Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential Monte Carlo method

Jie Liang and Jinfeng Zhang
Department of Bioengineering, SEO, MC-063, University of Illinois at Chicago, Chicago, Illinois 60607-7052

Rong Chen
Department of Information and Decision Sciences and Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois 60607-7124

(Received 7 March 2002; accepted 22 May 2002)

voids exist in proteins as packing defects and are often associated with protein functions. We study the statistical geometry of voids in two-dimensional lattice chain polymers. We define voids as topological features and develop a simple algorithm for their detection. For short chains, void geometry is examined by enumerating all conformations. For long chains, the space of void geometry is explored using sequential Monte Carlo importance sampling and resampling techniques. We characterize the relationship of geometric properties of voids with chain length, including probability of void formation, expected number of voids, void size, and wall size of voids.

We formalize the concept of packing density for lattice polymers, and further study the relationship between packing density and compactness, two parameters frequently used to describe protein packing. We find that both fully extended and maximally compact polymers have the highest packing density, but polymers with intermediate compactness have low packing density. To study the conformational effects of void formation, we characterize the conformational reduction factor of void formation and found that there are strong end-effect. Voids are more likely to form at the chain end. The critical exponent of end-effect is twice as large as that of self-contacting loop formation when existence of voids is not required. We also briefly discuss the sequential Monte Carlo sampling and resampling techniques used in this study. © 2002 American Institute of Physics.

I. INTRODUCTION

Soluble proteins are well-packed, and their packing densities may be as high as that of crystalline solids. Yet there are numerous packing defects or voids in protein structures, whose size distributions are broad. The volume ($v$) and area ($a$) of protein does not scale as $v \approx a^{3/2}$, which would be expected for models of tight packing. Rather, $v$ and $a$ scale linearly with each other. In addition, the scaling of protein volume and cluster-radius is characteristic of random sphere packing. Such scaling behavior indicates that the interior of proteins is more like Swiss cheese with many holes than tightly packed jigsaw puzzles.

What effects do voids have? Proteins are often very tolerant to mutations, which may suggest potentially stabilizing roles of voids in proteins. Voids in proteins are also often associated with protein function. The binding sites of proteins for substrate catalysis and ligand interactions are frequently prominent voids and pockets on protein structures. However, the energetic and kinetic effects of maintaining specific voids in proteins are not well understood, and the shape space of voids of folded and unfolded proteins are largely unknown.

In this paper, we examine the details of the statistical nature of voids in simple lattice polymers. Lattice models have been widely used for studying protein folding, where the conformational space of simplified polymers can be examined in detail. Despite its simplistic nature, lattice model has provided important insights about proteins, including collapse and folding transitions, influence of packing on secondary structure formation, and designability of lattice structures. However, one drawback is that the lattice model is not well-suited for studying void-related structural features, such as protein functional sites, since it is not easy to model the geometry of voids.

In this article, we first define voids as topological defects and describe a simple algorithm for void detection in the two-dimensional lattice. We then enumerate exhaustively the conformations for all $n$-polymers up to $n = 25$, and analyze the relationship of probability of void formation, expected packing density, and compactness, as well as expected wall interval of void with chain length. To study statistical geometry of long chain polymers, we describe a Monte Carlo sampling strategy under the framework of Sequential Importance Sampling, and introduce the general technique of resampling. The results of simulation of long chain polymers up to $N = 200$ for several geometric parameters are then presented. We further explore the conformation reduction factor $R$ of void formation, and describe the significant end-effect of...
void formation, as well as the scaling law of $R$ and wall interval of voids. In the final section, we summarize our results and discuss effective sampling strategy for studying the conformational space of voids.

II. LATTICE MODEL AND VOIDS

Lattice polymers are self-avoiding walks (SAWs), which can be obtained from a chain-growth model. Specifically, an $n$-polymer $P$ on a two-dimensional square lattice $Z^2$ is formed by monomers $n_i, i \in \{1, \ldots, N\}$. The location $x_i$ of a monomer $n_i$ is defined by its coordinates $x_i = (a_i, b_i)$, where $a_i$ and $b_i$ are integers. The monomers are connected as a chain, and the distance between bonded monomers $x_i$ and $x_{i+1}$ is 1. The chain is self-avoiding: $x_i \neq x_j$ for all $i \neq j$. We consider the beginning and the end of a polymer to be distinct. Only conformations that are not related by translation, rotation, and reflection are considered to be distinct. This is achieved by following the rule that a chain is always grown from the origin, the first step is always to the right, and the chain always goes up at the first time it deviates from the $x$-axis. For a chain polymer, two nonbonded monomers $n_i$ and $n_j$ are in topological contact if they intersect at an edge that they share. If two monomers share a vertex of a square but not an edge, these two monomers are defined as not in contact.

When the number of monomer $n$ is 8 or more, a polymer may contain one or more void [Fig. 1(a)]. We define voids as topological features of the polymer. The complement space $Z^2 - P$ that is not occupied by the polymer $P$ can be partitioned into disjoint components,

$$Z^2 - P = V_0 \cup V_1 \cdots \cup V_k.$$ 

Here $V_0$ is the unique component of the complement space that extends to infinity. We call this the outside. The rest of the components that are disjoint or disconnected to each other are voids of the polymer. Because nonbonded monomers intersecting at a vertex are defined as not in contact, they do not break up the complement space. As an example, the unfilled space contained within the polymer in Fig. 1(b) is regarded as one connected void of size 4 rather than two disjoint voids of size 2. This choice is arbitrary, but is consistent with the definition of contact. A simple algorithm for void detection can be found in the Appendix.

III. VOID DISTRIBUTION BY EXACT ENUMERATION

A. Probability of forming voids and expected number of voids

The number of conformations $\omega(n)$ for an $n$-polymer up to $n = 25$ obtained by exhaustive enumeration is shown in Table I. The numbers of conformations for polymers up to

| $n$ | $\omega(n)$ | $\omega_0(n)$ | $\omega_1(n)$ | $\omega_2(n)$ | $\omega_3(n)$ | $\omega_4(n)$ |
|-----|-------------|---------------|---------------|---------------|---------------|---------------|
| 3   | 2           | 2             | 0             | 0             | 0             | 0             |
| 4   | 5           | 5             | 0             | 0             | 0             | 0             |
| 5   | 13          | 13            | 0             | 0             | 0             | 0             |
| 6   | 36          | 36            | 0             | 0             | 0             | 0             |
| 7   | 98          | 98            | 0             | 0             | 0             | 0             |
| 8   | 272         | 270           | 2             | 0             | 0             | 0             |
| 9   | 740         | 734           | 6             | 0             | 0             | 0             |
| 10  | 2034        | 1993          | 41            | 0             | 0             | 0             |
| 11  | 5513        | 5393          | 120           | 0             | 0             | 0             |
| 12  | 15037       | 14508         | 529           | 0             | 0             | 0             |
| 13  | 40617       | 39078         | 1536          | 3             | 0             | 0             |
| 14  | 110188      | 104566        | 5602          | 20            | 0             | 0             |
| 15  | 296806      | 280599        | 16088         | 119           | 0             | 0             |
| 16  | 802075      | 748335        | 53149         | 591           | 0             | 0             |
| 17  | 2155667     | 2002262       | 151052        | 2353          | 0             | 0             |
| 18  | 5808335     | 5327888       | 470386        | 10051         | 10            | 0             |
| 19  | 15582342    | 14222389      | 1325590       | 34287         | 76            | 0             |
| 20  | 41889578    | 37784447      | 3973361       | 131298        | 472           | 0             |
| 21  | 11221246    | 100637371     | 11119456      | 416239        | 2680          | 0             |
| 22  | 301100754   | 267136710     | 32470871      | 1471874       | 12293         | 6             |
| 23  | 80557061    | 71063806      | 90361878      | 4479355       | 54998         | 24            |
| 24  | 2158326727  | 1883960171    | 259195774     | 14946910      | 223458        | 414           |
| 25  | 5768299665  | 5005591512    | 717505892     | 44337381      | 862748        | 2132          |
As the chain length grows, it is clear that both with chain length. The expected number of voids only by strongly connected unoccupied sites, namely, any void \( v \) and \( D \) is calculated as

\[
\pi_v = \frac{\sum_{k=1}^{5} \omega_k(n)}{\omega(n)}.
\]

The expected number of voids \( \bar{n} \) for a polymer is

\[
\bar{n} = \frac{\sum_{k=1}^{5} \omega_k(n) \cdot k}{\omega(n)}.
\]

As the chain length grows, it is clear that both \( \pi_v \) and \( \bar{n} \) increases [Figs. 2(a) and 2(b)].

**B. Void size**

The total size \( v \) of voids in a polymer is the sum of the sizes of all voids, namely, the total number of all unoccupied squares that are fully contained within the polymer. Let \( \omega_v(n) \) be the number of conformations of \( n \)-polymer with total void size \( v \). The expected total void size \( \bar{v} \) for the \( n \)-polymer is

\[
\bar{v} = \frac{\sum_v \omega_v(n) \cdot v}{\omega(n)}.
\]

Figure 2(c) shows that the expected void size \( \bar{v} \) increases with chain length \( n \).

**C. Wall size of void**

For a void \( V \) of size \( v \), what is the required minimum length \( l(V) \) for a polymer that can form such a void? Equivalently, what is the size of the wall of the polymer containing void \( V \)? Here we first restrict our discussion to voids formed only by strongly connected unoccupied sites, namely, any neighboring two sites of a void must be sharing at least one edge of the squares. We exclude voids containing weakly connected sites, where two neighboring sites are connected by only one shared vertex [Fig. 1(b)]. For \( v = 1, 2, \) and 3, it is easy from the geometry of the voids to see that \( l(v) = 8, 10, \) and 12, respectively. However, in general \( l(v) \) also depends on the shape of the void. A void of size 4 can have five different shapes. If the void is of the shape of a 2 \( \times \) 2 square, \( l(4) = 12 \). For the other four shapes, \( l(4) = 14 \).

For any strongly connected void, we find that the following general recurrence relationship for \( l(v) \) holds:

\[
l(v) = l(v-1) + \begin{cases} 
2, & \text{if } \Delta \partial V = 3 \\
0, & \text{if } \Delta \partial V = 2 \\
-1, & \text{if } \Delta \partial V = 1 
\end{cases}
\]

where \( \partial V \) represents the boundary edges of void \( V \), and \( \Delta \partial V \) represents the net gain in the number of boundary edges introduced by the newly added unoccupied site. Although the number and explicit shapes of strongly connected voids of size up to 5 can be found in Ref. 29, there is no general analytical formula known for the number of shapes of a void of size \( v \). This is related to the problem of determining the number of polyominos or animals (as in percolation theory) of a given size.

When weakly connected voids are also considered, there are more possible wall sizes for the void. For 22-mer, the number of different wall sizes observed for a void, strongly or weakly connected, at various size are shown in Fig. 3(a). Voids of size 5 have the largest diversity in wall size. This is of course due to the fixed chain length. A short chain such as the 22-mer has only a small number of ways for forming large voids. Figure 3(b) shows the average wall size for various void sizes in the 22-mer. The expected or average wall size \( \bar{w}(n) \) for a void in an \( n \)-polymer can be calculated as

\[
\bar{w}(n) = \sum_v w \cdot \frac{\omega_{v,w}(n)}{\omega_v(n)},
\]

where \( v \) is the void size, \( w \) is the wall size of the void, \( \omega_{v,w}(n) \) is the number of \( n \)-polymers containing a void of size \( v \) with wall size \( w \), and \( \omega_v(n) \) is the total number of \( n \)-polymers with a void of size \( v \). Figure 2(d) shows that \( \bar{w}(n) \) increases with chain length. Wall size and void size are analogous to the area and volume of voids in three-dimensional space.
when a chemists as a parameter for characterizing protein folding.3
Following this original definition, the packing density 
where
contains the molecule.3,30 It has been widely used by protein
An important parameter that describes how effectively
atoms fill space is the packing density 
Density 0.52 has only 11 conformations, whereas the highest
conformations with 2 and 3 voids.

FIG. 4. Packing density and compactness are two useful parameters describing packing of chain polymers. (a) The expected packing density decreases with chain length. (b) For the 22-mer, the majority of the conformations with 1-void have a high packing density, namely, the size of the void is small. Fewer conformations are found with large voids. The same pattern is observed for conformations with 2 and 3 voids. (c) The expected compactness fluctuates but in general decreases with chain length. (d) The relationship of average packing density \( p \) and average compactness \( \rho \) for the chain polymer of length \( N = 14 \)–\( 22 \).

D. Packing density

An important parameter that describes how effectively atoms fill space is the packing density \( p \). In proteins, it is defined by Richards and co-workers as the amount of the space that is occupied within the van der Waals envelope of the molecule, divided by the total volume of space that contains the molecule.3,30 It has been widely used by protein chemists as a parameter for characterizing protein folding.3 Following this original definition, the packing density \( p \) for the lattice polymer is

\[
p = n/(n + v),
\]
when a \( n \)-polymer has a total void size of \( v \).

The expected packing density \( \bar{p}(n) \) for an \( n \)-polymer can be calculated as

\[
\bar{p}(n) = \sum_p p \frac{\omega_p(n)}{\omega(n)},
\]
where \( \omega(n) \) is the number of all conformations of \( n \)-mer, \( \omega_p(n) \) is the number of \( n \)-mers with packing density of \( p \). The scaling of \( \bar{p}(n) \) with the chain length \( n \) decreases roughly linearly between \( n = 7 \) and \( n = 22 \) [Figure 4(a)]. Because it takes at least two additional monomers to increase the size of a void by one, \( \bar{p}(n) \) decreases only when \( n \) is an odd number for short chains.

Although voids are packing defects, most conformations with voids have high packing density, namely, the total size of the voids is small. Among all conformations of the 22-mer containing one void, the number of conformations increases monotonically with packing density. The lowest packing density 0.52 has only 11 conformations, whereas the highest packing density 0.92 has the largest number (6756751) of conformations [Fig. 4(b)]. A similar relationship is found among conformations with 2 and 3 voids [Fig. 4(b)].

E. Compactness

Another important parameter that measures the packing of the lattice polymer is the number of nonbonded contacts \( t \). It is related to the compactness parameter \( \rho \), defined by Chan and Dill12 as \( \rho = t/t_{\text{max}} \), where \( t_{\text{max}} \) is the maximum number of nonbonded contacts possible for an \( n \)-polymer. Compactness \( \rho \) has been studied extensively in seminal works by Chan and Dill.12,23,31 Although \( \rho \) is sometimes correlated with the compactness \( r \), these two parameters are distinct. The relationship between compactness and expected packing density for chain polymer of length 14–22 is shown in Fig. 4. For all chain lengths, both maximally compact polymer (\( \rho = 1 \)) and extended polymer (\( \rho = 0 \)) have maximal packing density (\( p = 1 \)), but polymers with low packing densities have intermediate compactness on average. Polymers with \( \rho \) between 0.4 and 0.6 have lowest packing density and therefore tend to have larger void size. The explanation is simple. An extended lattice chain polymer has no voids, it therefore achieves maximal packing density of \( p = 1 \), but its compactness \( \rho \) is 0. A maximally compact polymer with \( \rho = 1 \) also contains no voids, its \( p \) is 1. On the other hand, nonmaximally compact polymers can have a range of packing densities.

IV. OBTAINING VOID STATISTICS FOR LONG CHAIN POLYMERS VIA IMPORTANCE SAMPLING

Sequential Importance Sampling: Geometrically complex and interesting features emerge only in polymers of sufficient length, which are not accessible for analysis by exhaustive enumeration, due to the fact that the number of possible SAWs increases exponentially with the chain length. Monte Carlo methods are often used to generate samples from all possible conformations and obtain estimates of feature statistics using those samples. However, when chain length becomes large, the direct generation of the SAWs using the rejection method (i.e., generate random walks on the lattice and only accept those that are self-avoiding) from the uniform distribution of all possible SAWs becomes difficult. The success rate \( s_N \) of generating SAWs decreases exponentially, \( s_N \approx Z_N/(4 \times 3^{N-1}) \). For \( N = 48 \), \( s_N \) is only 0.79%.32 To overcome this attrition problem, a widely used approach is the Rosenbluth Monte Carlo method of biased sampling.26 The task is to grow one more monomer for a \( t \)-polymer chain that has been successfully grown from 1 monomer after \( t-1 \) successive steps without self-crossing, until \( t = n \), the targeted chain length. In this method, the placement of the \((t+1)\)th monomer is determined by the current conformation of the polymer. If there are \( n_t \) unoccupied neighbors for the \( t \)th monomer, we then randomly (with equal probability) set the \((t+1)\)th monomer to any one of the \( n_t \) sites. However, the resulting sample is biased toward more compact conformations and does not follow the uniform distribution. Hence each sample is assigned a “weight” to adjust for this bias. Any statistic can then be obtained from weighted aver-
age of the samples. In the case of the Rosenbluth chain growth method, the weight is computed recursively as $w_t = n_tw_{t-1}$.

Liu and Chen provided a general framework of Sequential Monte Carlo (SMC) methods which extend the Rosenbluth method to more general setting. Sophisticated but more flexible and effective algorithms can be developed under this framework. In the context of growing polymer, SMC can be formulated as follows. Let $(x_1, \ldots, x_t)$ be the position of the $t$ monomers in a chain of length $t$. Let $\pi_1(x_1), \pi_2(x_1, x_2), \ldots, \pi_t(x_1, \ldots, x_t)$ be a sequence of target distributions, with $\pi(x_1, \ldots, x_n) = \pi_n(x_1, \ldots, x_n)$ being the final objective distribution from which we wish to draw inference. Let $g_{t+1}(x_{t+1} | x_1, \ldots, x_t)$ be a sequence of trial distributions which dictates the growing of the polymer. Then we have:

**Procedure SMC**

1. Draw $x_1^{(j)}$, $j = 1, \ldots, m$ from $g_1(x_1)$
2. Set the incremental weight $w_1^{(j)} = \pi_1(x_1^{(j)}) / g_1(x_1^{(j)})$
3. For $j = 1$ to $m$
   - Sampling for the $(t+1)$th monomer for the $j$th sample
     - Draw position $x_{t+1}^{(j)}$ from $g_{t+1}(x_{t+1} | x_1, \ldots, x_t)
     - Compute the incremental weight $u_{t+1}^{(j)} = \frac{\pi_{t+1}(x_{t+1}^{(j)}, x_1^{(j)}, \ldots, x_t^{(j)})}{\pi_t(x_1^{(j)}, \ldots, x_t^{(j)})} \cdot g_{t+1}(x_{t+1}^{(j)}, x_1^{(j)}, \ldots, x_t^{(j)})$
     - Set $w_{t+1}^{(j)} = u_{t+1}^{(j)} \cdot w_t^{(j)}$
4. Resampling
5. At the end, the configurations of successfully generated polymers $(x_1^{(i)}, \ldots, x_t^{(i)})$ and their associated weights $w_n^{(i)}$ can be used to estimate any properties of the polymers, such as expected void size, compactness, and packing density. That is, the objective inference $\mu_n = E[h(x_1, \ldots, x_n)]$ is estimated with $\hat{\mu}_n = \frac{\sum_{j=1}^m h(x_1^{(j)}, \ldots, x_n^{(j)}) \cdot w_n^{(j)}}{\sum_{j=1}^m w_n^{(j)}}$, (1)

for any integrable function $h$ of interests.

The critical choices that affect the effectiveness of the SMC method are (1) the approximating target distribution $\pi_t(x_1, \ldots, x_t)$, (2) the sampling distribution $g_{t+1}(x_{t+1} | x_1^{(j)}, \ldots, x_t)$, and (3) the resampling scheme. In this study, we are interested in sampling from the uniform distribution $\pi_n(x_1, \ldots, x_n)$ of all geometrically feasible conformations of length $n$, which we call the final objective distribution. It can also be chosen to be the Boltzmann distribution when energy function such as the HP model is introduced.

The Rosenbluth method is a special case of SMC. Its target distributions $\pi_n(x_1, \ldots, x_t)$ is the uniform distribution of all SAWs of length $t$. Its sampling distribution $g_{t+1}(x_{t+1} | x_1, \ldots, x_t)$ is the uniform distribution among all $n_t(x_1, \ldots, x_t)$ unoccupied neighboring sites of the last monomer $x_t$, and the weight function is $w(x_1, \ldots, x_t, x_{t+1}) = n_t(x_1, \ldots, x_t) n_1(x_1, \ldots, x_t)$.

When there is no unoccupied neighboring sites ($n_t(x_1, \ldots, x_t) = 0$), there is no place to place the $(t+1)$ th monomer. In this case, the chain runs into a dead end and we declare the conformation dead, with weight assigned to be 0. In the case of Rosenbluth method, no resampling is used.

Similarly, the k-step look ahead algorithm chooses $\pi_{t+1}(x_1, \ldots, x_{t+k})$ being the marginal distribution of $\pi_n^k(x_1, \ldots, x_{t+k})$, the uniform distribution of all SAWs of length $t+k$. Hence $\pi_{t+1}$ is closer to the final objective distribution—the uniform distribution of all SAWs of length $n$. Specifically,

$$\pi_{t+1}(x_1, \ldots, x_{t+k}) = \sum_{x_{t+1}, \ldots, x_{t+k}} \pi_{t+k}(x_{t+1}, \ldots, x_{t+k}) \propto n_k(x_1, \ldots, x_{t+k})$$

where $n_k(x_1, \ldots, x_{t+k})$ is the total number of SAWs of length $t+k$ “grown” from $(x_1, \ldots, x_{t+k})$ [i.e., with the first $(t+1)$ positions at $(x_1, \ldots, x_{t+k})$]. In the k-step look-ahead algorithm, the sampling distribution is $g_{t+1}(x_{t+1} = x_{t+1}) = n_{t+k}(x_1, \ldots, x_k) / n_{t+k+1}(x_1, \ldots, x_k)$.

It chooses the next position according to what will happen $k$ steps later. Namely, the probability of placing the $(t+1)$ th monomer at $x$ is determined by the ratio of the total number of SAWs of length $t+k$ grown from $(x_1, \ldots, x_{t+k})$ and the total number of SAWs of the same length $t+k$ grown from one step earlier $(x_1, \ldots, x_k)$. The corresponding weight function is

$$w(x_1, \ldots, x_{t+k}, x_{t+1}) = \frac{n_k(x_1, \ldots, x_{t+k})}{n_{t+k+1}(x_1, \ldots, x_k)} \frac{n_t(x_1, \ldots, x_{t+k})}{n_{t+k}(x_1, \ldots, x_k)} = \frac{n_{t+k+1}(x_1, \ldots, x_k)}{n_{t+k}(x_1, \ldots, x_k)}$$

Although it has higher computational cost, it usually produces better inference on the final objective distribution, with less “dead” conformations. The standard Rosenbluth algorithm is a 1-step look ahead algorithm.

To compare geometric properties estimated from sequential Monte Carlo method and those obtained by exact enumeration, we examine the expected number of voids and expected void size for polymer of chain length 14–22. Figure 5 shows that sequential Monte Carlo can provide very accurate estimation of these geometric properties of voids. Here 2-step look ahead is used, with Monte Carlo sample size of 100,000 and no resampling is applied.

The resampling step is one of the key ingredients of the SMC method. There are many cases where resampling is beneficial. First, note that it is unavoidable to have some
dead conformations during the growth. These chains need to be replaced to maintain sufficient Monte Carlo sample size. Second, the weight of some chains may become so relatively small that their contribution in the weighted average [Eq. (1)] is negligible. When the variance of the weights is large, the effective Monte Carlo sample size becomes small.33,37,38 Third, for a specific function $h$, its value may become too small (even zero) for some sampled conformations. In all these cases, efficiency can be gained by replacing those conformations with "better" ones. This procedure is called "resampling." There are many different ways to do resampling. One approach is rejection control,39 which regenerates the replacement conformations from scratch. An easier approach is to duplicate the existing and good conformations. Specifically,

**Procedure RESAMPLING**

```plaintext
/m: number of original samples.
/m[(x_1^{(j)},...,x_m^{(j)})]_{j=1}^m: original properly weighted samples

for $j=1$ to $m$
   Set resampling probability of $j$th conformation $\propto \alpha^{(j)}$
endfor

for $j=1$ to $m$
   Draw $j$th sample from original samples
   $\{(x_1^{(j)},...,x_m^{(j)})\}_{j=1}^m$ with probabilities $\{\alpha^{(j)}\}_{j=1}^m$
   //Each sample in the newly formed sample is assigned a new weight.
   //#j-th chain in new sample is a copy of k-th chain in original sample.
   $w^{(k)}_j \leftarrow w^{(k)}_j / \alpha^{(k)}$
endfor
```

In the resampling step, the $m$ new samples $\{(x_1^{(j)},...,x_m^{(j)})\}_{j=1}^m$ can be obtained either by residual sampling or by simple random sampling. In residual sampling, we first obtain the normalized probability $\tilde{\alpha}^{(j)} = \alpha^{(j)} / \Sigma \alpha^{(j)}$. Then $[m \tilde{\alpha}^{(j)}]$ copies of the $j$th sample are made deterministically for $j = 1,...,m$. For the remaining $m - \Sigma [m \tilde{\alpha}^{(j)}]$ samples to be made, we randomly sample from the original set with probability proportional to $m \tilde{\alpha}^{(j)} - [m \tilde{\alpha}^{(j)}]$.

The choice of resampling probability proportional to $\alpha^{(j)}$ is problem specific. For general function $h$, such as the end-to-end extension $\|x_n - x_1\|$, it is common to use $\alpha^{(j)} = w^{(j)}$. In this case, all the samples in the new set have equal weight. When the function is irregular, a carefully chosen set of $\alpha^{(j)}$ will increase the efficiency significantly.

The method of pruning and enriching of Grassberger40 is a special case of the residual sampling, with $\alpha^{(j)} = 0$ for the $k$ chains with zero weight (dead conformations), $\alpha^{(j)} = 2$ for the top $k$ chains with largest weights, and $\alpha^{(j)} = 1$ for the rest of the chains. Residual sampling on this set of $\alpha$ is completely deterministic. The resulting sample consists of two copies of the top $k$ conformations (each of them having half of their original weight) and one copy of the middle $n - 2k$ chains with their original weight. The $k$ dead conformations are removed.

In our study of the relationship between compactness and packing density, we use a more flexible resampling method. Our focus is on the packing density among all conformations within certain range of compactness. In this case, our object target distribution is the uniform distribution among all possible SAW’s with compactness measure falling within a certain interval, i.e., a truncated distribution. Although compactness changes slowly as the chain grows, to grow into a long chain it is possible that the compactness of a chain evolve and cover a wide range during growth. Hence we choose the uniform distribution of all possible SAW’s of length $t$ as our target distribution at $t$, and only select those with the desired compactness at the end for our estimation of the packing density. In order to have larger number of usable samples (i.e., to achieve better acceptance rate) at the end, we encourage growth of chains with desirable compactness through resampling. Specifically,

**Procedure RESAMPLING (m,d,c_i)**

```plaintext
// m: Monte Carlo sample size, d: steps of looking back.
// c_i: targeting compactness.
// $k$: number of dead conformations.
Divide $m - k$ samples randomly into $k$ groups.
for group $i = 1$ to $k$
   Find conformations not picked in previous $d$ steps.
   //Pick the best conformation $P_j$, for example $P_j$ = polymer with min $|c - c_j|$.
   Replace one of the $k$ dead conformations with $P_j$.
   Assign both copies of $P_j$ half its original weight.
endfor
```

Here $d$ is used to maintain higher diversity for resampled conformations.

Most polymers sampled by sequential Monte Carlo with two step look-ahead but no resampling are well-extended with few voids, as shown in Figs. 6(a) and 6(b). In Fig. 6(b), the majority of the conformations have a higher packing density. They have small compactness (<0.5) and large packing density. There are not enough compact conformations. As a result, a small number of samples are accepted at the end whose compactness falls within the desired interval of higher than 0.5. By using the resampling step described above, we were able to generate more samples near the desired compactness value of 0.6 (Fig. 6(c)). Figure 6(c) is a pure histogram of compactness in the observed samples, without regarding the weight of the samples. Here resampling is applied at each sequential Monte Carlo growth step. Figure 6(d) shows that the resampling technique is also very effective in shifting the samples to small packing density values.
hence improve the inferences. Here resampling favoring lowest packing density is applied every 5 growth steps.

V. VOID DISTRIBUTION OF LONG CHAINS

We apply the techniques of sequential Monte Carlo with resampling to study the statistical geometry of voids in long chain polymers. Each Monte Carlo simulation starts with a sample size of 200,000, and we take the averaged values of 20 simulations. Resampling is carried out every 5 steps in the process of the chain growth. Figure 7(a) shows that the probability of void formation increases with the chain length. At chain length 105–110, about half of the conformations contain voids. At chain length 200, the standard deviation $(8.5 \times 10^{-3})$ is maximum. The expected number of voids [Fig. 7(b)] increases linearly with chain length. Similar linear scaling behavior is also observed in proteins. The expected wall size of void and void size also increase with chain length [Figs. 7(c) and Fig. 7(d)]. The expected packing density is found to decrease with chain length, which is consistent with the scaling relationship of void size and chain length shown in Fig. 7(d). The compactness $\rho$ of chain polymer has been the subject of several studies. The asymptotic value of $\rho$ we found is 0.18, slightly different from that reported in Ref. 41 ($\rho = 0.16$), and is within the range of 0.16–0.24 reported in Ref. 12. Different resampling strategies are applied where dead conformations are removed and other conformations with the targeted property is duplicated. Resampling favors conformations with small radius-of-gyration in Figs. 7(a)–7(e), and conformations with large weight in Fig. 7(f).

To explore the relationship of packing density $\rho$ and compactness $\rho$, we use sequential Monte Carlo with 2-step look-ahead to sample 200,000 conformations, each with an appropriate weight assigned. This is repeated 20 times, and the weighted average values of the packing density at various compactness for chains with 60–100 monomers are plotted (Fig. 8). The compactness value corresponding to the minimum packing density seems to have shifted from 0.462 for the 22-mer by enumeration to above 0.5 for the 100-mer as shown in Figs. 8 and 4. Data shown in Figs. 8 and 4(d) are not redundant. Rather, they complement each other and together provide a full picture of the relationship of $\rho$ and $\rho$ for chain length between $N = 14–22$ and $N = 30–100$. The lowest packing density $\rho_{\text{min}}$ of these lattice polymers occur at compactness $\rho = 0.525$. However, an accurate estimation of the value of asymptotic $\rho_{\text{min}}$ as $N \rightarrow \infty$ requires simulation of longer chains.
The accuracy of geometric properties of long chain polymers estimated by Monte Carlo can be assessed by the standard deviation obtained from multiple Monte Carlo runs.

VI. END EFFECTS OF VOID FORMATION

What is the effect of void formation on the size of conformational space? We consider the conformational reduction factor of voids. Following Refs. 12, 23, 31, we define the conformational reduction factor due to the constraint of a void as

$$R(n; i, j) = \frac{\omega(n; i, j)}{\omega(n)},$$

where $\omega(n; i, j)$ is the number of conformations that contains a void beginning at monomer ($i$) and ending at monomer ($j$), and $\omega(n)$ is the total number of conformations of $n$-polymers. $R(n; i, j)$ reflects the restriction of conformational space due to the formation of a void with wall interval of $k = |i - j|$. Figure 9(a) shows a 24-mer with one void that starts at $i = 4$ and ends at $j = 23$. Unlike self-contacts or self-loops, which was subject of detailed studies by Chan and Dill$^{12,23,31}$, all conformations analyzed here must contain a void. The polymer shown in Fig. 9(b) with a large loop has no void, and such polymers do not contribute to the numerator of $R$.

Figure 10(a) shows the reduction factor $R$ calculated by enumeration for voids at different starting positions with wall intervals $k = 7, 9$, and 11. There are clearly strong end-effects: The reduction factor of voids of the same wall interval depends on where the void is located. $R$ decreases rapidly as the void moves from the end of chain towards the middle. Void formation is much more preferred at the end of chain.

Similar end effects of void formation are also observed for the 50-mer sampled by sequential Monte Carlo [Fig. 10(b)]. Because of the effect of excluded volume, sterically it is less hindering to form a void at the end of a polymer. When a void is formed, the conformational...
space of the $k+1$ monomers between monomer $i$ and $j$, as well as the two tails become restricted. When the void is formed at the chain end, only one tail is subject to conformational restriction.

Void formation is different from self-contact. When monomer $i$ and $j$ form self-contact, it may involve the formation of a void, but it is also possible that there will be no unfilled space between $i$ and $j$. When a void is formed beginning at monomer $i$ and ending at monomer $j$, some monomers between $i$ and $j$ will have unsatisfied contact interactions. Self-contacting loops have been analyzed in previous theoretical studies\cite{22,42} which have been confirmed by exhaustive enumeration.\cite{23,31} An important study of the role of loop formation in understanding the disulfide bond and protein folding can be found in Ref. 44. Compared to non-bonded self-contact, the effect of conformational reduction is more pronounced for void formation. For the two-dimensional lattice, the ratios between reduction factors of self-contact at chain end and mid-chain of a sufficiently long polymer are 1.3, 1.4, 1.5, and 1.6 for $k = 3, 5, 7, 9$, respectively,\cite{23} whereas the ratios for voids at the chain end and the symmetric midpoint of the $N = 22$ polymer are 3.4, 4.0, and 4.4, for $k = 7, 9$, and 11.

We now consider the power-law dependence of $R(N;i,j)$ on the wall interval $k = |i−j|$. In the study by Chan and Dill,\cite{31} the scaling exponent $\nu$ of the reduction factor $R$ and loop length $k = |i−j|$ for $R(N;i,j) = k^{−\nu}$ is found to be dependent both on $k$ and the location of the cycle in the chain. The values of $\nu$ for self-contact range from 1.6 when $k = N$ to 2.4 when the loop is in the middle of a long chain with two long tails. Because void formation involves at least 8 monomers, its scaling behavior is less amenable to exhaustive enumeration, and application of Monte Carlo sampling is essential. Based on estimations from Monte Carlo simulation of void formation in the 50-mer, the value of $\nu$ depends on the void initiation position $l_0$ from the end of the polymer chain. $\nu$ ranges from 1.4±0.2 for $l_0 = 1$ to 3.0±0.2 for $l_0 = 8$ [Fig. 10(c)]. Our results show that the scaling exponent of $R$ with $k = |i−j|$ for void formation is similar to that of the self-contacting loop. This scaling exponent also depends on the location of the void. The exponent $\nu$ is estimated from the nonlinear least square regression fit of the data using the Gauss–Newton algorithm as implemented in the GNU package R. A cautionary note is that the estimation of standard error of $\nu$ is accurate asymptotically only for large samples,\cite{45} and therefore maybe overly optimistic in our case, where the number of data points is very small. An accurate estimation of confidence interval of $\nu$ for small sample nonlinear regression is beyond the scope of this work.

VII. CONCLUSION

In this work, we have studied the statistical geometry of voids as topological features in two-dimensional lattice chain polymers. We define voids as unfilled space fully contained within the polymer, and have developed a simple algorithm for its detection. We have explored the relationship of various statistical geometric properties with the chain length of the polymer, including the probability of void formation $\pi_v$, the expected number of voids $\bar{n}_v$, the expected void size $\bar{v}$, the expected wall size of voids $\bar{w}$, packing density $\rho$, and the expected compactness $\rho$. Our results show that for chains of $>105$–110 monomers, at least half of the conformations contain a void. At about 150 monomers, there will be at least one void expected in a polymer. The expected wall size scale linearly with the chain length, and about $10\%$ of the monomers participate in the formation of voids. We formalize the concept of the packing density for lattice polymers. We found that both the packing density and compactness decrease with chain length. The asymptotic value of compactness $\rho$ is estimated to be 0.18.

We have also characterized the relationship of packing density and compactness, which are two parameters that have been used frequently for studying protein packing. Our results indicate that packing density reaches minimum values between compactness 0.4–0.6. The effects of voids are studied by analyzing the conformational reduction factor $R$ of void formation. We found that there is a significant end-effect for void formation; the ratio of $R$ at chain end and at mid-chain may be twice as large as that of the $R$ factor for contact loops, where the formation of voids is not required.

In this study, we have applied sequential Monte Carlo sampling and resampling (SMC) techniques to study the statistical geometry of voids. SMC is essential for exploring the geometry of long chain polymers. The origin of SMC can be traced back to the work of Rosenbluth and Rosenbluth,\cite{26} where the idea of placing the current monomer with probability dictated by the outcome of future steps was first formulated. Grassberger was the first to apply the technique of resampling by weight to grow the chain polymer.\cite{33} Independently, Liu and Chen provided the general framework of Sequential Monte Carlo, which unifies for the first time the techniques of delayed sampling (look-ahead of future steps) and resampling by arbitrary statistical property. Although SMC has had many applications in science and engineering,\cite{46} we report in this paper the first application of SMC for sampling a variety of rare events in growing polymers. Specifically, we make concrete novel applications of SMC in sampling rare conformations with the prescribed value of radius of gyration and compactness. Because the general framework of the SMC has not been described before for growing chain polymer, we also provide in this publication details of the validation of the method.

Sequential Monte Carlo allows the generation of an increased number of conformations with a variety of interesting characteristics. For example, we can replace dead conformations with existing conformations of highest weight, or conformations with highest compactness, or with the smallest radius of gyration. Figure 6 provided examples where the distribution of conformations of polymers obtained by SMC can be adjusted by resampling. In Fig. 11, we further elaborate on the general flexibility of SMC for sampling that targets on a variety of statistical properties, as well as the important fact that SMC maintains properly weighted samples, which is essential for any statistical inference.

Figure 11(a) shows the histograms of conformations of 100-mer at different packing density generated without resampling. Figure 11(c) shows the histograms of conformations
when resampling by weight and resampling by compactness ρ are used. To resample by weight, dead conformations are replaced with conformations of the highest weight. To resample by compactness, dead conformations are replaced with conformations of the lowest compactness. Note that the total number of surviving conformations that reach a chain length of 100 is much higher than without resampling. Resampling by compactness generates many more conformations with higher compactness. In Figs. 11(c) and 11(d), resampling is applied to every step of the chain growth process. Unlike Fig. 6(d), where resampling favors the compactness value of ρ = 0.6, resampling in Fig. 11(c) favors samples with the highest value of compactness ρ. Other resampling schemes are possible, e.g., resampling by radius-of-gyration, by packing density. During resampling, the number k of dead conformations at each step of growth is identified and these are replaced with conformations of interest from k randomly divided groups. These conformations must not have been resampled in the previous four steps of the growth process to maintain sample diversity. Both histograms where resampling is used deviate from that of Fig. 11(a). Resampling by weight shifts the peak of the conformations to below 0.2, and resampling by compactness turns the histogram into bimodal. The latter produces a lot more conformations with compactness ρ > 0.4.

SMC sampling and resampling use biased samples since conformations are generated with a probability different from that of the target distribution. The bias is dictated by the different method of resampling and different choices of the number of steps of look-ahead in sequential Monte Carlo. An essential component of a successful biased Monte Carlo sampling is the appropriate weight assignment to each sample conformation. This is necessary because we need to estimate the expected values of the parameter such as packing density and void size under the target distribution of all geometrically feasible conformations. In Fig. 11(a), where each of the 100 000 starting conformations is generated by a two-step look-ahead without resampling, not every conformation is generated with the same probability and therefore is assigned a different weight accordingly. Figure 11(b) shows the weight-adjusted histogram, which is indicative of the probability density function at different compactness for the population of all geometrically feasible 100-mers. Figure 11(d) shows that when weights are incorporated and the area of the histogram normalized to the final number of surviving conformations, the weighted distributions of conformations using different resampling techniques have excellent agreement with the weighted distribution when no resampling is used [Fig. 11(b)]. All weighted histograms are normalized so the total area equals to the total number of surviving conformations reaching 100-mer. This example shows that by incorporating weights, the target distributions can be faithfully recovered even when the sampling is very biased.

Although sequential Monte Carlo sampling is very effective, the estimation of parameters associated with rare events remain difficult. In Fig. 10 where the conformational reduction factor R is plotted at various void initiation positions and wall interval lengths, voids starting at position 1 but with odd wall intervals (k ∈ {11, 13, ..., 25}) are much rarer, and it is unlikely that sequential Monte Carlo sampling with limited sample size can provide a large enough effective sample size for the accurate estimation of scaling parameters ν, where \( R(N; i, j) \approx k^{-ν} \).

In this study, we are interested in the statistics of void geometry, and our target distribution is the uniform distribution of all conformations of length n. With the introduction of an appropriate potential function and alphabet of monomers such as the HP model,11,34,35 we can study the thermodynamics, kinetics, and sequence degeneracy of chain polymers when voids are formed in polymers. In these cases, our target distributions will be chain polymers under the Boltzmann distribution derived from the corresponding potential functions.

**ACKNOWLEDGMENTS**

This work is supported by funding from National Science Foundation DMS 9982846, CMS 9980599, DMS 0073601, DIBI078270, and MCB998008, and American Chemical Society/Petroleum Research Fund.

**APPENDIX: VOIDS DETECTION IN TWO-DIMENSIONAL LATTICE POLYMERS**

To detect voids in a polymer, we use a simple search method. For an \( l \times l \) lattice, we start from the lower-left corner. Once we found an unoccupied site \( u \), we use the breadth-first-search (BFS) method to identify all other unoccupied sites that are connected to site \( u \). These sites are grouped together and marked as "visited." Collectively they represent one void in the lattice. We continue this process until all unoccupied sites are marked as visited.
Algorithm VOIDDETECTION (lattice, l)
\n\n$v = 0$ // Number of voids
\nfor $i = 1 \rightarrow l$
\n\n\nfor $j = 1 \rightarrow l$
\n\n\nif site $(i, j)$ is unoccupied and not visited
\n$v \leftarrow v + 1$
\n\nMark $(i, j)$ as visited.
\n\nBREADTHFIRSTSEARCH(lattice, $(i, j)$)
\nUpdate the size of void $(i, j)$
\n\n\nendfor
\n\nendfor
\n\nDetails of BFS can be found in algorithm textbooks such as Ref. 47.
\n1 F. M. Richards, Annu. Rev. Biophys. Bioeng. 6, 151 (1977).
2 C. Chothia, Nature (London) 254, 304 (1975).
3 F. M. Richards and W. A. Lim, Q. Rev. Biophys. 26, 423 (1994).
4 J. Liang and K. A. Dill, Biophys. J. 81, 751 (2001).
5 B. Lorenz, I. Orgzall, and H-O. Heuer, J. Phys. A 339, 31 (1989).
6 W. A. Lim and R. Sauer, Nature (London) 339, 31 (1989).
7 D. Shortle, W. E. Sites, and A. K. Meeker, Biochemistry 29, 8033 (1990).
8 D. D. Axe, N. W. Foster, and A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. 93, 5590 (1996).
9 R. A. Laskowski, N. M. Luscombe, M. B. Swindells, and J. M. Thornton, Protein Sci. 5, 2438 (1996).
10 J. Liang, H. Edelsbrunner, and C. Woodward, Protein Sci. 7, 1884 (1998).
11 K. F. Lau and K. A. Dill, Macromolecules 33, 6737 (1989).
12 H. S. Chan and K. A. Dill, Macromolecules 22, 4559 (1989).
13 K. A. Dill, Biochemistry 29, 7133 (1990).
14 E. Shakhnovich and A. Gutin, J. Chem. Phys. 93, 5967 (1990).
15 C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. 90, 6369 (1993).
16 V. S. Pande, C. Joerg, A. Yu Grosberg, and T. Tanaka, J. Phys. A 27, 6231 (1994).
17 N. D. Socci and J. N. Onuchic, J. Chem. Phys. 101, 1519 (1994).
18 K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. 4, 561 (1995).
19 A. Šali, E. I. Shakhnovich, and M. Karplus, Nature (London) 369, 248 (1994).
20 I. Shrivastava, S. Vishveshwara, M. Cieplak, A. Maritan, and J. R. Banavar, Proc. Natl. Acad. Sci. U.S.A. 92, 9206 (1995).
21 D. K. Klimov and D. Thirumalai, Phys. Rev. Lett. 76, 4070 (1996).
22 R. Melin, H. Li, N. Wingreen, and C. Tang, J. Chem. Phys. 110, 1252 (1999).
23 H. S. Chan and K. A. Dill, J. Chem. Phys. 92, 3118 (1990).
24 S. Govindarajan and R. A. Goldstein, Biopolymers 36, 43 (1995).
25 H. Li, R. Helling, C. Tang, and N. Wingreen, Science 273, 666 (1996).
26 M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. 23, 356 (1955).
27 D. Frenkel and B. Smit, Understanding Molecular Simulation: From Algorithms to Applications (Academic, San Diego, 1996).
28 D. P. Landau and K. Binder, Monte Carlo Simulations in Statistical Physics (Cambridge University Press, Cambridge, 2000).
29 S. W. Golomb, Polyominoes: Puzzles, Patterns, Problems, and Packings (Princeton University Press, Princeton, 1994).
30 F. M. Richards, J. Mol. Biol. 82, 1 (1974).
31 H. S. Chan and K. A. Dill, J. Chem. Phys. 90, 492 (1989).
32 J. S. Liu, Monte Carlo Strategies in Scientific Computing (Springer, New York, 2001).
33 J. S. Liu and R. Chen, J. Am. Stat. Assoc. 93, 1032 (1998).
34 K. A. Dill, Biochemistry 24, 1501 (1985).
35 H. S. Chan and K. A. Dill, J. Chem. Phys. 99, 2116 (1993).
36 H. Meirovitch, J. Phys. A 15, L735 (1982).
37 J. S. Liu and R. Chen, J. Am. Stat. Assoc. 90, 567 (1995).
38 A. Kong, J. S. Liu, and W. H. Wong, J. Am. Stat. Assoc. 89, 278 (1994).
39 J. S. Liu, R. Chen, and W. H. Wong, J. Am. Stat. Assoc. 93, 1022 (1998).
40 P. Grassberger, Phys. Rev. E 56, 3682 (1997).
41 T. Ishinabe and Y. Chikahisa, J. Chem. Phys. 85, 1009 (1986).
42 J. Des Cloizeaux, J. Phys. (Paris) 41, 223 (1979).
43 B. Duplantier, Phys. Rev. B 35, 5290 (1987).
44 C. J. Camacho and D. Thirumalai, Proteins: Struct., Funct., Genet. 22, 27 (1995).
45 D. M. Bates and D. G. Watts, Nonlinear Regression Analysis and Its Application (Wiley, New York, 1988).
46 A. Doucet, N. De Freitas, N. Gordon, and A. Smith, Sequential Monte Carlo Methods in Practice (Springer-Verlag, Berlin, 2001).
47 T. H. Cormen, C. E. Leiserson, and R. L. Rivest, Introduction to Algorithms (MIT Press, Cambridge, 1990).