Configurational entropy and mechanical properties of cross-linked polymer chains: Implications for protein and RNA folding

Dmitrii E. Makarov[1,2,*] and Gregory J. Rodin[2,†]
[1]*Department of Chemistry and Biochemistry, The University of Texas at Austin, Austin, Texas 78712*
[2]*Texas Institute for Computational and Applied Mathematics, The University of Texas at Austin, Austin, Texas 78712*

(Received 6 December 2001; published 19 July 2002)

We discuss the statistical mechanical properties of a single polymer chain that forms cross links among its monomers. Models of this type have served as prototypes in theories of RNA and protein folding. The chain is allowed to form pseudoknots and its monomers can each participate in multiple cross links. We demonstrate that the conformational free energy of such a chain can be estimated by using an algorithm that scales as a power of the number of cross links $N(N^1-N^3$, depending on the problem). Straightforward exact evaluation of the chain partition function via multidimensional integration scales exponentially with N and often is computationally prohibitive. Our approach can also be used to compute the "entropic force" generated by a cross-linked chain when it is stretched at its ends. Such forces can be directly measured by atomic force microscopy or by laser optical trap experiments performed on single RNA, DNA, and protein molecules.

DOI: 10.1103/PhysRevE.66.011908 PACS number(s): 87.15.−v

I. INTRODUCTION

Biopolymers such as RNA and proteins self-assemble into unique three-dimensional structures that are necessary for their function. Understanding how nature accomplishes this task is one of the most active topics of current biochemical and biophysical research. In order to attain the most stable, native state, the initially unfolded polymer must overcome a free energy barrier that is associated with the loss of configurational entropy in the folding process. In order to understand the free energy landscape of folding, one is thus required to compute the free energy for various conformations of the polymer.

Recently, several workers have developed a picture in which the folding of a single biopolymer molecule is viewed as a process in which it forms the required native contacts [1–5]. To be more specific, in this picture the conformation of a chain is specified by a list of contacts (or cross links) $\{\{i_1,j_1\},\{i_2,j_2\},\ldots,\{i_N,j_N\}\}$ formed among its monomers. Here the monomers $i=1$, 2,..., L are numbered along the chain. An example of such a conformation is shown in Fig. 1. The meaning of a cross link or a contact depends on the particular physical problem under consideration. In the case of RNA the cross links refer to base pairing between complementary nucleotides. In the case of proteins, one says that two amino acid residues form a contact if the distance between them is smaller than a prescribed distance [6]. Forming a contact between a pair of monomers may or may not be associated with the creation of an actual chemical bond between them. The conformational space of the polymer consists of all possible sets of contacts.

The advantage of such a representation for a polymer is that the conformational space associated with it is much smaller than that of the original problem (the full coordinate space of the molecule). This allowed several groups to perform kinetic Monte Carlo simulations of RNA [3–5] and protein [1,2] folding, a task that is computationally prohibitive with the current atomistic scale methods.

The most time consuming step in such simulations is computing the free energies of contact formation, which is required in order to calculate the probabilities of contact formation or breaking. One therefore needs an efficient algorithm for computing the free energy for any given set of contacts. A straightforward algorithm, in which one simply performs a brute force evaluation of the partition function of the chain, involves N-dimensional integration, and therefore is prohibitive when the number of contacts is large. Such an approach has been used in Ref. [3] for the case of an RNA molecule that is sufficiently small. Other approaches used the single loop approximation [7,8], a mean-field-type theory [9], or an interpolation between the two [7]. As we found in our simulations, such approximations are often inaccurate, although they do reproduce qualitative trends in the free energies of proteins. Other approaches to the statistics of cross-linked chains were developed in gel theory [10]. However, there cross links are considered to be random. Here we are concerned with a single chain forming a specific set of cross links that is dictated by its secondary/tertiary structure.

Calculating free energies becomes much easier if the

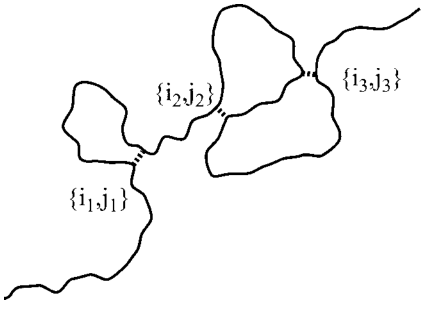

FIG. 1. An example of a polymer chain conformation with three contacts.

[*] Email address: makarov@mail.cm.utexas.edu

[†] Email address: gir@ticam.utexas.edu

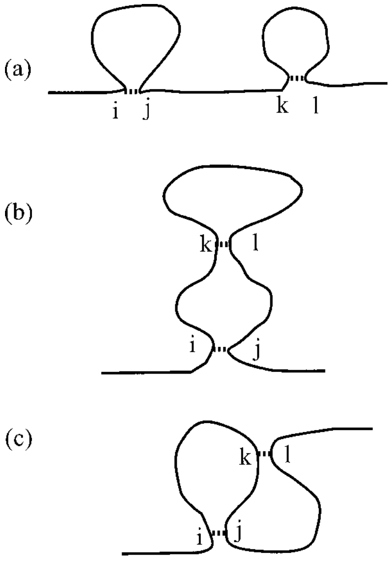

FIG. 2. (a) Independent, (b) nested, and (c) pseudoknotted arrangement of two contacts.

polymer does not form so called pseudoknots. Two contacts $\{i,j\}$ and $\{k,l\}$ do not form a pseudoknot if they are either independent $(i<j<k<l)$ or nested $(i<k<l<j)$, as shown in Fig. 2. In much of the work done on RNA folding, pseudoknot configurations of chains were disallowed [11], leading to very efficient dynamic programing algorithms for computing the partition functions [12–15]. With the discovery of pseudoknots in RNA (see, e.g., Ref. [16] and references therein) the need to take into account pseudoknotted RNA conformations has become apparent [3]. Pseudoknots complicate matters in two ways. First, the free energy of a cross-linked chain with pseudoknots is more difficult to calculate. Second, allowing pseudoknots leads to a dramatic increase in the total number of possible chain conformations. While the approach presented here does not solve the problem of searching all the pseudoknotted conformations, it facilitates the computation of their free energies.

In proteins, pseudoknotted arrangements of polypeptide chains are the rule rather than an exception. In addition, native contacts in proteins tend to "cluster." In other words, a single residue i may participate in multiple contacts, $\{i,j_1\}$, $\{i,j_2\},\ldots$. This happens because the contacts are not associated with molecular bonds but rather describe spatial proximity of amino acid residues. Hydrophobic effect favors the chain configurations, in which the side chains of the hydrophobic residues cluster together, thereby hiding from water and this results in the clustering of contacts. To illustrate this, Fig. 3 shows the contact map for the native state of the SH3 domain (protein databank file 1SHF.pdb). One can see that a single residue indeed forms multiple contacts.

The purpose of this paper is to demonstrate how to evaluate the configurational entropy of any cross-linked chain at a moderate computational expense. The specific questions that we will answer are as follows. Suppose we have a chain that has a given list of contacts $\{\{i_1,j_1\},\{i_2,j_2\},\ldots,\{i_N,j_N\}\}$, see Fig. 1. Suppose we also know the probability distribution $p_{0n}(\mathbf{r})$ for the end-to-end distance $\mathbf{r}$ vector of a chain of

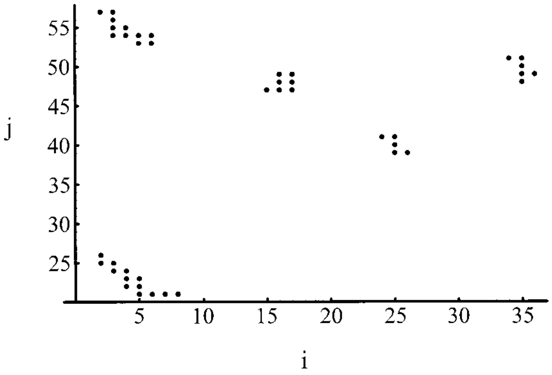

FIG. 3. A contact map of the SH3 domain protein. Amino acids number i and j are said to form a contact (represented as a point on this map) if the distance between their α-carbon atoms is shorter than 6 Å and if $|i-j|>12$ [1,2]

length n with no contacts. This probability can be approximated by various polymer models such as a Gaussian chain, the Kratky-Porod model, the wormlike chain, the freely jointed chain, etc [17]. What is the configurational entropy of such a cross-linked chain? Makarov and Metiu [1] have previously considered the case of Gaussian chain polymers. The method we propose here is applicable to a rather general class of polymers. The only requirement is that unconstrained polymer segments bounded by the "nodes" of the cross-linked network be statistically independent of each other. We explain this requirement in detail in Sec. II. Although this property may not be exactly satisfied for arbitrary chains, we argue there that it is a reasonable approximation for a broad range of polymer models including those conventionally used to describe biopolymers.

A related issue that is considered in this paper is calculating the force-extension curve of such a cross-linked polymer. This is of great interest in connection with the single molecule studies of individual RNA, DNA, and protein molecules stretched by the atomic force microscopy, laser optical trap, or magnetic bead techniques, see, e.g., Refs. [18–25]. The existing theoretical work on this problem for the case of RNA assumes the absence of pseudoknots [26–29]. We will show how to compute the force extension curves of single polymer chains that have cross links including those that result in pseudoknots. Of course, stretching a polymer molecule will result in the breaking of its cross links, which is not considered here. Instead, here we compute the force generated by the molecule with its cross links intact. To simulate single polymer extension experiments, one will have to supplement this treatment with a procedure describing the kinetics of cross-link formation and breaking (see, e.g., Ref. [30]).

This paper is organized as follows. Section II explains how one can calculate the contact formation probabilities and polymer force extension curves from the probability distributions of the distances between monomers. In Sec. III we derive Kirchhoff-type equations for calculating the probability distribution of the distance between any pair of monomers in a cross-linked chain. Section IV outlines a general algorithm for solving these equations for an arbitrary network of nonlinear chains and discusses how the computa-

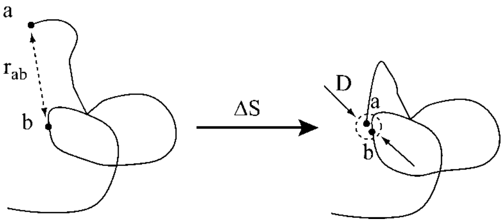

FIG. 4. Formation of a contact {*a,b*} in a cross-linked chain.

tional effort in this algorithm depends on the number of cross links. In Sec. V we describe our general algorithm for calculating conformational entropies of cross-linked chains. In Sec. VI we deal with a technical aspect that is concerned with the existence of redundant contacts. In Sec. VII we present a numerical example and calculate the entropy of a polymer that forms three cross links, two of which form a pseudoknot. Section VIII concludes with closing remarks.

II. CONTACT FORMATION PROBABILITIES AND FORCE EXTENSION CURVES OF CROSS-LINKED CHAINS

A. The entropy of contact formation

Consider the chain conformation shown in Fig. 4. What is the entropy change ΔS upon bringing points *a* and *b* in contact with one another? First, we need to specify more precisely what we call a contact. Here, we assume that *a* and *b* form a contact if the distance between them is less than a certain length *D*, $r_{ab}<D$. Let $p_{ab}(\mathbf{r}_{ab})$ be the probability distribution of the distance vector $\mathbf{r}_{ab}$. Then, assuming that *D* is small enough that this probability can be considered constant within a sphere of radius *D*, the probability of forming the contact is equal to

$$\exp(\Delta S/k_B)=v_0 p_{ab}(0), \tag{1}$$

where $v_0=4\pi D^3/3$. The total free energy change upon the formation of such a contact is

$$\Delta F=\varepsilon_{ab}-T\Delta S, \tag{2}$$

where ε_{ab} is the free energy of binding between *a* and *b*. We will not be concerned with this quantity here and will focus solely on ΔS. The details of the model such as the form of the short-ranged interaction between *a* and *b* and whether or not a contact is associated with a molecular bond, etc., will affect the factor v_0 but will not change the form of Eq. (1). We will therefore treat v_0 as a parameter here.

Central to the computation of ΔS is therefore the probability distribution $p_{ab}(\mathbf{r}_{ab})$ for the distance between two monomers. In what follows we will assume that this probability is known for any chain segment that does not form cross links. In the following section we will show how to calculate the probability of a cross-linked chain.

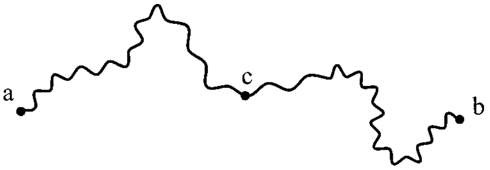

FIG. 5. A chain *ab* that consists of the segments *ac* and *cb*.

B. The force extension curve

The probability $p_{ab}(\mathbf{r}_{ab})$ is directly related to the force-extension curve of the chain when it is pulled at the points *a* and *b*. The pulling force $\mathbf{f}_{ab}$ is related to the extension of the chain by [10]

$$\mathbf{f}_{ab}=dF_{ab}/d\mathbf{r}_{ab}, \tag{3}$$

where

$$F_{ab}(\mathbf{r})=-k_BT\ln p_{ab}(\mathbf{r}) \tag{4}$$

is the free energy of the cross-linked chain with the constraint $\mathbf{r}_{ab}=\mathbf{r}$. The sign in Eq. (3) corresponds to f_{ab} being the force *exerted* on the chain. We note that one has to be careful applying Eq. (3) to single polymer molecules [31]. Depending on how the force extension curve is measured, there may be large fluctuation in the distance r_{ab} and/or in the force f_{ab}. For example, if one achieves polymer stretching by attaching a magnetic bead to its end then the force f_{ab} is determined by the magnetic field while the distance r_{ab} may undergo large fluctuations. Alternatively, one can force the polymer to have a given distance r_{ab} and then measure the resulting resistance force f_{ab} by monitoring the displacement of the cantilever attached to the polymer's end. In this case, the measured force may undergo fluctuations. If such fluctuations are large then the mean values of the force and/or extension are not necessarily the same as their most probable values. In this case, as shown in Ref. [31], the measured force extension curve may deviate significantly from Eq. (3).

C. Statistical independence of subchains

We see that we generally need to be able to compute the probability distribution of the end-to-end distance vector of a polymer. To make the treatment of complex cross-linked polymers feasible, we will make the assumption that subchains of such complex cross-linked chains are statistically independent. To illustrate the meaning of this, consider a chain shown in Fig. 5. We can view the chain *ab* as a composite chain that consists of the subchains *ac* and *cb*. The statistical independence implies

$$p_{ab}(\mathbf{r}_{ab})=\int d^3\mathbf{r}_c p_{ac}(\mathbf{r}_c-\mathbf{r}_a)p_{cb}(\mathbf{r}_b-\mathbf{r}_c). \tag{5}$$

Equation is satisfied exactly for any chain that consists of independent links. For a chain of *n* independent links that consists of monomers $k=0,1,\ldots,n$, one can write the end-to-end probability distribution as

$$p_{0n}(\mathbf{r})=\frac{\int d^3\mathbf{u}_1\cdots d^3\mathbf{u}_n \exp\{-[v(\mathbf{u}_1)+\cdots+v(\mathbf{u}_n)]/k_BT\}\delta(\mathbf{u}_1+\mathbf{u}_2+\cdots+\mathbf{u}_n-\mathbf{r})}{\int d^3\mathbf{u}_1\cdots d^3\mathbf{u}_n \exp\{-[v(\mathbf{u}_1)+\cdots+v(\mathbf{u}_n)]/k_BT\}}, \tag{6}$$

where $\mathbf{u}_i=\mathbf{r}_i-\mathbf{r}_{i-1}$ is the length of the ith link and $v(\mathbf{u}_i)$ its energy. If we set $a=0$, $b=n$ then one can easily see that the probability (6) satisfies Eq. (5) for any choice $0<c<n$.

Two examples of independent link chains commonly used in biopolymer modeling include (1) Gaussian chains where the potential v is harmonic, $v(u_i)=(1/2)\ \gamma u_i^2$, and (2) freely jointed chains where the length of the link is fixed, $|u_i|=1$.

When the links in the chain interact with one another, Eq. (5) is generally invalid. Such interactions can be of two types: local interactions that account for stiffness of the chains and nonlocal interactions leading to excluded volume effects [32].

If, for example, the chain is locally stiff, the direction of a link is correlated with that of its several neighbors. Consider two adjacent chain segments, *ac* and *cb*. They interact with one another via an interface whose size is of the order of the polymer persistence length l_p [32,33]. If the contour length of each chain segment, l_{ac} and l_{cb} is much longer than the persistence length l_p $(l_{ac},l_{cb}\gg l_p)$, then the free energy of the interface is much smaller than that of the segments; then Eq. (5) is approximately valid. In practice, we found Eq. (5) to be quite accurate even for modest values of the ratio l_{ac}/l_p. We have tested Eq. (5) for the wormlike chain model that is often used to model DNA or RNA chains [17] and found that while it is not exact it is generally quite accurate for the purpose of calculating the contact formation probabilities.

Another reason Eq. (5) may fail is excluded volume effects, which introduce nonlocal interactions among monomers. However, it has been argued (see Ref. [32]) that Eq. (5) must be valid for stretched chains where $\mathbf{r}_{ab}$, $\mathbf{r}_{ac}$, and $\mathbf{r}_{bc}$ are all long enough. In this limit, the probability distributions are known to follow the asymptotic behavior [10,32],

$$p_{ab}(\mathbf{r})\sim\exp(-c[r/R_{ab}]^{\delta}), \tag{7}$$

where R_{ab} is the mean distance between a and b and the scaling exponent δ is related to the dimensionality and is equal to 5/2 in three dimensions. In fact, Eq. (5) can be used to derive the relationship between δ and the dimensionality [32].

Thus we conclude that even in the presence of excluded volume effects the expression (5) may be a reasonable approximation. As mechanical tension tends to increase the distances among monomers, we expect Eq. (5) to become more accurate in the presence of large forces. The contact formation probabilities are less accurately described by Eq. (7) because those are dominated by small distances r, away from the asymptotic regime.

III. THE KIRCHHOFF RULES FOR CROSS-LINKED CHAINS

A. Sequential connection of chains

Consider two chains, *ac* and *cb*, connected as shown in Fig. 6(a). These either can be simple chains with no cross links or composite chains each having multiple cross links but there are no cross links between *ac* and *cb*. The probability distribution for the distance between a and b is equal to

$$\begin{aligned} p_{ab}(\mathbf{r})&=\int d^3\mathbf{r}_1 p_{ac}(\mathbf{r}_1)p_{cb}(\mathbf{r}-\mathbf{r}_1)\\ &\equiv\int d^3\mathbf{r}_1\exp[-F_{ac}(\mathbf{r}_1)/k_BT]\\ &\quad\times\exp[-F_{cb}(\mathbf{r}-\mathbf{r}_1)/k_BT]. \end{aligned} \tag{8}$$

We can evaluate the integral in Eq. (8) by the method of steepest descents. The integral is dominated by the region where the integrand achieves the maximum, which is given by the condition

$$F'_{ac}(\mathbf{r}_1)-F'_{cb}(\mathbf{r}-\mathbf{r}_1)=0. \tag{9}$$

Using Eq. (3), we can identify $\mathbf{f}_{ac}(\mathbf{r})\equiv F'_{ac}(\mathbf{r})$ as the "internal" force acting along the chain *ac*. Equation (9) then simply states that this force should be the same in both chains,

$$\mathbf{f}_{ac}(\mathbf{r}_1)=\mathbf{f}_{cb}(\mathbf{r}-\mathbf{r}_1). \tag{10}$$

This equation implicitly defines the individual chain extensions, r_1 and $r-r_1$ as a function of the total distance between a and b. A physical way to change the distance r is to apply an external force between the points a and b. The probability distribution of r in the presence of such a force will be

$$\begin{aligned} p_{ab}(\mathbf{r},\mathbf{f})&=\exp(\mathbf{f}\cdot\mathbf{r}/k_BT)p_{ab}(\mathbf{r})/Q(\mathbf{f})\\ &=Q(\mathbf{f})^{-1}\int d^3\mathbf{r}_1\exp(\mathbf{f}\cdot\mathbf{r}/k_BT)\\ &\quad\times\exp[-F_{ac}(\mathbf{r}_1)/k_BT]\exp[-F_{cb}(\mathbf{r}-\mathbf{r}_1)/k_BT], \end{aligned} \tag{11}$$

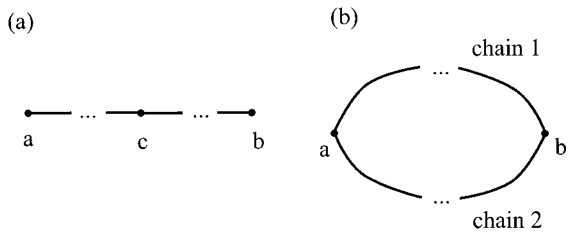

FIG. 6. (a) sequential and (b) parallel arrangements of two chains.

where the factor $Q(\mathbf{f})$ is required for normalization. Equation (11) follows from the fact that the energy of any chain conformation under the force $\mathbf{f}$ equals the energy of the same conformation without the force minus $\mathbf{f}\cdot\mathbf{r}_{ab}$. The average extension of the composite chain *ab* in the presence of the force is given by the integral

$$\langle\mathbf{r}\rangle=\int d^3\mathbf{r}p_{ab}(\mathbf{r},\mathbf{f})\cdot\mathbf{r}. \tag{12}$$

Substituting Eq. (11) into Eq. (12) and applying the method of steepest descents with respect to both $\mathbf{r}_1$ and $\mathbf{r}$ we find that the average extension satisfies the equation

$$\mathbf{f}=\mathbf{f}_{ac}(\mathbf{r}_1)=\mathbf{f}_{cb}(\langle\mathbf{r}\rangle-\mathbf{r}_1). \tag{13}$$

Equation (13) shows that the polymer springs *ac* and *cb* under the force $\mathbf{f}$ behave as mechanical springs with the force extension relationship,

$$\mathbf{f}_{ac}(\mathbf{r}_{ac})=F'_{ac}(\mathbf{r}_{ac})=-k_BT\nabla p_{ac}(\mathbf{r}_{ac})/p_{ac}(\mathbf{r}_{ac}),$$

$$\mathbf{f}_{cb}(\mathbf{r}_{cb})=F'_{cb}(\mathbf{r}_{cb})=-k_BT\nabla p_{cb}(\mathbf{r}_{cb})/p_{cb}(\mathbf{r}_{cb}), \tag{14}$$

where r_{ac} and r_{cb} are the mean end-to-end distances for these springs. Equation (13) simply states that these two springs must be in mechanical equilibrium under the force $\mathbf{f}$.

We are now in a position to calculate the force-extension curve of a composite chain *ab* and the entropy ΔS_{ab} of the formation of a contact between its ends *a* and *b*. The force-extension curve is simply obtained by solving, for every value of force $\mathbf{f}$, the equations

$$\mathbf{f}_{ac}(\mathbf{r}_{ac})=\mathbf{f},$$

$$\mathbf{f}_{cb}(\mathbf{r}_{cb})=\mathbf{f}$$

for $\mathbf{r}_{ac}$ and $\mathbf{r}_{cb}$ and then adding these displacements to calculate the total extension $\mathbf{r}_{ab}=\mathbf{r}_{ac}+\mathbf{r}_{cb}$. Thus we find the relationship $\mathbf{f}=\mathbf{f}_{ab}\cdot(\mathbf{r}_{ac}+\mathbf{r}_{cb})$ between the force $\mathbf{f}$ and the extension of the composite spring.

To calculate ΔS_{ab}, we need the probability distribution $p_{ab}(\mathbf{r})$ for the composite chain in the absence of the force. A convenient way to calculate this quantity is to use the force-extension curve for the composite chain $\mathbf{f}_{ab}(\mathbf{r})$ we just obtained. Using the relationship (3) we write

$$F_{ab}(\mathbf{r})=C+\int_0^{\mathbf{r}}d\mathbf{r}\mathbf{f}_{ab}(\mathbf{r})=C+\int_0^{r}drf_{ab}(\mathbf{r}). \tag{15}$$

Because the quantities $F_{ab}(\mathbf{r})$ and $|\mathbf{f}_{ab}\cdot(\mathbf{r})|$ are only dependent on the absolute value of $\mathbf{r}$ (assuming that the properties of the polymer do not depend on its orientation) the average chain extension $\mathbf{r}$ is along the force $\mathbf{f}$ and the integral in Eq. (15) is just a one-dimensional integral over $r=|\mathbf{r}|$. The constant C is determined from the condition that the probability

$$p_{ab}(\mathbf{r})=\exp[-F_{ab}(\mathbf{r})/k_BT]$$

$$=\exp\left[-\left(C+\int_0^r dr' f_{ab}(r')\right)/k_BT\right] \tag{16}$$

must be normalized,

$$\int d^3\mathbf{r}p_{ab}(\mathbf{r})=\int_0^\infty 4\pi r^2p_{ab}(r)dr=1, \tag{17}$$

which gives

$$\exp(-C/k_BT)$$

$$=\left[\int_0^\infty 4\pi r^2 dr\exp\left(-\int_0^r dr' f_{ab}(r')/k_BT\right)\right]^{-1}. \tag{18}$$

The contact formation probability is now given by Eq. (1).

Finally, we evaluate the loop closure probability, $p_{ab}(0)$, by using the steepest descent approximation. The dominant contribution to the integral in Eq. (18) comes from the vicinity of the equilibrium extension r^*, such that

$$f_{ab}(\mathbf{r}^*)=0. \tag{19}$$

Here we need to distinguish between two cases.

(1) $r^*=0$. This is often a good approximation for long flexible polymer chains. Furthermore, if both chains *ac* and *cb* have this property then the composite chain *ab* also has it.

(2) $r^*\neq 0$. For example, a short α helix in a protein will resist bending and so the average distance between its ends is obviously nonzero.

In case (1) we expand $F_{ab}(r)$ in power series around $r=0$,

$$F_{ab}(r)=(1/2)\left.\frac{df_{ab}}{dr}\right|_{r=0}r^2+\cdots. \tag{20}$$

The effective force constant of the chain is given by

$$\gamma_{ab}(r)\equiv\frac{df_{ab}}{dr}=\frac{df}{dr_{ac}+dr_{cb}}=\left[\left(\frac{df_{ac}}{dr_{ac}}\right)^{-1}+\left(\frac{df_{cb}}{dr_{cb}}\right)^{-1}\right]^{-1}$$

$$\equiv[1/\gamma_{ac}(r)+1/\gamma_{cb}(r)]^{-1}. \tag{21}$$

Again Eq. (21) is simply the rule of calculating the spring constant γ_{ab} of two sequentially connected mechanical springs with spring constants γ_{ac} and γ_{cb}.

Substituting Eqs. (20) and (21) into Eqs. (16)–(18), one obtains

$$p_{ab}(r)=\left(\frac{\gamma_{ab}(0)}{2\pi k_BT}\right)^{3/2}\exp\left(-\frac{\gamma_{ab}(0)r^2}{2k_BT}\right), \tag{22}$$

which is nothing but the probability distribution for the end-to-end distance of a spring with a force constant $\gamma_{ab}(0)$. This is, of course, only valid for small *r*. Thus the *ab* contact formation probability is proportional to

$$p_{ab}(0)=\left(\frac{\gamma_{ab}(0)}{2\pi k_BT}\right)^{3/2}. \tag{23}$$

In case (2) we similarly find the contact formation probability to be proportional to

$$p_{ab}(0)=\frac{1}{4\pi(r^{*})^{2}}\left(\frac{\gamma_{ab}(0)}{2\pi k_{B}T}\right)^{1/2}\exp\left[\int_{0}^{r^{*}}drf_{ab}(r)/k_{B}T\right]. \tag{24}$$

In this case, bringing a and b together requires overcoming a free energy barrier equal to $\int_{r*}^{0}drf_{ab}(r)$.

B. Parallel connection of chains

Consider now a parallel arrangement of two chains between the points a and b, Fig. 6(b). The probability distribution for the distance r between points a and b is equal to

$$\begin{aligned}p_{ab}(\mathbf{r})&=Np_{ab}^{(1)}(\mathbf{r})p_{ab}^{(2)}(\mathbf{r})\\&=N\exp\{-[F_{ab}^{(1)}(\mathbf{r})+F_{ab}^{(2)}(\mathbf{r})]/k_{B}T\},\end{aligned} \tag{25}$$

where the superscripts (1) and (2) indicate the two chains and N is a normalization factor. If one pulls on such a polymer at points a and b with a force $\mathbf{f}$ then the probability distribution becomes

$$p_{ab}(\mathbf{r},\mathbf{f})=N'(\mathbf{f})\exp\{-[F_{ab}^{(1)}(\mathbf{r})+F_{ab}^{(2)}(\mathbf{r})-\mathbf{f}\cdot\mathbf{r}]/k_{B}T\}, \tag{26}$$

where $N'(\mathbf{f})$ is another force-dependent normalization factor. The maximum of $p_{ab}(\mathbf{r},\mathbf{f})$ is achieved for the extension $\mathbf{r}$ satisfying the equation

$$\mathbf{f}=dF_{ab}^{(1)}(\mathbf{r})/d\mathbf{r}+dF_{ab}^{(2)}(\mathbf{r})/d\mathbf{r}\equiv\mathbf{f}_{ab}^{(1)}(\mathbf{r})+\mathbf{f}_{ab}^{(2)}(\mathbf{r}). \tag{27}$$

This is nothing but mechanical equilibrium condition for two parallel mechanical springs stating that the sum of forces at either a or b is zero. The arguments that led us to Eqs. (15)–(24) remain unchanged with the only exception that the total effective spring constant between the points a and b is now given by the sum of the individual spring constants, as is immediately seen from Eq. (27),

$$\gamma_{ab}(r)=\gamma_{ab}^{(1)}(r)+\gamma_{ab}^{(2)}(r)\equiv\frac{df_{ab}^{(1)}}{dr_{ab}}+\frac{df_{ab}^{(2)}}{dr_{ab}}. \tag{28}$$

C. The general case

One can break up a complex network that consists of elementary chains (each having no contacts) into multiple parallel or sequential connections. That is, one can build the network recursively by making sequential and parallel connections among various subchains. The rules for the composite chains are the same as those for the mechanical equilibrium of a system of mechanical springs. They are also the same as those for the conductance of a network of conductors. In fact, if we call γ_{ab}conductance between points a and b and $1/\gamma_{ab}$ the resistance between these points then Eq. (21) would state that the resistance of two sequentially connected conductors is the sum of their resistances and Eq. (28) would mean that the conductance of two parallel conductors is the sum of their conductances. To pursue this analogy even further, we can treat f_{ab} for an elementary chain as "current." Equations (13) and (27) then simply state that the total algebraic sum of currents in each node is zero. Finally, the distance r_{ab} is analogous to the voltage between points a and b in electric circuits. Stated above are the standard Kirchhoff rules for electrical circuits. Thus computing the force extension curve for a cross-linked chain that is pulled at points a and b is equivalent to calculating its conductance or computing the mechanical equilibrium of a network of mechanical chains. For linear springs with distance-independent force constants $\gamma_{\mathrm{ab}}(r)=\gamma_{ab}$ this could be achieved by solving a system of linear equations that express the equilibrium conditions for each node. The problem is somewhat more difficult for nonlinear springs, where these equations have to be solved self-consistently such that the resulting extension of each chain r_{ab} would be consistent with its spring constant γ_{ab}. We deal with this problem in the following section.

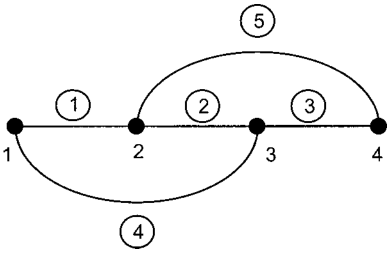

FIG. 7. Representation of the network used in Sec. IV. The network contains four nodes numbered from 1 to 4 and five elements (springs), whose numbers are encircled.

IV. NUMERICAL APPROACHES TO THE COMPUTATION OF EFFECTIVE FORCE CONSTANTS OF CROSS-LINKED NETWORKS

We have found from the preceding section that, for the purposes of calculating the effective spring constant and the force-elongation curve of a system of *entropic* springs, one can pretend that it is a system of *mechanical* springs. The energy of the equivalent mechanical spring connecting monomers i and j is given by $F_{ij}(|\mathbf{r}_i\text{-}\mathbf{r}_j|)|$. The physical origin of F_{ij} is entropic; however, this quantity can be treated as energy of an equivalent mechanical spring. If a force is applied to a pair of nodes of this network, all nodes will be displaced along the direction of the force. Thus our equivalent mechanical problem is one dimensional. The original problem of the cross-linked polymer is of course three-dimensional. However, as shown in Sec. III, the *average* displacement of each network node is equal to the displacement of the same node in the equivalent mechanical network. This displacement is along the applied force.

Outlined below is a general algorithm for computing the mechanical properties of a network of nonlinear mechanical springs. The representation of the network usind in this section is illustrated in Fig. 7. Notice that, as a practical trick, it is often convenient to represent cross links between pairs of monomers in the polymer chain as springs of infinite (or, in practice, very large) stiffness that can be incorporated into the network.

We view the network as a one-dimensional chain that is fixed at one end and loaded by an axial force f at another end. This chain is a collection of one-dimensional elements (or springs); each element connects two nodes of the network. We denote the nodal positions before application of the force by x_i, and after application of the force by $x_i + u_i$. The subscript i here enumerates all the chain nodes from left to right (so that $i=1,2,3,4$ in Fig. 7). We suppose that the total number of nodes is N, the first node is fixed ($u_i=0$) and the Nth node is loaded by the force. We use *superscripts* to enumerate the *elements* of the network. There are five elements in the network shown in Fig. 7, their numbers are encircled in the drawing. The energy of the jth element is given by $F^{(j)}(u_+^{(j)}-u_-^{(j)})$ and depends on its elongation $u_+^{(j)}-u_-^{(j)}$, where $u_+^{(j)}$ and $u_-^{(j)}$ are the displacements of the nodes that are connected by this element. We thus are using two different notations for the displacements of the same nodes; In the *local* notation, $u_+^{(j)}$ and $u_-^{(j)}$ are used to denote the displacement of the two nodes that are adjacent to the element number j. The same nodes are assigned a global index enumerating them along the chain. For example, consider element number 5 in Fig. 7. It connects nodes number 2 and 4. We can write the displacements of the node 2 in two different ways, as u_2 (using the global notation) or as $u_-^{(5)}$ (using the local notation with respect to element number 5). Similarly, we write the displacement of node 4 as $u_4 \equiv u_+^{(5)}$.

The total energy of the network is given by

$$F=\sum_j F^{(j)}(u_+^{(j)}-u_-^{(j)})-fu_N=E-fu_N. \tag{29}$$

The second term in this expression is the energy associated with the force. Minimizing F under the constraint $u_i=0$, one obtains a system of equations for computing the displacements u_i. If F were a quadratic functional, these equations would be linear and have the form

$$\begin{pmatrix} K_{11} & K_{12} & \cdots & K_{1N} \\ K_{21} & K_{22} & \cdots & K_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ K_{N1} & K_{N2} & \cdots & K_{NN} \end{pmatrix} \begin{pmatrix} 0 \\ u_2 \\ \cdots \\ u_N \end{pmatrix} = \begin{pmatrix} r \\ 0 \\ 0 \\ f \end{pmatrix}, \tag{30}$$

where the entries of the stiffness matrix are computed as the second partial derivatives of F with respect to the nodal displacements u_i, and r is the reaction force associated with the constraint $u_i=0$.

In the finite element method [34], the stiffness matrix is assembled from the elemental stiffness matrices. In our case, the stiffness matrix of the jth element, using the local notation, has the form

$$\begin{pmatrix} \gamma^{(j)} & -\gamma^{(j)} \\ -\gamma^{(j)} & \gamma^{(j)} \end{pmatrix}, \tag{31}$$

where $\gamma^{(j)}$ is the ordinary spring constant of the element. The assembly involves "stretching" of each elemental stiffness matrix, so that its entries are placed in the locations of the global stiffness matrix corresponding to the global numbers of the nodes adjacent to the element. Then the global stiffness matrix is obtained by summation of the stretched elemental stiffness matrices.

As an example, consider the chain shown in Fig. 7. It is made of four nodes and five elements. The elements 4 and 5 are actually straight, and shown as curved for clarity. The stretched elemental matrices are

$$\begin{pmatrix} \gamma^{(1)} & -\gamma^{(1)} & 0 & 0 \\ -\gamma^{(1)} & \gamma^{(1)} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \gamma^{(2)} & -\gamma^{(2)} & 0 \\ 0 & -\gamma^{(2)} & \gamma^{(2)} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma^{(3)} & -\gamma^{(3)} \\ 0 & 0 & -\gamma^{(3)} & \gamma^{(3)} \end{pmatrix}, \quad \begin{pmatrix} \gamma^{(4)} & 0 & -\gamma^{(4)} & 0 \\ 0 & 0 & 0 & 0 \\ -\gamma^{(4)} & 0 & \gamma^{(4)} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \gamma^{(5)} & 0 & -\gamma^{(5)} \\ 0 & 0 & 0 & 0 \\ 0 & -\gamma^{(5)} & 0 & \gamma^{(5)} \end{pmatrix},$$

so that the global stiffness matrix is

$$\begin{pmatrix} \gamma^{(1)}+\gamma^{(4)} & -\gamma^{(1)} & -\gamma^{(4)} & 0 \\ -\gamma^{(1)} & \gamma^{(1)}+\gamma^{(2)}+\gamma^{(5)} & -\gamma^{(2)} & -\gamma^{(5)} \\ -\gamma^{(4)} & -\gamma^{(2)} & \gamma^{(2)}+\gamma^{(3)}+\gamma^{(4)} & -\gamma^{(3)} \\ 0 & -\gamma^{(5)} & -\gamma^{(3)} & \gamma^{(3)}+\gamma^{(5)} \end{pmatrix}. \tag{32}$$

To calculate the displacement of each node we insert this into Eq. (30) and solve the resulting linear system of equations. For example, if we take $\gamma^{(1)}=\gamma^{(2)}=\gamma^{(3)}=\gamma^{(4)}=\gamma^{(5)}=\gamma$ then we find

$$r=-f, \quad u_2=f/2\gamma, \quad u_3=f/2\gamma, \quad u_4=f/\gamma, \quad F=-f^2/2\gamma. \tag{33}$$

If we pull this network at nodes 1 and 4, it behaves as a single spring with a stiffness equal to γ.

For nonlinear problems, which are of primary interest to this work, the formulation of the governing equations is not that much different. Upon variation of the *unconstrained* free energy, we obtain the system of nonlinear equations,

$$\begin{pmatrix} \partial E/\partial u_1 \\ \partial E/\partial u_2 \\ \cdots \\ \partial E/\partial u_N \end{pmatrix} - \begin{pmatrix} r \\ 0 \\ \cdots \\ f \end{pmatrix} \equiv \begin{pmatrix} g_1 \\ g_2 \\ \cdots \\ g_N \end{pmatrix} - \begin{pmatrix} r \\ 0 \\ \cdots \\ f \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}. \tag{34}$$

By applying Newton's iterative method to this system, one obtains a system of equations whose stiffness matrix is computed according to the same rules as that of the linear system. The key difference between the two matrices is that the stiffness matrix associated with the linear problem is independent of the loading and displacements, whereas the stiffness matrix associated with the nonlinear problem depends on the displacements u_i and reaction r at the beginning of the iteration. Imposing the constraint $u_1=0$, we obtain the following system of equations for the corrections to the displacements δu_i and reaction δr:

$$\begin{pmatrix} K_{11}(u_i,r) & K_{12}(u_i,r) & \cdots & K_{1N}(u_i,r) \\ K_{21}(u_i,r) & K_{22}(u_i,r) & \cdots & K_{2N}(u_i,r) \\ \cdots & \cdots & \cdots & \cdots \\ K_{N1}(u_i,r) & K_{N2}(u_i,r) & \cdots & K_{NN}(u_i,r) \end{pmatrix} \begin{pmatrix} 0 \\ \delta u_2 \\ \cdots \\ \delta u_N \end{pmatrix} = \begin{pmatrix} \delta r \\ 0 \\ 0 \\ f \end{pmatrix} - \begin{pmatrix} r \\ g_2(u_i,r) \\ \cdots \\ g_N(u_i,r) \end{pmatrix}. \tag{35}$$

Newton's method exhibits a quadratic rate of convergence, provided the initial guess is reasonably close to the solution. Replacing all elements by linear springs may, for instance, provide a reasonable initial guess. If a close initial guess is not available, one may resort to iterative methods that are less sensitive to the quality of the initial guess [35].

In general, it would be difficult to estimate the cost of this approach in terms of the number of arithmetic operations. Optimal estimates are $O(N)$ and pessimistic estimates are $O(N^3)$. The former estimate corresponds to chains whose stiffness matrix is sparse (small number of cross links) and the latter estimate corresponds to dense stiffness matrices (large number of cross links). The minimization problem considered here can be treated by a variety of methods of nonlinear programming [36], and their relative advantages and disadvantages are significantly problem dependent. Nevertheless, it is safe to claim that the approach proposed here is capable of handling problems with hundreds of thousands of unknowns, which is far more than is required in any practical problem involving an RNA or a protein molecule.

V. COMPUTING THE CONFIGURATIONAL ENTROPY

A. The procedure

We are now ready to formulate our general recipe how to compute the total configurational entropy of an arbitrary polymer chain with the contacts $\{\{i_1,j_1\},\{i_2,j_2\},\ldots,\{i_N,j_N\}\}$. We start with a chain conformation that has no contacts. We denote this conformation $\{\}$. Then we compute the entropy change $\Delta S[\{\}\rightarrow\{\{i_1,j_1\}\}]$ of forming the first contact. This is given by Eq. (1),

$$\exp(\Delta S[\{\}\rightarrow\{\{i_1,j_1\}\}]/k_B)=v_0 p_{i_1,j_1}(0). \tag{36}$$

Using Eqs. (16) and (18),

$$p_{i_1j_1}(0)=\left[\int_0^\infty 4\pi r^2 dr \times\exp\left(-\int_0^r dr' f_{i_1j_1}(r')\Big/ k_BT\right)\right]^{-1}. \tag{37}$$

The force extension curve $f_{i_1j_1}(r)$ of the chain is known. Depending on the model, it may be computed with molecular dynamics or by a Monte Carlo method from an atomistic model or approximated by one of the many available models of biopolymers (wormlike chain, freely jointed chain, etc.).

Next we compute the change of entropy $\Delta S[\{\{i_1,j_1\}\}\rightarrow\{\{i_1,j_1\},\{i_2,j_2\}\}]$ upon the addition of a second contact, $\{i_2,j_2\}$. This is given by Eq. (36) and (37) except the force $f_{i_1j_1}(r)$ is now replaced by $f_{i_2j_2}(r)$, the force-extension curve of the chain between points i_2 and j_2 *in the presence of the cross link* $\{i_1,j_1\}$. The latter is computed as the force in a composite chain using Kirchhoff's rules as described above.

We next add the third cross link, $\{i_3,j_3\}$, recompute the forces and calculate the entropy change. This procedure is repeated until all the required contacts are created and the total configurational entropy is the sum of the entropy change in each step,

$$S=\Delta S[\{\}\rightarrow\{\{i_1,j_1\}\}]+\Delta S[\{\{i_1,j_1\}\} \rightarrow\{\{i_1,j_1\},\{i_2,j_2\}\}]+\cdots. \tag{38}$$

B. The Gaussian chain approximation

The procedure is simplified greatly if all the elementary chains satisfy the condition $f_{ab}(0)=0$. This condition ensures that the probability distribution $p_{ab}(\mathbf{r})$ has a maximum at $r=0$. Such a condition is satisfied by many models of random coils (including, for example, the common wormlike chain model). Then using the arguments that lead to Eq. (23), we find that the entropy ΔS of forming a new contact $\{i,j\}$ is given by

$$\exp(\Delta S/k_B) = v_0 p_{ij}(0) = v_0 \left(\frac{\gamma_{ij}(0)}{2\pi k_B T} \right)^{3/2}. \qquad (39)$$

The advantage of using Eq. (39) instead of Eq. (37) or Eq. (24) is obvious: there is no need to calculate the distance between *i* and *j* for different forces in order to compute the integrals in Eq. (37). There is no need to solve the nonlinear system of Kirchhoff's equations to compute the extensions of nonlinear springs. Instead, one replaces all the chains with *linear* springs whose effective force constants are those computed for zero extensions. The problem is thus reduced to the computation of the configurational entropy of a system of Gaussian chains with the force constants $\gamma_{ij}(0) = d^2F_{ij}/dr^2|_{r=0}$. Such a problem has already been solved in Ref. [1].

Replacing all nonlinear chains by linear springs may seem to be a rather drastic approximation. It is well known that entropic forces measured, e.g., in single DNA, RNA, or protein molecules are strongly nonlinear [21,22,25,31]. We stress that this replacement is accurate *for the purposes of calculating the chain entropy* (in the absence of a force) while it would be totally inadequate for the calculation of force extension curves of such a chain. If the condition $f_{ab}(0)=0$ is satisfied for each elementary chain, then conformations where any one of them is strongly extended will not be likely to be sampled by the entire cross-linked chain without an external force. This is why the Gaussian approximation is appropriate if one needs to calculate the entropy of this chain. When a force is applied, extended conformations of the chain will be sampled where deviations from the Gaussian distribution become significant.

To illustrate this point, in Fig. 8(a) we plot the exact probability distribution $p(r)$ for the distance between the ends of a freely jointed chain. The chain contains 100 links and the length of each link is 0.3 nm. This probability is indistinguishable from the Gaussian distribution for an equivalent Gaussian spring. The probability distribution eventually becomes non-Gaussian when *r* is large; however, for such large distances $p(r)$ is essentially zero. Since in calculating the loop closure probability only $p(0)$ is needed, this probability, for all practical purposes, is exactly the same as that of an equivalent Gaussian chain. However, if one wants to use Eq. (3) to calculate the force $f(r)$ for large extensions *r*, then the non-Gaussian tail of $p(r)$ is essential regardless of how small $p(r)$ is. Thus one sees strong nonlinearity in the force extension curve for the freely jointed model, plotted in Fig. 8(b). Comparing Figs. 8(a) and 8(b), we find that nonlinear behavior of $f(r)$ sets in for the extensions so large that $p(r)$ is essentially zero.

VI. REDUNDANT CONTACTS

The contact map for the SH3 domain protein shown in Fig. 3 has been computed from the protein databank file containing the coordinates of each atom in the protein as follows. We identify the position of a monomer (x_i, y_i, z_i) with that of its α carbon. For every pair of monomers $i<j$ we say they form a contact if

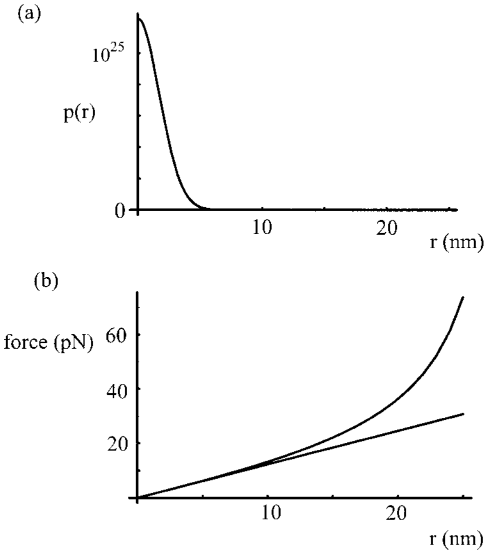

FIG. 8. (a) The probability distribution of the distance between the ends of a freely jointed chain (see text for the parameters). It is indistinguishable from the same distribution calculated in the Gaussian approximation. (b). The force extension curve of the freely jointed chain. The straight line is the result obtained in the Gaussian approximation.

$$\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2} < D, \qquad (40)$$

where the contact radius was taken to be $D=6$ Å, and if the distance between them along the chain $|i-j|$ satisfies the inequality $|i-j|>C$ (where $C=12$ for Fig. 3). The second condition was needed in Ref. [2] to exclude short-range contacts.

A list of contacts generated in this way, however, will contain redundant contacts. To explain what we mean by this consider the contact list (see also Fig. 9) $\{\{a,b\},\{b,c\},\{a,c\}\}$. An attempt to compute the conformational entropy of such a chain configuration using the algorithm described in the preceding section will cause trouble. Indeed, in this algorithm we would first close the loop between *a* and *b*, then form the contact $\{b,c\}$. Then we would be supposed to compute the entropy change for the formation

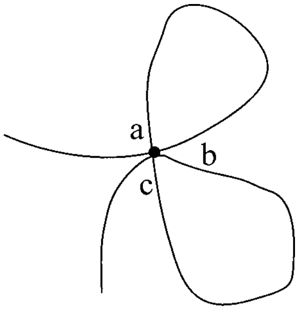

FIG. 9. A chain that forms contacts $\{a,b\}$ and $\{b,c\}$.

of the contact $\{a,c\}$, for which we would need the probability $p_{ac}(\mathbf{r})$. But the points a and c are already in contact (if a is in contact with b and b is in contact with c then a is in contact with c) and so this quantity is meaningless. The calculation will go awry when trying to form such a contact that already exists.

To avoid this difficulty, in our algorithm we eliminate the redundant contact by the following procedure. We first combine all contacts into *clusters*. For example, the list of contacts $\{\{a,b\},\{b,c\},\{a,c\},\{d,e\}\}$ will be equivalent to the following list of clusters $\{\{a,b,c\},\{d,e\}\}$, where the contacts $\{a,b\}$, $\{b,c\}$, and $\{a,c\}$ have been combined to a single cluster $\{a,b,c\}$, where each monomer appears only once. We now break up each cluster back into contacts. For a cluster of the form $\{a_1,a_2,\ldots,a_m\}$ the resulting contacts will be $\{a_1,a_2\},\ldots,\{a_{m-1},a_m\}$. Thus a cluster consisting of m elements will be broken up into $m-1$ contacts. Thus the original list $\{\{a,b\},\{b,c\},\{a,c\},\{d,e\}\}$ undergoes the following two transformations

$$\begin{aligned}&\{\{a,b\},\{b,c\},\{a,c\},\{d,e\}\}\\&\quad\rightarrow\{\{a,b,c\},\{d,e\}\}\\&\quad\rightarrow\{\{a,b\},\{b,c\},\{d,e\}\},\end{aligned}$$

as a result of which one redundant contact has been eliminated. The list of contacts that results from this procedure has the same entropy as the original one but it has no redundant contacts.

VII. A NUMERICAL EXAMPLE: A FOLDING CHAIN WITH A PSEUDOKNOT

To illustrate the ideas described in the previous sections, we consider here a simple example, calculating the entropy cost of forming a cyclic polymer that contains a single pseudoknot, as shown Fig. 10(a). Our model is a polymer that consists of $L-1=20$ links. We will assume that each segment of this polymer is a Gaussian chain. More specifically, we model each link as a spring with a force constant γ_1. The probability distribution for the end-to-end distance of a free chain of n links is then given by

$$p_{0n}(r)=\left(\frac{\gamma_n}{2\pi k_BT}\right)^{3/2}\exp\left(-\frac{\gamma_n r^2}{2k_BT}\right), \tag{41}$$

where $\gamma_n=\gamma_1/n$. The mean square distance between the ends of an n-link chain is then

$$\langle r^2\rangle=3k_BT/\gamma_n\equiv ns^2,\quad s^2=3k_BT/\gamma_1. \tag{42}$$

The list of contacts in the folded chain configuration whose entropy we would like to calculate is {{0, 20}, {4, 12}, {8, 16}}. We will assume that a contact is formed when two monomers are within a distance D and the resulting entropy loss is given by Eq. (1) with $v_0=4\pi D^3/3$.

Although in the previous sections we have described a brute force algorithm for calculating the entropy of such a chain with any number of contacts, this particular problem is simple enough that the result can be obtained from a back-of-an-envelope calculation.

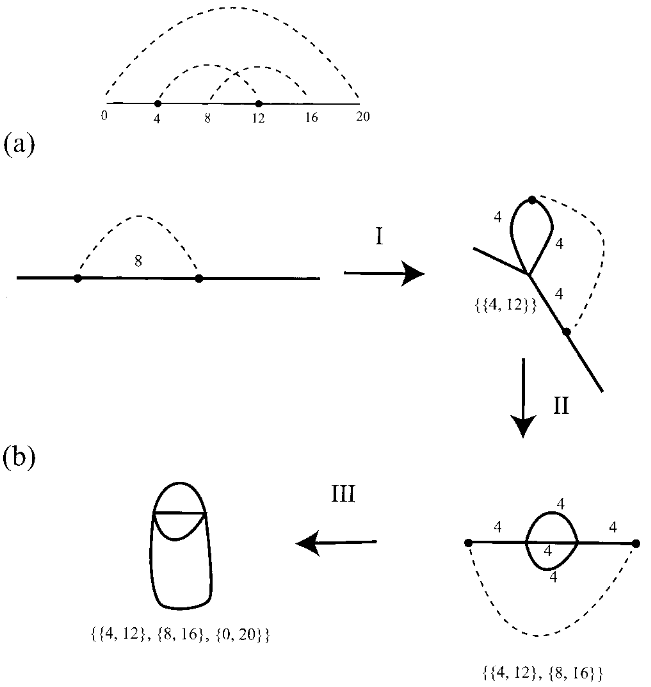

FIG. 10. (a) A chain forming contacts {{0, 20}, {4, 12}, {8, 16}}. The contacts to be formed are shown as dashed lines. (b) Possible steps that result in the formation of this chain configuration. The contacts to be formed at each step are shown by dashed lines.

To see this, consider a specific folding path of the chain that results in the desired set of contacts, see Fig. 10(b). It consists of three steps.

Step I. Form the contact {4, 12}. To find the entropy change, we need the probability distribution for the distance between monomers number 4 and 12. This is given by a Gaussian distribution of a spring whose length is $n=8$ links and therefore the spring constant is $\gamma_1/8$. The resulting entropy change is found from

$$\begin{aligned}\exp(\Delta S_{\mathrm{I}}/k_B)&=v_0p_{0n}(0)\\&=(4\pi D^3/3)\left(\frac{\gamma_1/8}{2\pi k_BT}\right)^{3/2}\\&=(4\pi D^3/3)\left(\frac{3}{16\pi s^2}\right)^{3/2}.\end{aligned} \tag{43}$$

Step II. Form the contact {8, 16}. By examining Fig. 9(b) we find that the effective spring between monomer 8 and monomer 16 (in the presence of the contact {4, 12}) consists of two parallel chains four links each that are connected sequentially with another chain of length 4. Using the rules formulated in Sec. III we find that the effective spring constant for the equivalent chain would be given by

$$\gamma_{\mathrm{eff}}^{-1}=(\gamma_1/4)^{-1}+[2(\gamma_1/4)]^{-1}$$

or $\gamma_{\mathrm{eff}}=\gamma_1/6$. The resulting entropy change is

$$\exp(\Delta S_{\mathrm{II}}/k_B)=(4\pi D^3/3)\left(\frac{\gamma_1/6}{2\pi k_B T}\right)^{3/2}$$
$$=(4\pi D^3/3)\left(\frac{1}{4\pi s^2}\right)^{3/2}. \qquad (44)$$

Step III. We now see from Fig. 10(b) that the pseudoknot formed by the pair of contacts {4, 12} and {8, 16} appears as three chains each of length 4 connected in parallel. The resulting pseudoknot spring constant is $3\gamma_1/4$. We finally complete folding by closing the loop between monomers 0 and 20. The effective spring between these monomers consists of the above pseudoknot spring connected, in sequence, with two chains each including four links. The resulting spring constant satisfies the relation

$$\gamma_{\mathrm{eff}}^{-1}=(\gamma_1/4)^{-1}+(\gamma_1/4)^{-1}+[3(\gamma_1/4)]^{-1},$$

which gives $\gamma_{\mathrm{eff}}=3\gamma_1/28$, and, for the entropy change

$$\exp(\Delta S_{\mathrm{III}}/k_B)=(4\pi D^3/3)\left(\frac{3\gamma_1/28}{2\pi k_B T}\right)^{3/2}$$
$$=(4\pi D^3/3)\left(\frac{9}{56\pi s^2}\right)^{3/2}. \qquad (45)$$

The resulting entropy change is

$$\exp(\Delta S/k_B)=\exp(\Delta S_{\mathrm{I}}/k_B)\exp(\Delta S_{\mathrm{II}}/k_B)\exp(\Delta S_{\mathrm{III}}/k_B)$$
$$=\frac{3\sqrt{3/14}}{896\pi^{3/2}}\left(\frac{D}{s}\right)^3\approx 0.000\,278\,35(D/s)^3. \qquad (46)$$

One can check that by taking a different folding path, i.e., by adding the contacts in a different order, one obtains different entropy changes in each individual steps but identically the same final entropy (46). Of course, one does not have to perform these steps manually. Our algorithm does them automatically.

In Fig. 11, we have summarized the different folding paths and plotted the entropy of each intermediate state. For definiteness, we assumed $D/s=1$. The entropy of the initial state without contacts is set to be the origin, $S=0$. All paths lead to the same final state with all the contacts formed. Because of the symmetry, the contacts {4, 12} and {8, 16} are equivalent. This results in only four distinct intermediate states shown in Fig. 11. The lines between the states indicate the possible folding paths.

It is instructive to compare the exact result, Eq. (46), with the existing approximations. In the independent loop approximation [7] one writes the total entropy as

$$\Delta S\approx\sum_{i_1,j_k}\Delta S_{i_k j_k}, \qquad (47)$$

where the entropy of the formation of a single contact is given by the single-loop expression

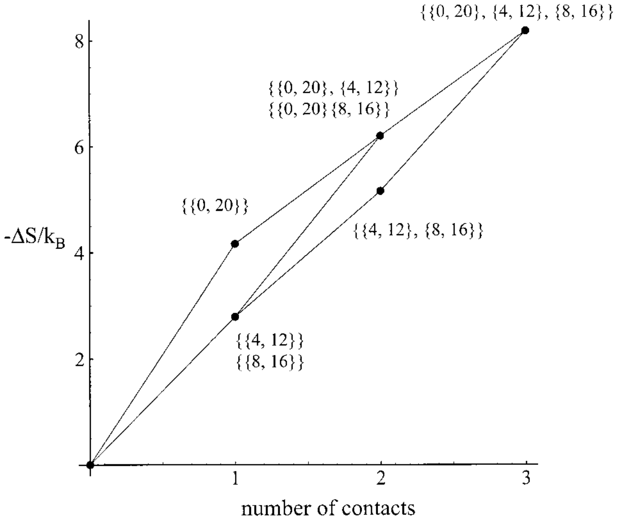

FIG. 11. Possible pathways resulting in the formation of the chain configuration with the three contacts, {{0, 20}, {4, 12}, {8, 16}}. The chain entropy is plotted as a function of the number of contacts for each path.

$$\exp(\Delta S_{ij}/k_B)=v_0 p_{ij}(0)=\left(\frac{6}{\pi}\right)^{1/2}\left(\frac{D}{s}\right)^3|i-j|^{-3/2}. \qquad (48)$$

Applying Eq. (47) to the present problem gives

$$\exp(\Delta S/k_B)\approx 0.000\,057\,635(D/s)^3. \qquad (49)$$

Comparing this with Eq. (46), the independent loop approximation underestimates the probability of forming our configuration by about a factor of five. This is not surprising as this approximation ignores the fact that forming one contact makes the chain more compact and thus facilitates the formation of others.

We next analyze the mean-field approximation [7,9]. In this approximation one uses Eq. (47) with ΔS_{ij} set to be a constant,

$$\exp(\Delta S_{ij}/k_B)=\left(\frac{6}{\pi}\right)^{1/2}\left(\frac{D}{s}\right)^3 l_{\mathrm{eff}}^{-3/2}, \qquad (50)$$

where

$$l_{\mathrm{eff}}=L/N. \qquad (51)$$

This approximation is expected to work well in the limit of a large number of contacts N, while in our case $N=3$. Using Eqs. (47), (50), and (51), we find

$$\exp(\Delta S/k_B)\approx 0.000\,415\,49(D/s)^3, \qquad (52)$$

which overestimates the probability of forming this configuration by a factor of ~1.5.

Finally, Shoemaker and Wolynes [7] used an interpolation formula between the single-loop limit and the mean-field formula, Eqs. (48) and (50),

$$\exp(\Delta S_{ij}/k_B)=\left(\frac{6}{\pi}\right)^{1/2}\left(\frac{D}{s}\right)^3(l_{\text{eff}}^{-3/2}+|i-j|^{-3/2}) \quad (53)$$

Substituting this into Eq. (47), we find

$$\exp(\Delta S/k_B)\approx 0.001\,659(D/s)^3, \quad (54)$$

which is larger than the exact value.

VIII. CONCLUDING REMARKS

The representation, in which a polymer conformation is specified by a set of contacts it forms, provides a convenient way to discretize the conformational space of an RNA or a protein molecule thereby reducing its size and making it tractable. Any state of the polymer is represented as a contact map such as the one in Fig. 3. Other discrete models of biopolymers were proposed, lattice models being most notable among them [37,38]. We note that lattice proteins are ''models'' while contact maps are coarse grained representations of ''true'' proteins or RNA molecules.

The folding pathways, for a contact representation of a protein or an RNA molecule can be plotted as diagrams in Fig. 11. Some of the pathways may be blocked because of high free energy barriers along them. One can further study the dynamics of such models by assuming that the transition between any two adjacent points of a diagram is a first order kinetic process, with forward and backward rate constants satisfying the principle of detailed balance [1]. We also note that contact formation rate constants have been measured experimentally in the case of simple loops in some polypeptides [39–41].

The use of our approach by itself does not solve the RNA or protein ''folding problem.'' The total size of the conformational space and therefore the number of possible folding pathways is still exponentially large. Thus an exhaustive search for the minimum free energy state is still a prohibitive problem. However, use of our algorithm will greatly enhance kinetic Monte Carlo simulations [1–3,5,30,42] that sample the kinetically probable pathways rather than all possible pathways. A kinetic Monte Carlo algorithm mimics the evolution of a single molecule en route to its native state and can be directly related to the single molecule observations of protein or RNA kinetics [43].

We finally note that our results concerning the mechanical properties of cross-linked chains imply that the mechanical response of individual protein and RNA molecules is controlled by their native topology. Klimov and Thirumalai [44] arrived at the same conclusion on the basis of an off-lattice simulation of the force induced unfolding of globular proteins. The diversity of protein tertiary structures thus accounts for the diversity of the mechanical properties exhibited by proteins in living organisms.

ACKNOWLEDGMENTS

This work was supported by the Army Research Office (G.J.R.) and the Robert A. Welch Foundation (D.E.M.).

[1] D. E. Makarov and H. Metiu, J. Chem. Phys. **116**, 5205 (2002).

[2] D. E. Makarov, C. Keller, K. W. Plaxco, and H. Metiu, Proc. Natl. Acad. Sci. U.S.A. **99**, 3535 (2002).

[3] H. Isambert and E. D. Siggia, Proc. Natl. Acad. Sci. U.S.A. **97**, 6515 (2000).

[4] O. V. Galizitskaya and A. V. Finkelstein, J. Chem. Phys. **105**, 319 (1995).

[5] C. Flamm, W. Fontana, I. L. Hofacker, and P. Schuster, RNA **6**, 325 (2000).

[6] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).

[7] B. A. Shoemaker and P. G. Wolynes, J. Mol. Biol. **287**, 657 (1999).

[8] H. Jacobson and W. H. Stockmayer, J. Chem. Phys. **18**, 1600 (1950).

[9] P. J. Flory, J. Am. Chem. Soc. **78**, 5222 (1956).

[10] P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).

[11] R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999).

[12] M. Zuker and P. Stiegler, Nucleic Acids Res. **9**, 133 (1981).

[13] J. S. McCaskill, Biopolymers **29**, 1105 (1990).

[14] R. Nussinov and A. B. Jacobson, Proc. Natl. Acad. Sci. U.S.A. **77**, 6309 (1980).

[15] S.-J. Chen and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **97**, 646 (2000).

[16] I. Tinoco and C. Bustamante, J. Mol. Biol. **293**, 271 (1999).

[17] T. Strick, J.-F. Allemand, V. Croquette, and D. Bensimon, Prog. Biophys. Mol. Biol. **74**, 115 (2000).

[18] M. B. Viani, T. E. Schaffer, G. T. Paloczi, I. Pietrasanta, B. L. Smith, J. B. Thompson, M. Richter, M. Rief, H. E. Gaub, K. W. Plaxco, A. N. Cleland, H. G. Hansma, and P. K. Hansma, Rev. Sci. Instrum. **70**, 4300 (1999).

[19] L. Tskhovrebova, J. A. Trinic, J. A. Sleep, and R. M. Simmons, Nature (London) **387**, 308 (1997).

[20] B. L. Smith, T. E. Schaffer, M. Viani, J. B. Thompson, N. A. Frederick, J. Kindt, A. Belcher, G. D. Stucky, D. E. Morse, and P. K. Hansma, Nature (London) **399**, 761 (1999).

[21] T. Strick, J. Allemand, V. Croquette, and D. Bensimon, Prog. Biophys. Mol. Biol. **74**, 115 (1999).

[22] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, Science **276**, 1109 (1997).

[23] J. F. Marko and E. Siggia, Macromolecules **28**, 209 (1995).

[24] T. E. Fisher, A. F. Oberhauser, M. C. Vezquez, P. E. Marsalek, and J. Fernandez, Trends Biochem. Sci. **24**, 379 (1999).

[25] K. Wang, J. G. Forbes, and A. J. Jin, Prog. Biophys. Mol. Biol. **77**, 1 (2001).

[26] H. Zhou, Y. Zhang, and Z.-C. Ou-Yang, Phys. Rev. Lett. **86**, 356 (2001).

[27] U. Gerland, R. Bundschuh, and T. Hwa, Biophys. J. **81**, 1324 (2001).

[28] A. Montanari and M. Mezard, Phys. Rev. Lett. **86**, 2178 (2001).

[29] D. K. Lubensky and D. R. Nelson, e-print cond-mat/0107423.

[30] D. E. Makarov, P. K. Hansma, and H. Metiu, J. Chem. Phys. **114**, 9663 (2001).

[31] D. E. Makarov, Z. Wang, J. B. Thompson, and H. G. Hansma, J. Chem. Phys. **116**, 7760 (2002).

[32] J. des Cloizeaux and G. Jannink, *Polymers in Solution* (Clarendon, Oxford, 1990).

[33] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1953).

[34] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method* (Butterworth-Heinemann, Boston, 2000).

[35] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (Prentice-Hall, Englewood Cliffs, NJ, 1983).

[36] M. S. Bazaraa and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms* (Wiley, New York, 1979).

[37] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[38] H. S. Chan and K. A. Dill, Proteins: Struct., Funct., Genet. **30**, 2 (1998).

[39] S. J. Hagen, J. Hofrichter, A. Szabo, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **93**, 11 615 (1996).

[40] L. J. Lapidus, W. A. Eaton, and J. Hofrichter, Proc. Natl. Acad. Sci. U.S.A. **97**, 7220–7225 (2000).

[41] W. A. Eaton, V. Munoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter, Annu. Rev. Biophys. Biomol. Struct. **29**, 327 (2000).

[42] O. V. Galzitskaya and A. V. Finkelstein, Proc. Natl. Acad. Sci. U.S.A. **96**, 11 299 (1999).

[43] C. Bai, C. Wang, X. S. Xie, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **96**, 11 075 (1999).

[44] D. K. Klimov and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **97**, 7254 (2000).