein conformation, or fold, is represented as a self-avoiding random walk (SAW) [6]. A site on the lattice visited by the walk represents an amino acid. Two sites of the SAW are in contact if they are nearest neighbors and they are *non-consecutive* along the walk.

To search for a protein's native structure in the space of contact maps (as has been proposed by several groups), it is important to have general knowledge about the size and nature of this space. Recent studies [7,8] of the dynamics of naturally occurring proteins has shown that the contact maps along with simple energetics is enough information to reproduce the vibrational spectrum with some accuracy. This makes it important to understand the statistics of the contact map representation. In particular, one would like to know how the number of different *physical* contact maps depends on the chain length N. Clearly one has $2^{N(N+1)/2}$ distinct $N \times N$ symmetric matrices of binary elements $S_{ij} = 0, 1$. Most of these, however, do not correspond to physical structures; these matrices cannot be realized as contact maps of real, physical chains or SAW's.

In fact, as we shall see, N_M, the total number of physical maps obtainable for a chain of length N on a lattice satisfies the bounds

$$e^{c_\ell N} \le N_M \le e^{c_u N} \ . \tag{2}$$

The upper bound in (2) follows trivially from the bound on N_{SAW}, the total number of SAW's, which is asymptotically given by [9]

$$N_M \leq N_{SAW} \sim N^{\gamma-1}\mu^N \sim e^{c_u N}\ , c_u = \log\mu, \tag{3}$$

where μ is the connectivity constant of the lattice, and γ is a critical exponent.

A simple construction of a special set of walks, each with a distinct contact map provides the lower bound. Start the chain at the the origin, $i = 1$; the first step and all odd-indexed steps are in the positive horizontal direction $+x$, whereas every even indexed step is in the vertical direction, either $+y$ or $-y$. The decision taken for step $2i$ either brings the site $2i+1$ into contact with $2i-2$, in which case $S_{2i-2,2i+1} = 1$, or this contact is absent and $S_{2i-2,2i+1} = 0$. Hence for every choice of the set of vertical steps we get a walk whose contact map does not appear for any different walk from the set. Clearly, the maps constructed this way must have $S_{2i-k,2i+1} = 0$ for $k > 1$. In this way we obtain N'_{SAW} SAW's and the following exponential lower bound for the number of contact maps

$$N_M \geq N'_{SAW} \sim 2^{N/2} \sim e^{c_\ell N}\ . \tag{4}$$

Clearly the argument works for any dimension and can be extended for the triangular lattice. This (rather poor) bound can be improved by including walks whose maps can have other non-vanishing elements, e.g. with $S_{2i-4,2i+1} = 1$. A better lower bound for the square lattice is obtained by an explicit construction given in Section 3.

We actually expect that $\log(N_M)/N$ approaches a limit,

$$\frac{\log(N_M)}{N} \to a \tag{5}$$

as N becomes large (the existence of a limit does not follow directly from (2)). To estimate a we computed, for $N \leq 20$, the precise numbers N_M on the square and triangular lattices. This is done by exact enumeration of all possible distinct SAW's, i.e. not related by symmetry operations of the lattice, and the corresponding contact maps.

Using this enumeration for low N and sampling for larger N, we also computed various other statistics of contact maps, such as the number of maps with particular density of contacts, the number of SAW's that correspond to this set of maps, etc. We also obtained explicit results about the corresponding quantities for walks on a special "ladder" lattice.

II. EXACT ENUMERATION IN $D = 2$

In the upper curve of Fig. 1 we plot the number of walks N_{SAW}, obtained by complete enumeration [11], versus N, fitted (for $N \leq 25$) with the known [9] estimates $\mu = 2.6381585$ for the connective constant and $\gamma = 43/32$ for the critical exponent on the $2D$ square lattice. The lower curve is the total number N_M of contact maps, corresponding to all possible SAW's with $N \leq 20$ on a $2D$ square lattice. Fitting Eq.(2), we obtain $a = 0.83(1)$. This result was obtained previously, by enumerating walks with $N \leq 14$, by Chan and Dill [10]. For comparison, we note that a straightforward fit of N_{SAW} with Eq.(3) gives the upper bound prefactor $c_u = 1.00(1)$, and that the lower bound prefactor, as from Eq.(4), is $c_\ell = 0.346$. We obtained the corresponding results for the triangular lattice. But in this case due to the higher density of contacts, we were able to obtain results only for $N \leq 11$ as shown in Fig. 1. Our fit gave $c_u = 1.47(1)$ and $a = 1.28(1)$. To address the question whether in $D = 2$ the constant a for the contact maps is strictly less than $\log(\mu)$, we present in Fig. 2 a plot of the running value of the connective constant μ for the walks and the running value of $\exp(a)$ for the maps as the size of the walk increases. The running slope $\mu(N)$ is computed from enumeration data using the standard formula

$$\log\mu(N) = \frac{1}{2}\left[\log N_{SAW}(N+1) - \log N_{SAW}(N-1)\right]\ , \tag{6}$$

and an analogous one for the factor a for maps.

This figure is consistent with $a < \log\mu$.

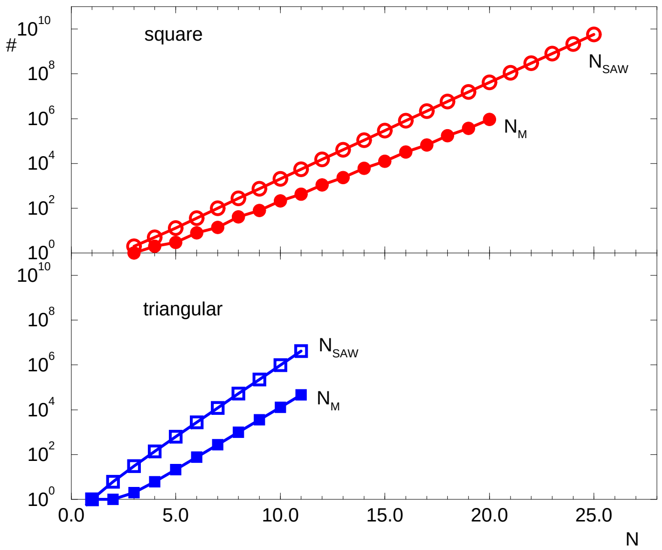

FIG. 1. Upper curve: N_{SAW}, the number of SAW's versus their length N, obtained by exact enumeration for $N \leq 25$ on the $2D$ square lattice and fitted with Eq. (2). The lower curve shows the exponential variation of N_M, the number of contact maps corresponding to all possible SAW's with $N \leq 20$. Data were obtained from complete enumeration.

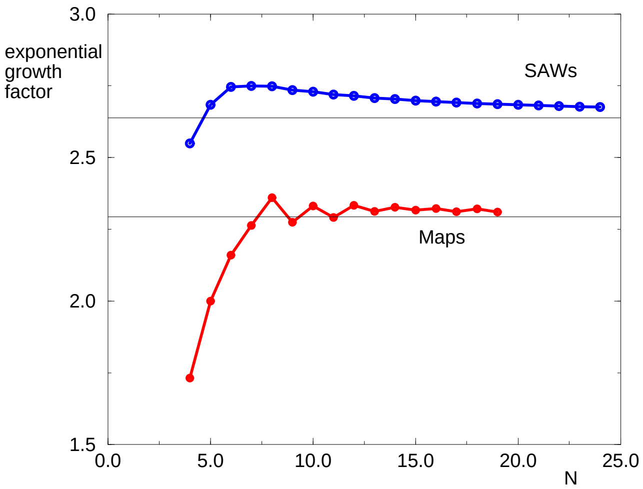

FIG. 2. A comparison of the connective constant μ for the walks and the exponential growth factor $\exp{(a)}$ for the contact maps generated on a square lattice evolving with the size of the walk. Horizontal lines are the known value $\mu = 2.638$ [9] and our estimate $e^a = 2.3$, based on the data for all N's.

Most biologically functional proteins fold into remarkably compact conformations, with very few solvent molecules in the interior. Therefore it is of interest to consider how the number of contact maps and their corresponding walks varies with the number of contacts.

Denote by $N_{SAW}(c)$ the number of walks with a fixed number Nc of contacts. When there is an interaction energy u associated with each contact, then $\log(N_{SAW})$ is identical to the entropy of the chain at energy $E = Ncu$. In Fig. 3 we plot the fractions

$$n_{SAW}(c) = \log(N_{SAW}(c)/N) \tag{7}$$

for chains of different lengths N on the $2D$ square lattice.

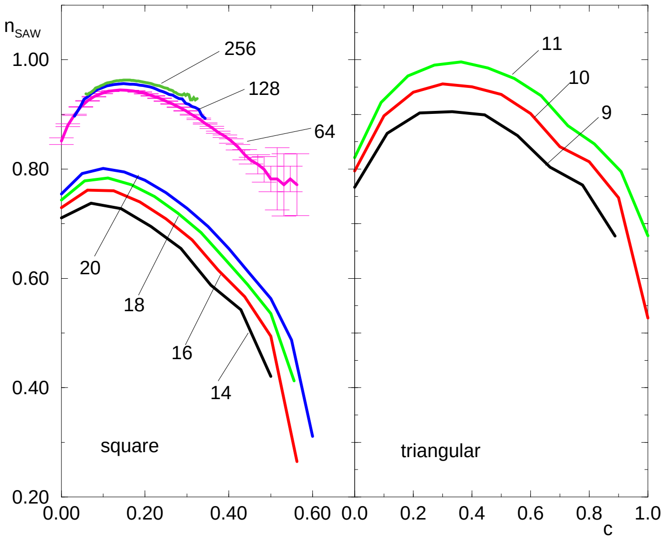

FIG. 3. Logarithm $n_{SAW}(c)$ of the fraction of walks with a given fraction c of contacts. On the square lattice, we show data obtained from exact enumeration for chain lengths $N = 14, 16, 18, 20$, and data obtained from sampling for $N = 64, 128, 256$. For clarity, errors on data from sampling are shown only for $N = 64$. On the triangular lattice, we show the fraction $n_{SAW}(c)$ of walks with a given number Nc of contact for $N = 9, 10, 11$.

The time required to enumerate walks and maps increases exponentially with the size N and it becomes impractical to use this method. However we want to generate the statistics for larger values of N, which is the actual physical situation. Standard techniques are routinely used [9] to generate unbiased samples of SAWs on the lattice. We use the method of incomplete enumeration (Redner-Reynolds) to generate our sample of unbiased SAW's.

We use our sample to generate the distribution of the fraction of the walks with a given number of contacts $n_{SAW}(Nc)$ as introduced before for SAW's of length N=64,128,256. In Fig. 3, we plot the result.

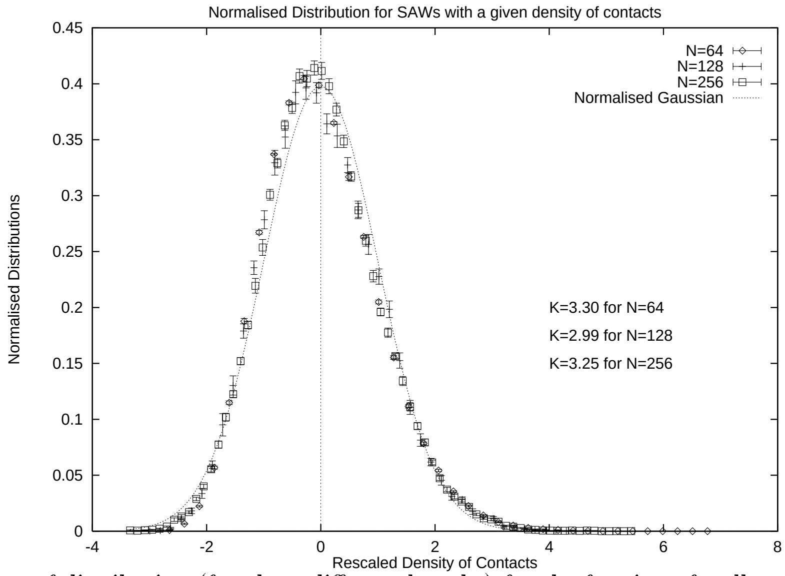

FIG. 4. The collapse of distributions(for three different lengths) for the fraction of walks with a given number of contacts after rescaling the finite-size variables.

One would like to say something about how this distribution looks in the asymptotic limit. We try to analyze this by rescaling the finite-size variables such that the distributions collapse on top of each other. If Fig. 4, we observe that normalizing the variance to 1 and the mean to 0, results in the collapse of the distributions (for the three different lengths of N=64, 128,256). We compare this to the normalized Gaussian. From the data obtained, it appears that we cannot rule out either possibility (Gaussian or non-Gaussian). We also list the kurtosis values K obtained for the different data sets. For a Gaussian distribution, we expect an exact value of 3.00.

It is however not clear how one should generate the distribution of the maps with a given number of contacts $N_M(Nc)$. While we have standard and efficient algorithms to generate SAWs with the desired weight, it seems difficult to generate contact maps which are equally weighted in the sample and not biased with their degeneracies.

Let now $N_M(c)$ be the number of distinct contact maps with Nc contacts. We show in Fig. 5 how the fractions

$$n_M(c) = \log\left(N_M(c)/N\right) \tag{8}$$

vary with c, again for various chain lengths.

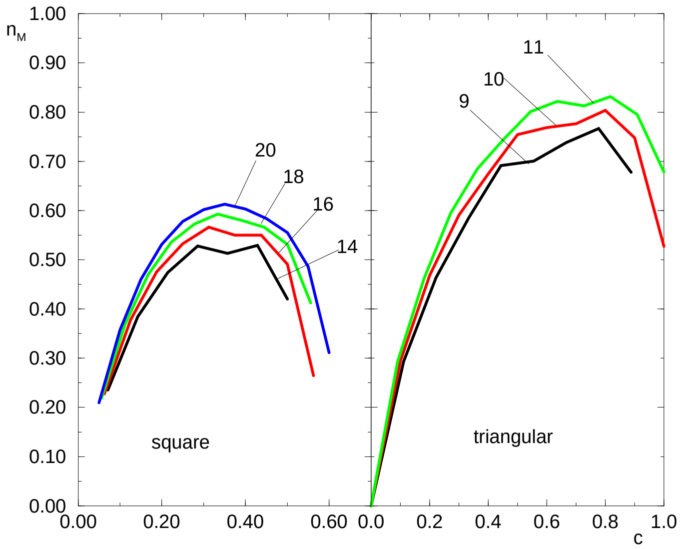

FIG. 5. Logarithm $n_M(c)$ of the fraction of contact maps with a given fraction c of contacts, shown for 4 different walk lengths on the 2D square lattice, and for 3 different walk lengths on the triangular lattice.

The main difference between Figs. 3 and 5 is that the distribution of walks has its maximum at smaller values of c than the distribution of contact maps. This can be understood intuitively by a noting that for small c the number of maps is suppressed in comparison to $N_{SAW}(c)$: for example, there is only a single contact map with $c = 0$, whereas there are many walks with no contacts. In general, the degeneracy of contact maps has a non-trivial dependence on the number of their contacts.

Consider walks of length N and denote by

$$G = e^{Ng} \tag{9}$$

the degeneracy of a map, i.e. the number of walks corresponding to that contact map. For each value of g we determined $H(g)$, the number of maps whose degeneracy is g. This information is shown in Fig. 6 where we present $h(g) = \log H(g)/N$ versus g, for walks of length $N = 20$ on the square lattice. We further analyze the degeneracy by concentrating on the *subset of maps* with a fixed number of contacts, Nc. In Fig. 6 we show results for $Nc = 3, 4, 5, 6, 9$, i.e. $c = 0.15, 0.2, 0.25, 0.3, 0.45$. Not surprisingly, the maps with large number of contacts which correspond to the typical native folds of proteins generally have small degeneracy. It is the maps with few contacts which account for the large degeneracy. In general the map with $c = 0$ (all zero entries in the matrix) has $G > 2^N$ corresponding to all the directed walks with no contacts. The walks that correspond to maps with different degeneracies differ in the lengths of contact-free segments that the walk has. For $N = 20$ and $Nc = 6$ on the square lattice, we measured the length L of the longest contact-free stretch at the ends of the walk. Maps with low degeneracy have, on the average, $L \simeq 1$, whereas for highly degenerate maps we found, typically, $L \simeq 7$ (there are also highly degenerate maps and walks with long contact-free stretches far from the ends). Clearly, the presence of long stretches free of contacts is responsible for the high degeneracy of a map.

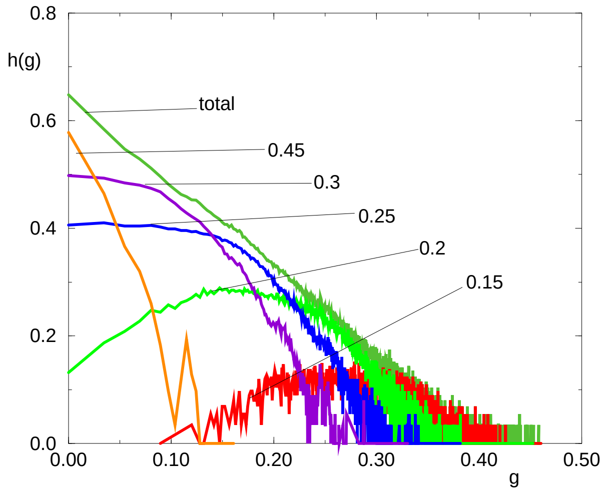

FIG. 6. Histogram of the degeneracy; $h(g) = \log H(g)/N$, where $H(g)$ is the number of maps with degeneracy $G = e^{Ng}$, for walks of length $N = 20$ on the square lattice. Separate curves are shown for subset of maps with $c = 0.15, 0.2, 0.25, 0.3, 0.45$.

Let now $\bar{G}(Nc)$ denote the average degeneracy over all the maps with Nc contacts. We studied $\bar{G}(Nc)$ as a function of Nc. As already mentioned, contact maps corresponding to maximally compact walks have, on the average, a very small degeneracy. It seems reasonable to assume that for a fixed $c = \frac{N_c}{N}$, $\bar{G}(Nc)$ will grow exponentially with N, such that

$$\log \bar{G}(N, Nc) = \alpha N f(c) \ . \tag{10}$$

The enumeration results seem to support this assumption as seen in the collapse plot in Fig. 7 with $\alpha = 0.86$ for the square lattice and $\alpha = 1.07$ for the triangular lattice. The value of α is extracted by fitting $\bar{G}(N, 0) \sim e^{\alpha N}$. As we can see the assumption Eq. (10) seems to hold to good accuracy.

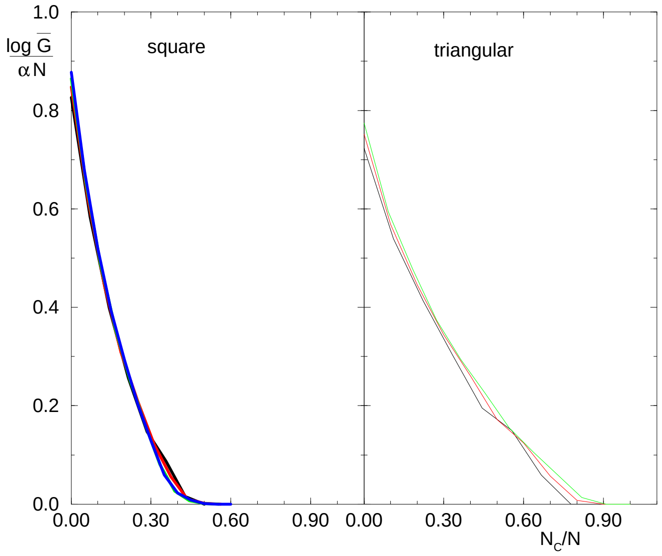

FIG. 7. Scaling plot of the degeneracy function $\bar{G}(c)$ averaged over all the contact maps with Nc contacts, plotted for different chain lengths N on the triangular and square lattices.

III. EXACT RESULTS FOR WALKS AND MAPS ON A LADDER

In this section, we introduce and solve the problem exactly on a toy lattice. The lattice is a ladder of two rows of sites, at points (x, y) with $y = 0, 1$ and $x = 0, 1, 2, \ldots$.. We consider all walks starting at the origin with horizontal steps in the positive x direction. We first show that the numbers of SAW's and contact maps is exponential in N, with different coefficients a. Denote by $A(N)$ the number of walks of N steps;

$$A(N) = A_h(N) + A_v(N)$$

where $A_h(N)$ is the number of walks that end with a horizontal step and $A_v(N)$ walks end with a vertical step. Since a vertical step *must* be preceded by a horizontal one we have

$$A_v(N) = A_h(N-1)$$

On the other hand, to every walk one can add a horizontal step so that

$$A_h(N) = A(N-1)$$

Thus we get, using these three relationships, the recursion for the Fibonacci numbers:

$$A(N) = A(N-1) + A(N-2) \tag{11}$$

and hence the number of walks grows, for large N, exponentially

$$A(N) \sim e^{a_w N} \qquad\qquad a_w = \ln\frac{1+\sqrt{5}}{2} \approx 0.481 \tag{12}$$

A recursion for the number of contact maps can be calculated as well. One way to do this is by representing $B(N)$, the total number of distinct contact maps of N steps as a sum

$$B(N) = B_0(N) + B_1(N)$$

where $B_0(N)$ is the number of contact matrices (maps) whose first row contains only zeroes (i.e. the first site does not have a contact); $B_1(N)$ is the number of those maps for which the first site does have a contact. Since to every map we can add a first row (and column) of zeroes, we have

$$B_0(N) = B(N-1)$$

For all maps that start with a contact, the first four steps are fixed; the corresponding walks can be continued in two different ways, either with a vertical step or with a horizontal one. These two possibilities give rise to a recursion of the form

$$B_1(N) = B_1(N-2) + B(N-5)$$

With a little algebra the last three equations yield the final recursion

$$B(N) = B(N-1) + B(N-2) - B(N-3) + B(N-5) \tag{13}$$

If we now assume that

$$\frac{\log(B(N))}{N} \rightarrow e^{a_m}$$

as N becomes large, we find that e^{a_m} is the solution of the equation

$$q^5 - q^4 - q^3 + q^2 - 1 = 0$$

which yields

$$a_m \approx 0.367 < a_w$$

Having counted the number of walks and maps, we turn to calculate various statistical features of maps and walks on a ladder. For example, we can consider the fraction of maps with a given number of contacts; the degeneracy of maps, i.e. the number of different walks that have the same map,etc. Analytical examination of such quantities sheds light on the origin of results obtained from exact enumeration of walks in two dimensions and indicate the extent to which the relatively short chains that can be enumerated represent the true two dimensional behavior.

A walk of N steps taken according to the rules given above can be characterized by the sequence of the contact-free intervals between all pairs of consecutive contacts. We denote by m the number of steps needed to walk from the end site of contact $n-1$ to the end site of contact n. Let $D(m)$ denote the degeneracy of such a contact-free walk, i.e. the number of different SAWs of length m. To calculate $D(m)$, we introduce a transition matrix L, among six possible "states", of pairs of consecutive steps, referred to as "2-steps". The six possible 2-steps that can occur on a ladder are shown in Fig. 8.

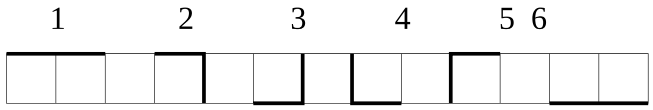

FIG. 8. The six possible two successive steps on a ladder.

The fact that $L_{42} = 1$ shows that it is possible to have a 3-step walk whose first and second steps constitute a 2-step is of type 2 and the second and third step constitute a 2-step of type 4. Note that only those transitions that do not terminate the walk (i.e. do not generate a contact) are designated as possible by the matrix L - for example we have $L_{34} = 0$ since a 4 followed by a 3 generates a contact.

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \tag{14}$$

Note that to have a contact by adding 2-step n, the 2-step $n-1$ must be either of type 4 or 5; the corresponding vectors are $V_4 = (000100)$ and $V_5 = (000010)$. The degeneracy of walks of length m in between contacts is then given by

$$D(m) = [(V_4)^T + (V_5)^T] L^{m-1} V_2 \quad or \quad [(V_4)^T + (V_5)^T] L^{m-1} V_3 \tag{15}$$

The possible lengths for m are 2, 5, 6, 7, ..., and the corresponding degeneracies are given $1, 1, 1, 1, 2, 3, 4, 6, 9, 13, 19, 28, 41, 60....$ Note that $D(100) \sim 10^{16}$ and asymptotically

$$D(m) \propto (1.465)^m = e^{0.382m} \tag{16}$$

where 1.465 is the largest eigenvalue of the matrix L.

An $N \times N$ contact map is completely specified by the set of inter-contact intervals $\{m\}$. If for a given map an interval of length m appears $N(m)$ times, denote

$$P(m) = N(m)/N$$

The logarithm of the number of SAW associated with this particular map is then given by

$$\log N_{SAW}(\{m\}) = N \sum_m P(m) \log D(m) \tag{17}$$

The number of contacts of this map is given by

$$N_c(\{m\}) = \sum_m N(m) = N \sum_m P(m) = Nc \tag{18}$$

where the number of contacts per step,

$$c = \sum_m P(m) = \frac{N_c}{N} \tag{19}$$

The normalization of the $P(m)$ is

$$\sum_m P(m)m = 1 \tag{20}$$

The number of maps, N_M, characterized by the same set of fractions $\{P(m)\}$ (with different orderings of the contact-free intervals) is

$$\log N_M = -N \sum_m P(m) \log P(m) + Nc \log c \tag{21}$$

and therefore the number of SAW, N_W, associated with all maps characterized by the same fractions $\{P(m)\}$ is

$$\log N_{SAW}(P(m)) = -N \left\{ \sum_m P(m)[\log P(m) - \log D(m)] - c \log c \right\} \tag{22}$$

The interplay between these two terms is clear. As the distance m between contacts increases, the number of SAW corresponding to such a map increases exponentially, but at the same time the number of contacts in the map decreases and the number of such maps (permutation of the distances) decreases exponentially. Some limiting cases can be analyzed as follows. For the case densest with contacts, i.e. $c = 0.5$, there are only two possible maps and hence $log(N_{SAW})/N \to 0$. On the other hand, for maps with $O(1)$ contacts, and hence $c \to 0$, m scales with N and $D(N) \propto e^{0.382N}$, and therefore $N_{SAW} \sim e^{0.382N}$. Since in both limiting cases $log(N_{SAW})$ does not scale as $0.481N$ (see eq. (12)), the quantity $\log(N_{SAW})$ is expected to have a maximum at some intermediate value of c.

The number of SAW associated with maps that have Nc contacts can be studied analytically;

$$N_{SAW}(c) = \int_0^1 \pi dP(m) \delta \left[\sum_m mP(m) - 1 \right] \delta \left[\sum_m P(m) - c \right] e^{-N\{\sum_m P(m)[\log P(m) - \log D(m)] - c \log c\}} \tag{23}$$

The integrals are evaluated by the saddle point method; the resulting equations can be reduced to the following coupled equations for $P(2)$ and $P(5)$

$$\begin{aligned} 1 &= P(2) \sum_m D(m)[P(5)/P(2)]^{(m-2)/3} m \\ c &= P(2) \sum_m D(m)[P(5)/P(2)]^{(m-2)/3} \end{aligned} \tag{24}$$

where for every allowed $m = 2, 5, 6, 7, ...$, the degeneracy $D(m)$ is determined by eq. (15); these are supplemented by

$$P(m) = P(2)D(m)[P(5)/P(2)]^{(m-2)/3} \tag{25}$$

The numerical solution of the saddle point equations gives $\frac{1}{N} \log N_{SAW}$ as a function of c, the concentration of contacts, is presented in Fig. 9. The maximum $\frac{1}{N} log(N_{SAW}) = 0.481$, as expected, is obtained for $c \sim 0.105$.

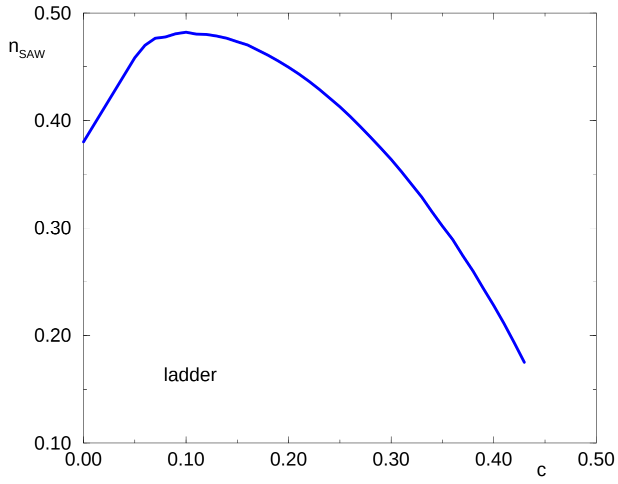

FIG. 9. $n_{SAW} = \log\left(N_{SAW}/N\right)$ versus c, for maps of Nc contacts

One can calculate $\log(N_M)$ as a function of c in a similar fashion. All one has to do is to set $D(m) = 1$ in eq. (23); the resulting saddle point equations are obtained from eq. (24- 25), by using there, again, $D(m) = 1$.

The numerical solution for $\frac{1}{N}\log(N_M)$ as a function of c, with the trivial end points $(0, 0)$ and $(0.5, 0)$, are presented in Fig. 10.

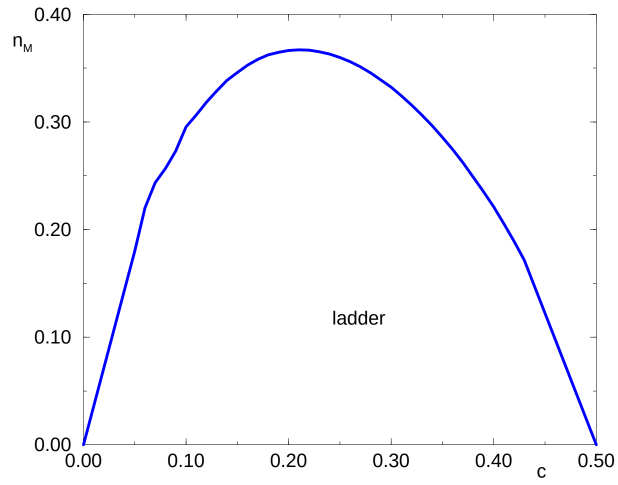

FIG. 10. $n_M = \log\left(N_M/N\right)$ versus c; Nc is the number of contacts

The final property of walks and maps on a ladder that we calculate deals with the degeneracy of a map with Nc contacts. Denote by $G = e^{Ng}$ the number of walks that have the same map and by $H(g, c)$ the number of maps of Nc contacts and this value of the degeneracy. The quantity $H(g, c)$ is given by

$$H(g,c)=\int_0^1 \pi dP(m)\delta\left[\sum_m P(m)\log D(m)-g\right]\delta\left[\sum_m mP(m)-1\right]\delta\left[\sum_m P(m)-c\right]e^{-N\left[-\sum_m P(m)\log P(m)-c\log c\right]} \tag{26}$$

and the saddle point equations for $\{P(m)\}$ are

$$P(m)=P(2)D(m)\left[\frac{P(2)P(8)}{P(5)^2}\right]^{[\log D(m)/\log 2]}\left(\frac{P(5)}{P(2)}\right)^{(m-2)/3} \tag{27}$$

The three unknown fractions $P(2),\ P(5),\ P(8)$ are determined through the three global constraints

$$c=\sum_m P(m),\qquad 1=\sum_m P(m)m,\qquad g=\sum_m P(m)\log D(m)$$

A typical result for $c=0.2$ are presented in Fig. 11.

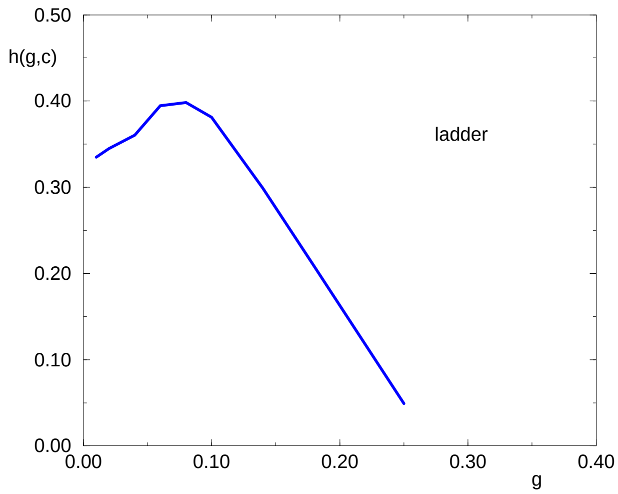

FIG. 11. Histogram of degeneracy: $h(g,c)=\log H(g,c)/N$ versus g, for $c=0.2$ on a ladder.

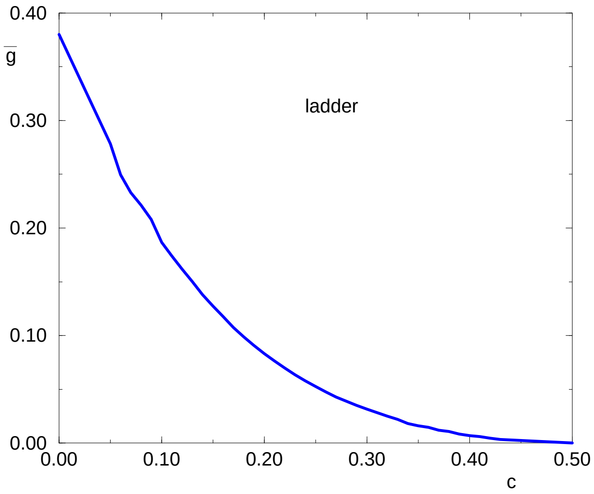

FIG. 12. Histogram of average degeneracy $\overline{g}$ versus c on a ladder.

IV. SEMIDIRECTED RESTRICTED WALKS

A related problem is that of semidirected restricted walks (SRW) on a square lattice. These walks are defined as follows: all horizontal steps are directed - in the $+x$ direction. Vertical steps are restricted so that the number of consecutive vertical steps never exceeds k. The k=1 case is already a superset of walks on a ladder.

The number of SRWs can be computed as follows.

Denote the total number of walks by $A(N)$. As before, $A_h(N)$ of these walks end with a horizontal step and $A_v(N)$ walks end with a vertical step.

$$A(N) = A_h(N) + A_v(N)$$

$A_v(N)$ can be further classified into k classes. ${A^i}_v(N)$ corresponds to walks that end with exactly i vertical steps.

$$\begin{aligned} A_v(N) = {A^1}_v(N) + {A^2}_v(N) + \cdots + {A^k}_v(N) \\ {A^i}_v(N) = {A^{(i-1)}}_v(N-1) \\ {A^1}_v(N) = 2A_h(N-1) \end{aligned}$$

A little algebra gives the following recursive relation

$$A(N) = A(N-1) + 2\left(A(N-2) + \cdots + A(N-k-1)\right)$$

So the connective constant (of exponential growth) is given by the root of the following polynomial equation:

$$(y-1)y^k = 2\frac{1-y^k}{1-y}$$

For $k=1$ this reduces to $(y-1)y=2$, i.e. $y=2$, whereas in the $k\to\infty$ limit it simplifies to $(y-1)^2=2$ so that the connective constant increases to $y=1+\sqrt{2}\approx 2.42$.

Computing the number of contact matrices for a general k seems slightly more tedious, but it is possible to do it explicitly for $k=1$. We denote $B(N)$ by the number of distinct maps of size N. It can be classified into maps with either one contact or no contact in the first row. The number of the former is $B_0(N)$ and the latter $B_1(N)$.

$$
\begin{aligned}
B(N) &= B_0(N) + B_1(N) \\
B_0(N) &= B(N-1) \\
B_1(N) &= B(N-4) + B_1(N-2)
\end{aligned}
$$

A little algebra leads to the following recursive relation:

$$B(N) = B(N-1) + B(N-2) - B(N-3) + B(N-4)$$

which, in turn, leads to the following polynomial equation:

$$q^4 - q^3 - q^2 + q - 1 = 0$$

The root, $q\approx 1.51$, corresponding to the growth factor for the maps, is slightly higher than that of the ladder (≈ 1.44). We have not found a simple way to compute $B(N)$ for general k values.

V. SUMMARY

Contact maps are a compact and useful representation of a protein's structure. Contact maps are used for screening candidate structures from a database. More recently attempts were made to use them to fold proteins, i.e. determine the map of a protein of known sequence by minimizing some energy function.

In order to have a handle on the work involved in searching the subspace of physical maps, it is important to know various statistics. For example, how the number of physical maps increases with the protein's length, the dependence of various properties on the number of contacts, etc. In this paper we studied these issues on several lattices; for an essentially one-dimensional ladder the results were obtained analytically and in two dimensions we studied the square and triangular lattices by exact enumeration and sampling. In addition we provide exact bounds on the number of distinct physical maps, valid in any dimension.

Our main findings can be summarized as follows:

- The number of physical contact maps scales exponentially with the length N of the walk.
- The number of contact maps (and of walks as well) is a non-monotonic function of the number of contacts.
- The average degeneracy of contact maps that have N_C contacts decreases as N_C increases.
- Contact maps corresponding to very compact walks (i.e. highest N_C) have low degeneracy. The ground state of a protein is most likely to be found among these maps.

Acknowledgments

This work was initiated during a visit of ED to Rutgers, supported by DIMACS. BS and JL thank J. Kahn and O. Penrose for very helpful comments and discussions. BS and JL wish to acknowledge research supported in part by NSF grant and DIMACS. The work of MV and ED was partially supported by a grant from GIF (German Israel Science Foundation), the Israeli Ministry of Science and the Minerva Foundation.

[1] S. Lifson and C. Sander, *Nature* **282**, 109 (1979).
[2] T. F. Havel, G. M. Crippen and I. D. Kuntz, *Biopolymers* **18**, 73 (1979).
[3] L Mirny and E. Domany, *Proteins* **26**, 391 (1996).
[4] G. Crippen and T. F. Havel, *Distance geometry and molecular conformation.* Wiley, New York (1988).
[5] M. Vendruscolo, E. Kussell and E. Domany, *Folding & Design* **2**, 295 (1997).
[6] K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
[7] T. Haliloglu, I. Bahar and B. Erman, *Phys. Rev. Lett.* **79**, 3090 (1997).
[8] I. Bahar, A. R. Atilgan and B. Erman, *Folding and Design.* **2** 173 (1997).
[9] N. Madras, and G. Slade, G., *The Self-Avoiding Walk*, Birkhauser, Boston (1993).
[10] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991). Thirumalai and collaborators recently obtained $a \simeq 1.22$ on the cubic lattice; H. J. Bussemaker, D. Thirumalai and J. K. Bhattacharjee, Phys. Rev. Lett. **79**, 3530 (1997).
[11] D. C. Rapaport, *Comp. Phys. Rep.* **5**, 265 (1987).