Journal of Computer Science & Systems Biology - Open Access
www.omicsonline.com Research Article JCSB/Vol.2 February 2009

Modeling the Chain Entropy of Biopolymers: Unifying Two Different Random Walk Models under One Framework

Wayne Dawson* and Gota Kawai

Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino-shi, Chiba 275-0016, Japan

*Corresponding author: Wayne Dawson, Bioinformation Engineering Laboratory, Department of Biotechnology, Graduate School of Agriculture and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, E-mail: dawson@bi.a.u-tokyo.ac.jp

Received December 18, 2008; Accepted January 31, 2009; Published February 03, 2009

Citation: Dawson W, Kawai G (2009) Modeling the Chain Entropy of Biopolymers: Unifying Two Different Random Walk Models under One Framework. J Comput Sci Syst Biol 2: 001-023.

Abstract

Entropy plays a critical role in the long range structure of biopolymers. To model the coarse-grained chain entropy of the residues in biopolymers, the lattice model or the Gaussian polymer chain (GPC) model is typically used. Both models use the concept of a random walk to find the conformations of an unstructured polymer. However, the entropy of the lattice model is a function of the coordination number, whereas the entropy of the GPC is a function of the root-mean square separation distance between the ends of the polymer. This can lead to inconsistent predictions for the coarse-grained entropy. Here we show that the GPC model and the lattice model both are consistent under transformations using the cross-linking entropy (CLE) model and that the CLE model generates a family of equations that include these two models at important limits. We show that the CLE model is a unifying approach to the thermodynamics of biopolymers that links these incompatible models into a single framework, elicits their similarities and differences, and expands beyond the models allowing calculation of variable flexibility and incorporating important corrections such as the worm-like-chain model. The CLE model is also consistent with the contact-order model and, when combined with existing local pairing potentials, can predict correct structures at the minimum free energy.

Introduction

Modeling the entropy of a biopolymer is usually handled in two steps. In the first step, the coarse-grained interactions and conformations are modeled. This typically involves representing the monomers as featureless blobs that are connected to each other like links on a chain and interact with one another by thermodynamic potentials. In the second step, the detailed, local, context-dependent interactions are included. Both steps are essential to estimating the precise entropy of a biopolymer; however, the coarse-grained and fine-grained interactions appear to be independent enough that they can be handled in an additive approach (Honig et al., 1976). This is the basis of the hierarchical folding concept (Baldwin and Rose, 1999ab; Tinoco and Bustamante, 1999). The coarse-grained interactions tend to affect the global entropy of the biopolymer whereas the fine-grained interactions affect the local thermodynamics. In this work, we focus on modeling the coarse-grained contributions that affect the global entropy.

A true representation for the coarse-grained entropy of any polymer remains unknown. Two models that are frequently applied to the problem are the lattice model and the Gaussian polymer chain (GPC) model.

Lattice models are quite successful at predicting simple protein folds and the funnel shape in the folding landscape (Dill and Stigter, 1995 ; Chan and Dill, 1997; Go, 1999; Onuchic et al., 2000; Kolinski et al., 2003; Pokarowski et al., 2003), protein evolution (Mirny et al., 1998), protein-protein docking (Zhang et al., 1997; Mintseris and Weng,

2003) and explaining surface adhesion of a protein in a very intuitively simple manner (Liu and Haynes, 2005). Numerous hybrid models with more real world examples also exist; such as real structure based models of RNA (Chen, 2008) and proteins (Day and Dagget,t 2003; Ding et al., 2008).

Lattice models offer a convenient computational approach to reduce the number of conformations of a small biopolymer to a manageable size. For example, for a protein, one might try a lattice model with a coordination number 3, yielding $O(3^N)$ configurations. The base is known as the coordination number (q). The coordination number is usually associated with the Ramachandran angles of a protein that are mainly distributed within 3 principal sectors of the plot; right handed alpha-helices(α_R), left-handed alpha-helices (α_L) and beta-strands (β) (Lesk, 2001). Crudely speaking, one can chose a single pair of Ramachandran angles within each sector from the average of the observed angles to "represent" that sector. Similar observations can be made for RNA (Takasu et al., 2002; Murray et al., 2003; Chen, 2008).

Yet such selections are ultimately subjective. In addition to these 3 regions; one could justifiably insist on adding a variety of other protein secondary structure elements such as beta-turns, 3_{10} helices, parallel and anti-parallel beta strands and effectively an infinite host of other possibilities. These can even be constructed based on some reasonable criteria (Pappu and Rose, 2002). The choice on how to cordon off these sectors to assign such angles is often decided based on computational considerations. Nevertheless, whatever criteria is used, all such cordoning is subjective and not unique.

The GPC model is based on the experimentally observed physical tendencies of real polymers (Flory 1969; Grosberg and Khokhlov, 1994). The concept of a "chain" comes from the image of a real chain where the links would roughly resemble the coarse-grained features of individual monomers (or "mers" for short). Experimental parameters such as the radius of gyration and the polymer stretching interaction can be directly associated with parameters in the GPC model (Flory, 1953). The GPC-model consists of a polymer chain in which each link of the chain is free to rotate over the entire 4π solid angle (able to rotate through every angle of latitude and longitude including back on itself, which even a real polymer chain cannot do). Hence, the GPC-model also contains peculiarities that are not aesthetically satisfactory – though other standard models such as "the free electron gas" contain similar disquieting artifacts yet yield correct conclusions (Ashcroft and Mermin, 1976).

The lattice model and the GPC model are inconsistent. Admittedly, because the lattice has fixed dimensions and constraints, it produces a similar root-mean-square end-to-end separation distance to the GPC model (App A; Section A2). However, the entropy of a lattice model is proportional to the number of residues N ($\Delta S \propto N \ln q$) yet the entropy of the GPC is proportional to the end-to-end distance ($\Delta S \propto 2\ln r - \beta r^2$) where β is a constant (App A4, Equations (A9-11)). These expressions have little in common, especially since r is a variable.

The cross-linking entropy (CLE) model was developed to model the *coarse-grained* entropy of biopolymers. This model has been shown to significantly improve the calculations of the minimum free energy and has been applied to prediction of RNA pseudoknots and simple proteins (Dawson et al., 2005; Dawson et al., 2007). The model also offers ways to understand the flexibility, polymer-solvent effects and correlation effects within a given biopolymer. A similar estimate of the configuration entropy derived from the CLE model (Dawson et al., 2001ab; Dawson et al., 2006; Dawson et al., 2007) has been recently reported by another research group (Hnizdo et al., 2008).

Here we show a consistent way to connect all of the best features of both the lattice model and the GPC-model via the CLE-model so that at one end of the spectrum we find a lattice model and at the other end, we see a GPC-model. In addition, we show how the worm-like-chain model (Flory, 1969) and the contact order model (Ivankov et al., 2003) fit into this context. The discussion in this work is not restricted to biopolymers. It applies to all polymers where binding of diverse parts can be found. Although we often focus on the more familiar Ramachandran angles common to protein research, such concepts can be generalized (Taylor, 1948; Takasu et al., 2002; Murray et al., 2003) such that the model applies to all polymers where similar interactions are observed. This work focuses mainly on the theoretical aspects of the models because the practical applications of the CLE model have already been demonstrated elsewhere previously (Dawson et al., 2001ab; Dawson et al., 2005; Dawson et al., 2006; Dawson et al., 2007).

Results and Discussion

The Lattice Model can Fail to Correctly Estimate the Conformations of an Ideal Polymer

There are good reasons to use a lattice model. It is difficult to do computer simulations on a true GPC-model because one must consider all the angles in the solid angle. Typically, a real polymer is not free to rotate over this entire

region; even in principle. Hence, choosing a lattice of some tractable coordination number is a very sensible strategy to use in *approximating* the essential conformations of a given polymer.*

However, is the lattice model a reasonable one-to-one approximation of a simple folded polymer? To examine this, we turn to the example of protein beta sheets (β-sheets) because the concepts of protein structure are likely to be

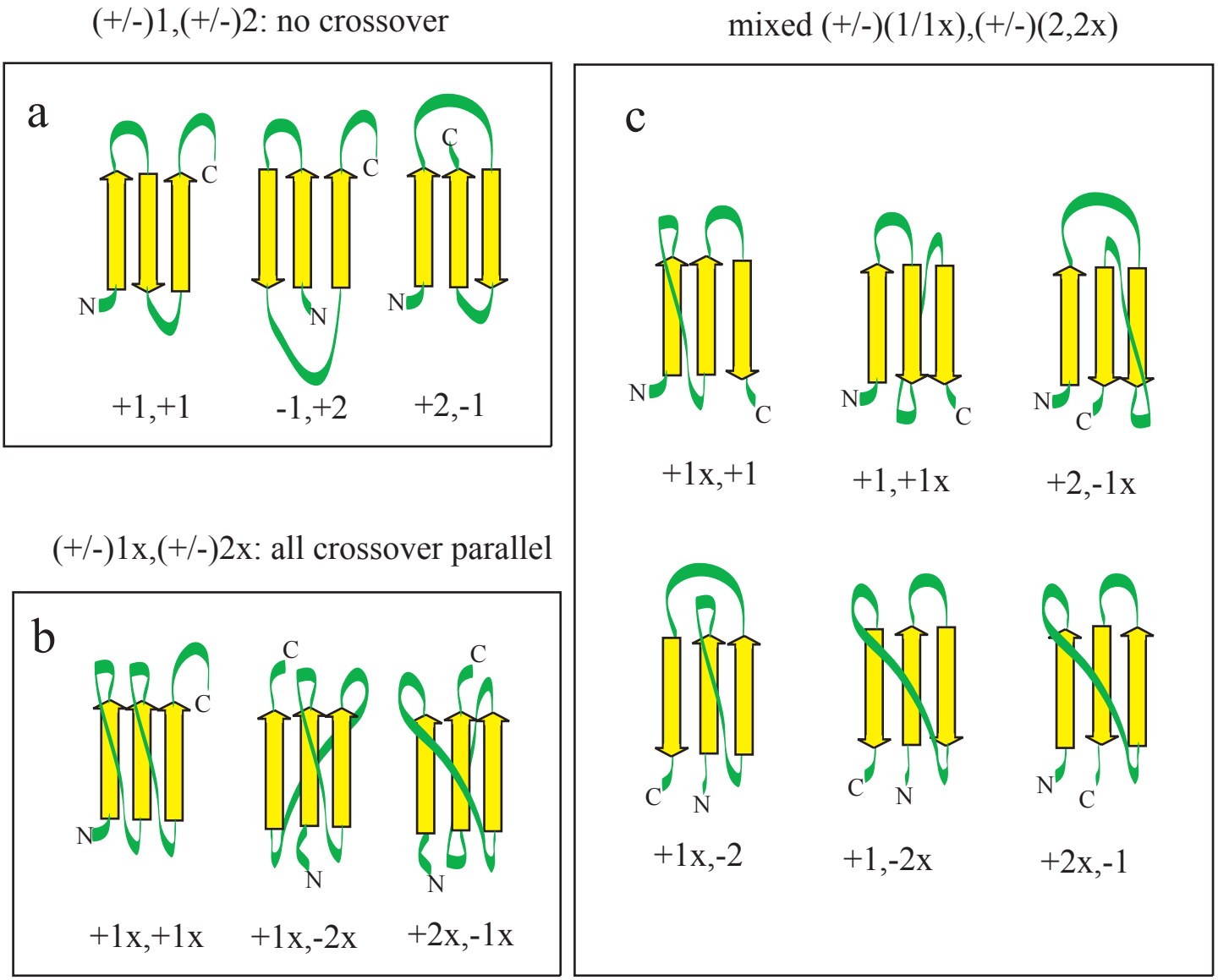

Figure 1: The full number of arrangements that can be generated from a protein composed of 3 beta-strands when right handed crossover of the beta-strands is included: (a) no crossover, (b) all crossovers and (c) a mixture of (a) and (b). The notation below indicates the location of the next strand (±1, ±2) and the x (±1x, ±2x) indicates a crossover beta-strand (Richardson, 1977). The total number of arrangements follows the rule $2^{n-1}n!/2$ (Cohen et al., 1982), where n is the number of beta-strands.

*It is important to distinguish between the GPC-model and the freely jointed polymer chain (FJPC) model, because there are overlaps and similarities. Both *can* assume that the bond angle is free to rotate over the entire 4π solid angle. The main difference is that the FJPC-model still treats the monomers in the polymer chain as individual units (often including physically justified fixed bond angles) and therefore links between monomers are equal to the physical distance between them. The GPC-model goes one step further in abstraction allowing the distances between effective monomers to consist of non-integral distances of the physical monomer to monomer (mer-to-mer) separation distance. This distance is known as the Kuhn length (App A). The advantage of GPC-model is that it captures the physical behavior of real polymers which generally do not tend to bend on the same length scale as the mer-to-mer distance. Real polymers tend to be stiff and unable to bend on such short length scales. On the other hand, the FJPC takes into consideration the contour length of the polymer chain and therefore, it vehemently resists stretching beyond the contour length. The GPC can be stretched to infinite length. The FJPC-model is sometimes simplified to a finite set of fixed angles using molecular geometry as a guide. Effectively, this reduces the FJPC-model to the form of a lattice model.

familiar to most readers. Since this should also apply to RNA, we first must show that an equivalent β-sheets-like configuration can be generated with RNA pseudoknots as well. For convenience, in this Section, we assume that the Kuhn length (ξ: App A1) is of unit value; $\xi \equiv 1$ mer.

A set of n beta-strands (β-strands) − without other types of protein secondary structure − yields a total of $(1+n!)n!/2$ unique β-sheets. Usually, the left handed and mixed left and right handed crossovers are excluded because they are far less common (Richardson, 1977). The arrangement of three β-strands that exclude all left handed and mixed cross-over structures is shown in Fig 1. The notation (Richardson, 1977) indicates the positioning of the next strand relative to a given strand where (±1, ±2) indicates non-crossover β-strands and (±1x, ±2x) indicates right-handed crossover β-strands. The total number of ways that n β-sheets can be arranged (excluding left handed and mixed crossover) is shown by (Cohen et al., 1982) to be

$$\Omega_{\beta} = 2^{n-1}\,(n!/2) \qquad (1)$$

where the factor 2^{n-1} accounts for the β-sheets that contain right handed crossover β-strands. The second term $(n!/2)$ is the combinatorial patterns of parallel and anti-parallel β-sheets with no crossover (Fig 1a). All the conformational patterns in Equation (1) can certainly exist and can be found by X-ray crystallography or NMR techniques.

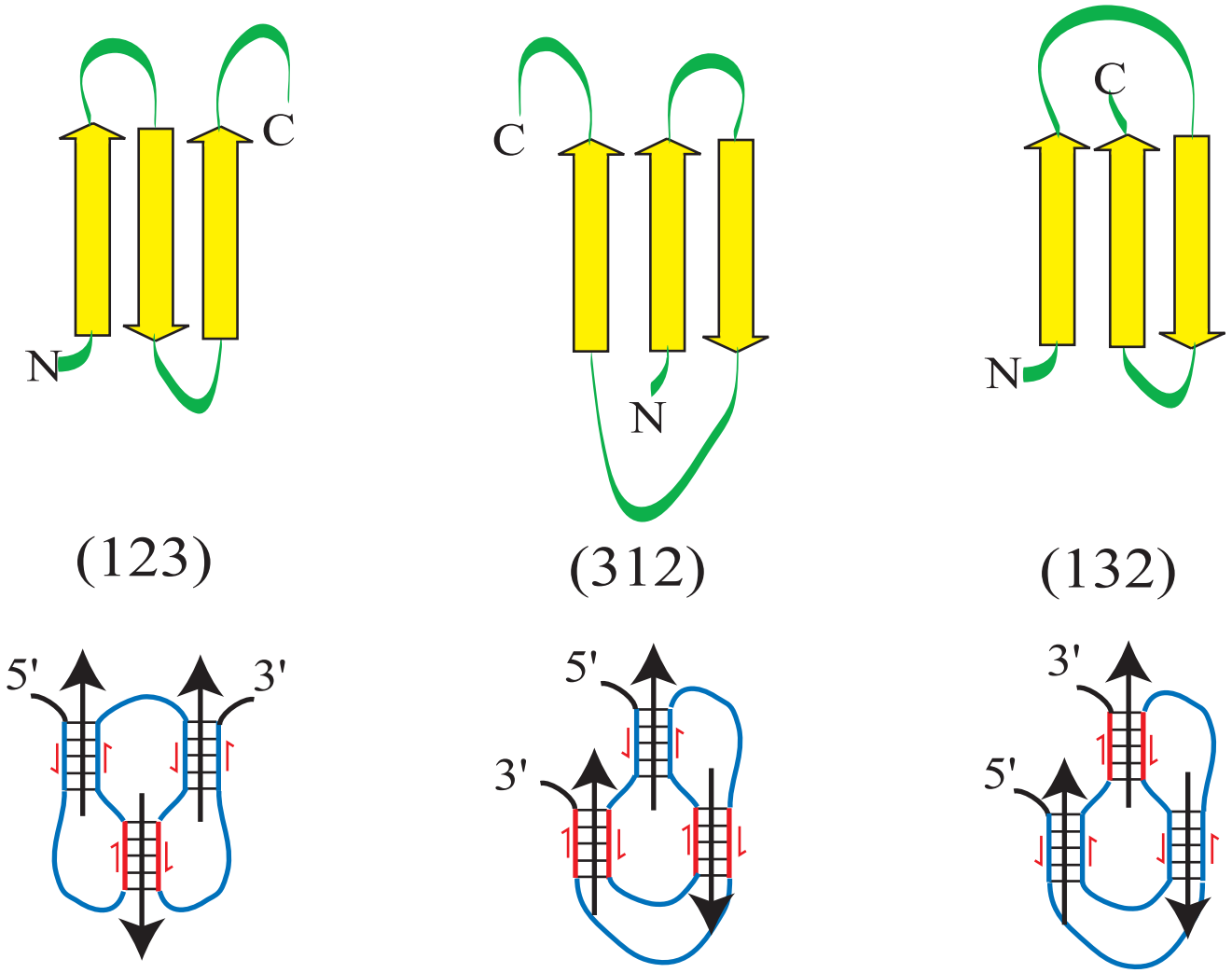

Figure 2: A one-to-one comparison between the patterns generated by 3 protein β-strands and an equivalent arrangement of RNA stems in the form of various RNA pseudoknots. The RNA cannot form a direct neighbor as happens with the beta-strands; however, by shifting them in the pattern of an ABACBC type pseudoknot (and other patterns), a similar pattern of structure can be found. The blue and red stems suggest an order in which the structure might form; blue suggests initially formed stems and red represents subsequent stem formation; see (Dawson et al., 2007) for details.

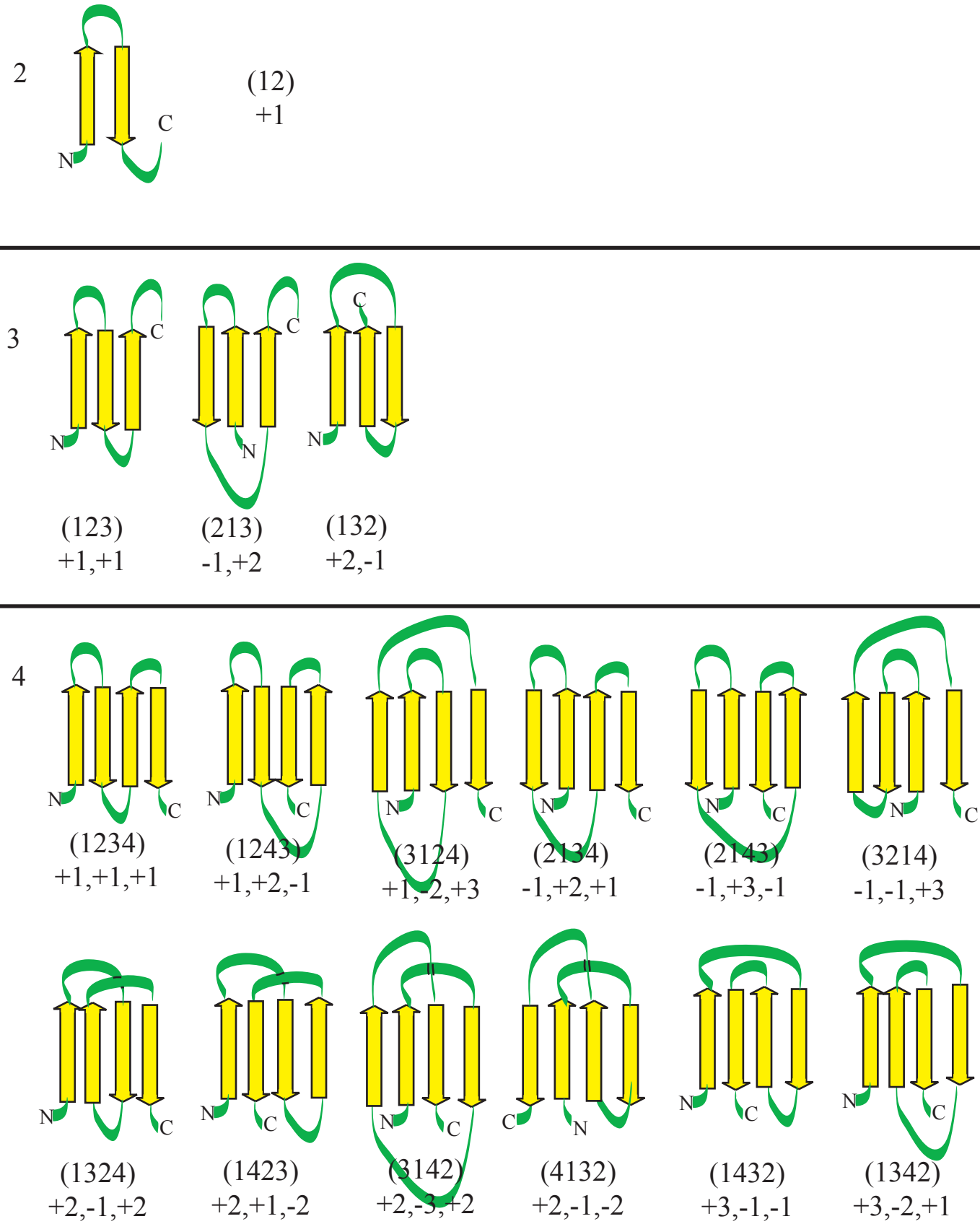

Figure 3: The number of unique arrangements of beta-sheets (excluding crossover parallel beta-sheets) for n=2,3,4, where n is the number of beta-sheets. Because of a plane of 2-fold symmetry, the arrangements show factorial increase of $n!/2$.

In Fig 2, a pattern of 3 stems forming various RNA pseudoknots is compared with the equivalent pattern of β-sheet patterns *that contain no right handed crossovers* (Fig 1a). For RNA, the small red arrows point along the RNA chain in the 5' to 3' direction of the chain. An RNA stem is represented by the contiguous cross hatched regions and indicates base pairing. (The double strand helix of A-RNA is assumed in the ladder shape.) For a chain numbered from 1 to N, the large arrows point toward the tail of the stem, which is located at the base pair closing with the largest 3' index. Comparing the two patterns, they are effectively equivalent. The color coding on the RNA stems is meant to imply the linkage stem (or stems) and the root domain (or domains). A full description of the concept of linkage stem and root domain are discussed in (Dawson et al., 2007). Hence, we focus on this smaller subset displayed in Fig 2. The combinatorics of this subset of protein beta-sheet structures (containing no crossovers) follows a $n!/2$ rule (Fig 3).

Therefore, with no loss of generality in using protein beta-strands, we consider a protein structure Ramachandran plot. We can divide the Ramachandran plot into three major regions α_R (alpha-helix), α_L (left-handed alpha-helix) and β (beta-sheet) sectors. Then the coordination number (q) is 3 and we would say that the number of possible configuration for an N residue protein is 3^{N-1}. However, this is not the only way we can cordon off the structure angles. For example, it would be reasonable to divide the beta-sheet regions into parallel- ($\beta_{\uparrow\uparrow}$) and anti-parallel ($\beta_{\downarrow\uparrow}$) beta-sheets, triple-helix (β_{3x}) and polyproline conformations ($\beta_{\mathrm{Pr\,II}}$) (Adzhubei and Sternberg, 1994; Pappu et al., 2000; Pappu and Rose, 2002). The coordination number is now increased to 6 and the N amino acids (aa) have 6^{N-1} conformations available to them. We could also add 3_{10} helices and beta-turn (β_θ) backbone configurations to the list (Richardson, 1981). Indeed, we could choose many additional ways to cordon off the Ramachandran plot in some mutually exclusive set of sectors; see for example (Pappu and Rose, 2002) where they deduced a partitioning of 10 regions or (Pokarowski et al., 2003) where $q = 12$. The coordination number is not unique.

For $N < q$, it is easy to see that the configurations can overestimate the maximum number of configurations of N free particles ($N!$). For example, let $N = 7$ and $q = 8$, where we chose α_R, α_L, $\beta_{\downarrow\uparrow}$, $\beta_{\uparrow\uparrow}$, β_{3x}, $\beta_{\mathrm{Pr\,II}}$, 3_{10} and β_θ conformations — all reasonable choices for a 7 monomer protein. Then 6! = 720 but 86 = 262144. The fixed coordination number far exceeds the configurations for N free particles.[†] Hence, it is not consistent for $N < q$.

For $N >> q$, it is perhaps less apparent to realize that the fixed coordination number should *underestimate* the number of conformations accessible to a polymer.

Suppose we only permit a subset of combinatorial beta-sheet patterns that contain no crossover (Fig 3) and we make sequences in which the average β-strand plus turn segment is 6 residues. This fully satisfies the structures in Fig 2 and yields $n!/2$ possible configurations (Fig 3). Then we write $n = N/6$ and we compare it with the total number of configurations. Since the arrangement of the β-strands depends on the number of conformations, the total combinatorial number of β-strand arrangements must not exceed the total number of conformations predicted by the lattice model and, indeed, should be much less than that

$$n!/2 = (N/6)!/2 << 3^{N-1} \qquad (2)$$

where we assume a nominal coordination number of $q = 3$. Taking the logarithms of Equation (2) and applying Sterling's approximation, we obtain

$$\ln(n!/2) \approx \ln\left(\sqrt{2\pi}\right) + \left(n+\tfrac{1}{2}\right)\ln n - n - \ln(2) \qquad (3)$$

Rearranging and taking the limit on the dominant terms (with $n = N/6$) yields the following inequality

$$\lim_{N\to\infty} \frac{\left(\frac{N}{6}+\frac{1}{2}\right)\ln\left(\frac{N}{6}\right)-\frac{N}{6}}{N-1} = \frac{1}{6}\left(\ln\left(\frac{N}{6}\right)-1\right) << \ln 3 \qquad (4)$$

Equation (4) cannot be true for all N given a finite segment length containing a beta-strand and turn length (here a sum length of 6 aa). Solving Equation (4) yields $\ln N = 6 \ln 3 + \ln 6 + 1$, or $N = 11980$ aa. This is a very large number; however, for any finite q, Equation (4) cannot be satisfied for some N large enough. Including the prefactor 2^{n-1} in Equation (1) does not satisfy Equation (4) either; in fact, N will become smaller because Equation (3) will grow even faster. Likewise, including left-handed crossovers and mixed right-handed and left-handed crossovers causes the left hand side to blow up even faster still. Therefore, we have a contradiction. Moreover, with factorial growth in the arrangement of beta-strands, the total number of conformations must also be on the order of $N!$ or $q \sim O(N)$ for a

[†]It just so happens that 36 = 729 which is close to 6!; however, there is no physically objective reason to exclude 8, 10, 20 or even more such coordination numbers from the possible list.

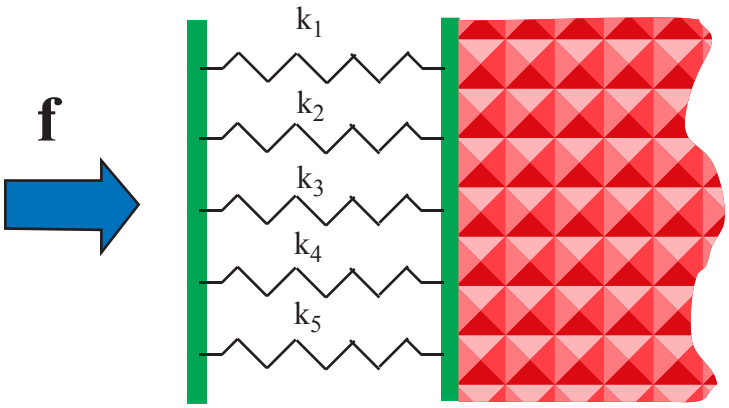

Figure 4: Example of a group of springs arranged in parallel with a force applied along the axis of the spring: k_l (l=1,…5). On the right hand side is a wall and a force f is applied from the left hand side. The response of these parallel springs is the sum of their respective spring constants.

distinguishable arrangement of conformations.[‡]

The number of conformations of a lattice model cannot be universally estimated for all N using a single fixed coordination number (though it may approximate that number *for some N*). Furthermore, the choice of q is not unique. Lattice models were originally intended for crystals where crystal packing certainly defines and limits the orientations. They were applied to biopolymers to *approximate* the expected structure and reduce the computational load. We will show later that we can fix this issue by taking the degeneracy into account.

A Derivation of the Cross Linking Entropy Model

Here we derive the cross linking entropy model (CLE-model). For simplicity, we assume a Kuhn length of $\xi \equiv 1$ mer (App A1).

First we consider the defined parameters in a Gaussian polymer chain (App A4). The extensive parameters (*i.e.*, *measurable*) are the radius of gyration that yields an estimate for the root-mean-square (rms) end-to-end distance (r) and the force (f) acting on the terminal ends of the polymer chain. From the definitions, the heat flow due to the work done by a polymer chain consisting of N monomers with state parameters r and f (in a reversible reaction) is

$$TdS = dU + fdr, \qquad (5)$$

where U is the internal energy, S is the entropy and T is the temperature. For a polymer, $\Delta U \sim 0$ (Flory 1953). Equation (5) takes a form similar to the work done by an ideal gas in which fdr replaces PdV (where V is the volume and P is the pressure and the response behavior is also analogous). The Helmholtz free energy is $F = U - TS$. For the state parameters r and f

$$dF = fdr - SdT \qquad (6)$$

with

$$S = -\left(\frac{\partial F}{\partial T}\right)_r \quad \text{and} \quad f = \left(\frac{\partial F}{\partial r}\right)_T \qquad (7)$$

Then $\partial S / \partial r = -\partial f / \partial T$. From the definition of the GPC (App A4&5), $S(r)$ is independent of T, hence, the virial equation of state has the form

$$f = \left(\frac{\partial F}{\partial r}\right)_T = -T\left(\frac{\partial S}{\partial r}\right)_T \qquad (8)$$

For constant T, Equation (6) becomes

$$-TdS_T = -T\left(\frac{\partial S}{\partial r}\right)dr_T = fdr_T \qquad (9)$$

Using Equation (A16) with $\delta = 2$, $\gamma = 1$, and $\nu = 1/2$ and

[‡]Given the statistical definition of the models, on a square lattice, a pair of successive monomers can overlap on themselves. Hence, one possible pattern is a chain that folds back and forth on itself over and over again. However, besides being entirely unphysical (though allowed in the definition), such patterns are not *distinguishable* because a large fraction of entities will occupy exactly the same positions in the lattice over and over again. Furthermore, *even if we admit such arrangements*, Equation (4) shows that the lattice model still cannot satisfy all the possible configurations for large enough N. This section is dealing strictly in statistics. All the elements are accounted for on the left hand side and no restrictions are placed on the right hand side of Equation (4). Under these conditions, the right hand side *must* satisfy Equation (4).

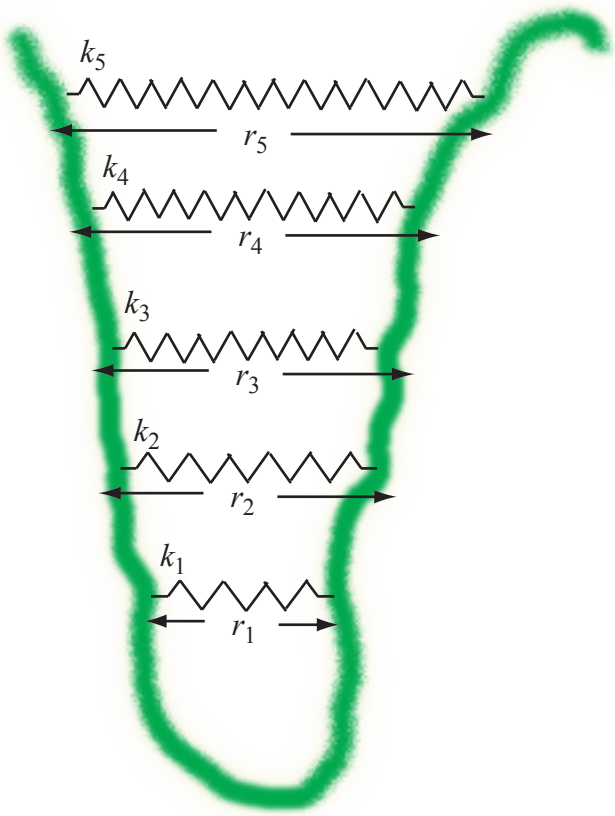

Figure 5: Based on the analogy presented in Figure 4, a group of springs arranged in parallel and set in equilibrium on a polymer chain.

applying Equation (8), the force response between the terminal end-points of a polymer is

$$f(r) = -T\left(\frac{\partial S}{\partial r}\right)_T = 2k_B T\left(\frac{\gamma}{r} - \alpha r\right) \qquad (10)$$

where $\alpha = (\gamma+1/2) / \langle r^2\rangle$, $\langle r^2\rangle$ is the root mean square end-to-end separation distance (Equation (A3)) and k_B is the Boltzmann constant. The minimum for Equation (10) is located at $r_o = \sqrt{\gamma/\alpha}$. This expresses the minimum in the end-to-end separation distance (not the rms-distance $\langle r^2\rangle = \xi N b^2$ indicated in Equation (A3)). When $r < r_o$ or $r > r_o$, a force drives the end-to-end distance back to r_o. For a GPC, $\gamma \equiv 1$ and $r_o = \sqrt{3\langle r^2\rangle / 2}$.

So far, we have only considered the end-to-end distance. However, Equation (10) is not restricted only to the ends of the chain. For any (reasonable) number of residues (N) in a polymer chain, this same $\langle r^2\rangle = \xi N b^2$ relationship holds. First, from Equation (A4), the relationship can be translated. Second, for every length of sequence, this property holds. Hence, it is a general rule that applies to every point on a polymer chain. In short, for any indices i and j, where $i < j, \langle r^2\rangle_{ij} \approx \xi\,(j - i + 1)b^2$.

We now ask what happens if we chose a set of specific coordinate pairs on the chain and apply a constraining interaction on them. The force equation resembles the displacement of a spring. Figure 4 shows a bank of 5 springs aligned in parallel to which a force is applied. The effective spring constant for an array of springs in parallel is the sum of the individual spring constants. Hence, the effective force is an additive property of the effective spring constant times the displacement

$$f = f_1 + \ldots + f_n = (k_1 + k_2 + \ldots + k_n)\,\Delta\chi \qquad (11)$$

where $\Delta\chi$ is displacement $k_1, k_2, \ldots, k_n$ are the individual spring constants, and $f_1, f_2, \ldots, f_n$ the individual contribution of spring k_l (l = 1,2,…n) to the force.

By analogy, we build a similar model in which a polymer chain is folded out to its equilibrium configuration and held in place by an array of springs (Figure 5). Just as pressure pushes against the surface of a container (force/area), so

the equilibrium condition for the correlated motion of the monomers push and pull the contour of the polymer chain back to the equilibrium state. Labeling the interaction force between monomers *ij* with the index *k*, if we now force interaction between any pair of residues *k*, we observe a force f_k in response. The dependence of *i* and *j* on $r_k(=r_{ij})$ and $f_k(=f_{ij})$ is only with respect to the number of residues separating them, and there is no explicit dependence of the other residues on the value of r_{ij} (a general characteristic feature of models like the GPC). Given this behavior, it follows that when *n* such forces are applied, we should expect a similar expression as Equation (11) to emerge: namely,

$$f=f_1(r_1)+f_2(r_2)+\ldots+f_n(r_n) \tag{12}$$

Substituting this with Equation (9), we find

$$TdS=T\left\{\left(\frac{\partial S}{\partial r_1}\right)dr_1+\left(\frac{\partial S}{\partial r_2}\right)dr_2+\ldots+\left(\frac{\partial S}{\partial r_n}\right)dr_n\right\} \tag{13}$$

and equating individual terms with Equation (12) and integrating leads to

$$\int f dr=-\int TdS=\int f(r_1)dr_1+\int f(r_2)dr_2+\ldots+\int f(r_n)dr_n \tag{14}$$

Factoring out the constant *T* and solving for the entropy we obtain

$$\Delta S=\Delta S_1+\Delta S_2+\ldots+\Delta S_n=\sum_k \Delta S_k \tag{15}$$

which is exactly the same as Equation (A23) in App A6 for $\xi \equiv 1$ mer.

Normally, we should expect that $\xi > 1$ mer. Because the CLE model averages the entropy contributions of each interaction over the Kuhn length (ξ), for $\xi > 1$ mer, the entropy in Equation (15) should be scaled by a factor $1/\xi$ (explained in App A and (Dawson et al., 2001a)). Doing so, the result exactly agrees with the expressions found in Equations (A11), (A16-A20) and therefore Equation (A23).

There are at least four independent ways to arrive at Equation (15). In (Dawson et al., 2001a) the CLE-model was derived directly from physical considerations of the entropy and in (Dawson et al., 2001b) it was derived by assuming that each connection leads to the creation of a new loop. In this Section, we have derived Equation (15) from consideration of the force on a chain (which is physically analogous to the pressure of an ideal gas). Equation (15) can also be derived qualitatively from considerations of diffusion.

The CLE Model Satisfies the Coordination Number Using the GPC Model

The entropy of a folded polymer is known to have the form of an integral expression (Dill and Stigter, 1995; Chan and Dill, 1997). Here we show that the summation rule in Equation (15) has the properties of integration and that it satisfies Equations (2 – 4) with the GPC model. We show that the conformations of a polymer chain composed of *N* segments has a coordination number that is a function of *N*; i,e., $q=f(N)$. Hence, under these conditions, the CLE model satisfies both distinguishability and Equation (4).

The summation rule in Equation (15) is a consequence of integrating the correlated interactions (the cross links) in the model. From the derivation of each ΔS_k (App A3, Equation (A7)), ΔS_k represents the probability that state *k* in configuration $r_{k[i]}$ should acquire a configuration $r_{k[f]}$: $p(r_{k[f]} \cap r_{k[i]})\Delta r$, where *p* is the probability and $k \Leftrightarrow (i,j)$ describes the interaction between monomers *i* and *j* (App A, Equation (A3)ff). The entropy $S_k(r_k)$ corresponds to the probability of the configuration r_k: $P(r_k)\Delta r$. The ratio of these

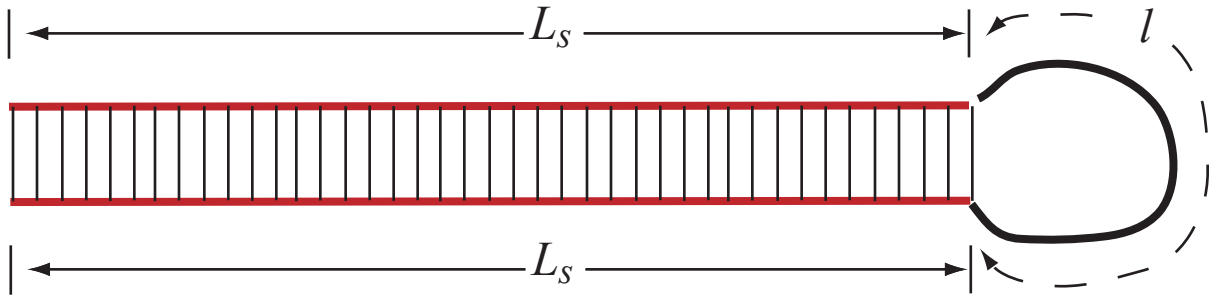

Figure 6: Example of a single hairpin containing L_s base pairs and a loop of length *l* nt. The total length of the sequence is *N*.

states (associated with $r_{k[i]}$ and $r_{k[f]}$) form a conditional probability[§]

$$p\left(r_{k[f]} \mid r_{k[i]}\right) = \frac{p\left(r_{k[f]} \cap r_{k[i]}\right)}{p\left(r_{k[i]}\right)} \approx \frac{p\left(r_{k[f]}\right)}{p\left(r_{k[i]}\right)} \tag{16}$$

Writing Equation (16) in terms of the entropy, we see that Equation (15) is measuring the likelihood that each state k will transition from an initial state [i] to a final state [f]

$$\Delta S = k_B \sum_k \ln\left(\frac{p\left(r_{k[f]}\right)}{p\left(r_{k[i]}\right)}\right) \tag{17}$$

where $k[\ ]$ denotes the state of the interactions between ij. We transform the summation into integration by exchanging the state label k for the enclosed sequence length ($N(k)$)

$$\Delta S = k_B \int \ln\left(\frac{p\left(r_{N(k)[f]}\right)}{p\left(r_{N(k)[i]}\right)}\right) dk = \int \left(S\left(r_{N(k)[f]}\right) - S\left(r_{N(k)[i]}\right)\right) dk \tag{18}$$

Equation (18) calculates the total change in entropy due to forcing a polymer into a specific configuration that is a function of $N(k)$.

In general, Equation (15) is much easier to arrange and evaluate than the integral form in Equation (18). However, for an RNA chain forming a hairpin in a single stem from 5' to 3' (Figure 6) or two anti-parallel beta-strands joined via a loop, the summation in Equation (15) can be easily written as an integral. For the GPC with parameters γ and $\delta \equiv 2$, Equation (15) becomes

$$\Delta S = -k_B \sum_{k=0}^{L_s} \left\{\gamma \ln\left[\psi(2k+l)\right] - \zeta(\gamma+1/2)\left(1 - \frac{1}{\psi(2k+l)}\right)\right\} \tag{19}$$

and, converting to an integral,

$$\Delta S \sim -k_B \int_0^{L_s} \left\{\gamma \ln\left[\psi(2k+l)\right] - \zeta(\gamma,2)\left(1 - \frac{1}{\psi(2k+l)}\right)\right\} dk \tag{20}$$

Now, using the relationship that $N = 2L_s + l$,

$$\Delta S \sim -k_B \left\{\frac{\gamma}{2}\left(N[\ln(\psi N) - 1] - l[\ln(\psi l) - 1]\right) - \zeta(\gamma,2)L_s + \frac{\zeta(\gamma,2)}{2\psi}\ln(N/l)\right\}$$

$$\sim -\frac{\gamma k_B}{2}\left(N[\ln(\psi N) - 1]\right) \sim O\left(\ln(N^N)\right) \to O\left(\ln(N!)\right) \tag{21}$$

Hence, $q \propto N$ easily satisfies Equation (4). Equation (21) can also be derived directly from the summation (see (Dawson et al., 2001a)).[**]

What does this mean?

First, equation (21) shows factorial-order growth with the number of binding pairs (i.e., cross-links) that are formed. The RNA-stem has strong correlation due to the fixed and stabilized structure. This also suggests that the "penalty" for RNA-stem formation should go mostly to stem formation rather than to loop formation for the *coarse-grained* entropy. (The fine grained entropy counts in the loops and in the pairing interactions.) This relationship also would apply to various beta-sheet conformations.

Second, Equation (21) is consistent with the fact that the number of ways that N distinguishable particles can be arranged is $N!$. It is consistent with the Gaussian (and Gamma) function statistics because its *maximum* value is always less than or equal to the normalization constant (Equation (A11)). Moreover, the global entropy is known to be an integral property for this type of system (Dill and Stigter, 1995; Chan and Dill, 1997). Hence, Equation (15) is consistent with the concept of integration and consistent with textbook statistics.

Equation (21) is also consistent with the fact that x-ray

[§]The independence of the conditional probability for this Gaussian model is understood because it can be worked out from a Markov chain rule where successive steps in the configuration depend only on the given configuration at the immediate previous step and are independent on any steps prior to that point (Montroll, 1950; Feller, 1968 and 1971). In other words, knowledge of previous steps is restricted to the state of the current step and the next step to be assigned. We are, therefore, justified to use this strategy on the grounds that it is the definition. Further, the theorem on the subadditivity of entropy assures us that $S_{12} \le S_1 + S_2$. Hence, at worst, we have consistently erred on the side of *overestimation* of the entropy. One can visualize that the effective coupling dies off with distance; hence, for large enough Kuhn length, the Markov model is reasonable. This is the concept of renormalization theory discussed in App A1. Whereas the model can certainly be further refined, it does not change these concepts.

[**]It is true that the combinatorics of RNA *secondary structure* stems follow a 2^{N-1} rule. This is only the RNA *secondary structure*. Furthermore, whereas this explains adequately the computational combinitorics of RNA *secondary structure*, it does not justify equating the combinatorics with the entropy because these systems involve *distinguishable* entities. The combinitorics only consider the *pairing*; not the *distinguishable* relationships or how it got in a particular configuration.

diffraction can *distinguish* these indexed monomers and produces a structural factor proportional to the number of monomer (N). Were the true structures that of a lattice, a coordination number (q) should emerge from the lattice parameters and structural factor of the x-ray diffraction data and most of the monomers in the protein structure or RNA structure could not be uniquely identified and assigned because of this degeneracy. We would observe dispersion akin to a crystal with many defects. We observe unique angles that are distinguishable (e.g., for proteins, the Ramachandran plots all show non-degenerate distinguishable residues).

For lattices that use the self avoiding random walk (App B) with a large enough coordination number, degenerate (but distinguishable) conformations have been observed (Pokarowski et al., 2003). In this case, the example contained approximately 12 residues and the lattice was a face centered cubic (i.e., $q = 12$). Hence, even the lattice model predicts degeneracy when a large enough coordination number is used.

Finally, this model satisfies the inconsistencies in Equation (4). A unique coordination number ($q \sim O(N)$) is always obtained from the CLE model combined with the GPC. [††]

Making the Lattice Model Consistent

We have shown that the CLE model satisfies Equation (4). In this Section, we show how to unify the lattice model and the GPC.

The relationship between the lattice model and the CLE-based GPC can be expressed as a family of equations having the following form

$$C_N \approx \left(\frac{q(N)}{w}\right)^{\alpha(N)} \left(\frac{g(N)}{h(N)}\right)^{\beta} \qquad (22)$$

where $q(N)$, $g(N)$, $h(N)$ and $\alpha(N)$ are increasing functions of N, and w and β are a constants and we have assumed a unit Kuhn length ($\xi \equiv 1$ mer).

For example, if $q(N) = \Psi N$, $g(N) = (\Psi N)^{1/\Psi}$, $h(N)=e^N$, $\alpha(N) = \gamma N$, $w = e$ and $\beta = (\gamma+1/2)$, then

$$C_N \approx \left(\frac{\Psi N}{e}\right)^{\gamma N} \left(\frac{(\Psi N)^{1/\Psi}}{e^N}\right)^{(\gamma+1/2)} \qquad (23)$$

and changing to the logarithm form, we obtain

$$\ln(C_N) \approx \gamma N \ln(\Psi N) - N\left(2\gamma+\frac{1}{2}\right)+\left(\frac{\gamma+1/2}{\Psi}\right)\ln(\Psi N) \qquad (24)$$

The derivative with respect to N is

$$\frac{\partial(\ln(C_N))}{\partial N} \approx \left\{\gamma\ln(\Psi N)-(\gamma+1/2)\left(1-\frac{1}{\Psi N}\right)\right\} \qquad (25)$$

which is the same form as Equation (A19) and easily transforms into

$$\Delta S(N) \approx -k_B\left\{\gamma\ln(\Psi N)-(\gamma+1/2)\left(1-\frac{1}{\Psi N}\right)\right\}.$$

When Equation (22) uses the values $q(N) = N$, $g(N) / h(N) = N$, $\alpha(N) = N$, $w=e$ and $\beta = 1/2$, we obtain

$$C_N \approx \left(\frac{N}{e}\right)^N (N)^{1/2} \qquad (26)$$

which is very close to the asymptotic approximation known as Stirling's formula $N! \approx (2\pi)^{1/2}\ (N/e)^N\ N^{1/2}$, where $N!=1\cdot2\cdot3\cdots N$. [‡‡] The total number of ways one can arrange N *distinguishable* objects is also $N!$ in size.

All the amino acids in a protein (or RNA) are distinguishable using X-ray crystallography or NMR spectroscopy. Such monomers are semi-classical enough in size and mass to obey Maxwell-Boltzmann statistics which are used when computing the statistics of *distinguishable* objects. It follows that the true number of conformations must also be of

[††] In terms of Equation (4), the CLE-model permits N unique angles; hence, all entities in these models are theoretically distinguishable. Nevertheless, in this Section, we currently are still ignoring the fact that there is real physical space involved with a real polymer. This must limit the set of possible conformations and will be addressed later in this work.

[‡‡] The derivation of Stirling's formula is via the Gamma function: $\Gamma(x) = \int_0^\infty t^{x+1}e^{-t}dt$. However, in $\Gamma(x)$, N plays the role of x and the integration of t is from 0 to ∞. In this work, the probability density function and its weight ($(e^{-t})t^{x+1}\ dt$) have the same form (after exchange of variables) as Equation (A12). However, x corresponds to $\delta\gamma$ and the integration is from 1 to N. Therefore, whereas both arrive at almost the same formula, the meaning behind the operations is completely different. A detailed derivation of Stirling's formula is found in Lebedev (1965).

order $N!$ in size. In the previous section, we have shown that the summation in Equation (15) is also a form of integration. Equation (21) leads to Equation (23); whence the number of conformations is of order N^N. Equation (26) shows that this is of similar size to a factorial expression. The CLE model yields a family of equations that are consistent with a system of distinguishable particles.

The lattice model requires that $q(N) = \text{constant} \equiv q_o$, $g(N) / h(N)=1$ and $\beta=1$. With the exception of a range of values around q_o, this does not match conceptually with a simple integration of Equation (25). Neither does it return anything remotely resembling a Gaussian model when we try to evaluate its derivative:

$$d\left[\ln(q_o^N)\right] / dN = \ln(q_o) \tag{27}$$

A variant of the lattice model has the form $C_N \propto q_o^N \ln N^\gamma$ (Arinstein, 2005). This conforms to the second term in Equation (25). However, it still fails to satisfy Equation (4) for N large enough.

What is missing is the degeneracy. For a lattice constant q_o and sequence length N, the degeneracy $\sigma(N)$ is

$$\sigma(N) = \frac{q(N)}{q_o w}\left(\frac{g(N)}{h(N)}\right)^{\beta/\alpha(N)} \tag{28}$$

and so it grows monotonically with N. Equation (28) removes the fact that we generate too many states when $N<<q_o$ and too few when $N>>q_o$. In Equation (28), the coordination number is $q(N)=N$, the root mean square deviation expands to $(g(N)/h(N))^\beta = \sqrt{N}$ and the scaling factor is the exponential base $w=e$. This is a standard Gaussian distribution.

The reason why the lattice model has often been successful is because the sequence lengths that are used are typically of similar order to q_o. The extreme computational costs usually restrict the use of lattice models to $4 < N < 20$ mers. For such cases, Equation (27) has the form $q_o \sim N$ and, as a result, it tends to yield conformations of the approximate order; disguising the issue.

We have shown that Equations (22-25) are consistent with a Gaussian-type model and satisfy the inequality on the right hand side of Equation (4). Equation (27) has the same form as Equations (22-25) and therefore, the CLE can incorporate this model. Equation (4) is satisfied if we weight the coordination number by Equation (28). Therefore, the CLE model embraces both forms and shows the route of transformation between them.

Including Variable Flexibility: the Kuhn Length

Most functional biopolymers have different flexibilities in different parts of their structures that reflect their function and all such polymers have a Kuhn length larger than one mer. A scaffold should be stiff ($\xi \sim 10$ mers) whereas a protein-protein docking region would require "shock absorbers" and flexible interfaces ($\xi \sim 3$ to 5 mers) to help the subunits bind. Mechanical parts require flexible joints ($\xi \sim 3$ mers). All this indicates that we need a model that accounts for different flexibilities of a real polymer.

For the case where $\xi \neq 1$ mer, Equation (22) must be renormalized (see App A1). For $\xi > 1$, Equation (22) is transformed as follows

$$C_N \xrightarrow{\xi>1} \Upsilon_\xi[C_N] = \left\{\left(\frac{q(N)}{w}\right)^{\alpha(N)}\left(\frac{g(N)}{h(N)}\right)^{\beta}\right\}^{1/\xi} \tag{29}$$

where $\Upsilon_\xi[\,]$ functions as an operator for scaling the global entropy (Equation (A18)).

In Equation (25), we obtained the isolated entropy for a particular binding pair (or region of binding pairs) in the biopolymer. Because the conformations and their derivatives are related and separable, we can handle each of these parts separately and add them together according to Equation (15).

We can now generalize these findings. From Equation (15), we can sum the entropies. From Equation (25), the derivatives express the instantaneous long range entropy contribution between mers ij. A full transformation for the CLE model for a given binding pair (bp) configuration (Equation (A23); App A6) becomes

$$\Delta S_{cle}(\xi > 1) = \sum_{\text{all}(\xi_k)}\left\{\frac{\xi_k}{N}\Delta S_{\gamma\delta}(\xi_k)\right\} - \sum_{\text{group}(\xi_k)}\left\{k_B \sum_{bp_k(ij)} \frac{\partial \ln\left(C_{N(ij)}\right)}{\xi_k \partial N(ij)}\right\} \tag{30}$$

where $N(ij)$ corresponds to individual binding pairs, ξ_k refers to successive segments of mers k each of which has a Kuhn length of ξ_k. The first summation of Equation (30) scales the contribution of $\Delta S_{\gamma\delta}(\xi_k)$ (the local coarse-grained

entropy; App A6) for all segments of mers k in the polymer chain and the second summation scales the entropy of a group of binding pairs (the global coarse-grained entropy; App A6), where ξ_k can vary depending on the location of ij. Here we presume that $\xi_k > 1$ mer. Hence, the model is easily adapted to a variable Kuhn length from first principles; unlike either the lattice model or the GPC.

We can now understand from the total entropy that branching in RNA structures *reduces* the entropy loss. Consider two branches of length N_1 and N_2 such that $N_1+N_2 \le N_3$, where N_3 is the closing point of the two branches. It is clear that even $q_o^{N_1}.q_o^{N_2} = q_o^{N_1+N_2} \le q_o^{N_3}$, surely therefore, $N_1^{N_1} \cdot N_2^{N_2} << N_3^{N_3}$. Hence, Equation (22) shows that branching is a way to reduce entropy loss in a complex structure. We should *expect* multibranch loops in slowly folding polymers like RNA to branch if there is any reasonable option to do so. It is possible therefore, to scale these contributions independently allowing a variable Kuhn length within the same structure via Equation (30), yielding a variable flexibility in the final structure.

For a pure lattice model where no correction for degeneracy is necessary, $\Delta S_{\xi\gamma\delta}$ involves a small correction proportional to $\ln(q_o)$. The entropy in Equation (30) is then

$$\Delta S_{cle}(\xi > 1;\ \text{pure lattice model}) = \frac{k_B}{\xi}\ln(q_o) - \frac{k_B}{\xi}\sum_{bp(ij)}\ln(q_o) \qquad (31)$$

which is a linear function of N: (N-1)$\ln(q_o)$ (White et al., 2005).

Using the CLE model, not only have we found a way to transform the lattice model so that it is consistent, we have shown how to evaluate a lattice when the structure has a variable flexibility.

Squeezing a more Realistic Model from the Boundaries

In previous Sections, we have already made issue with the Markov chain approximation used in the GPC-model and the lattice model. The lattice model and the GPC are merely statistical models that ignore the physical realities of the systems they model. These models have largely been successful because the physically impossible configurations just so happen to have a small enough probability that ignoring their "possibility" does not significantly affect many results. Nevertheless, outrageously absurd configurations can be imagined that become ever more possible with increasing length. Hence, an attrition of such conformations is expected particularly for large N. Here we consider how to build a more realistic model for estimating the total number of conformations of a biopolymer that considers real polymers with self avoiding interactions, coordination limits and chain-winding limits.

Equation (23) and (29) express a family of equations to which both models belong. We have seen in Equation (4) that the lattice model can underestimate the number of conformations for large sequence length. Likewise, because the Markov model involves non-interacting particles, the GPC-model can overestimate the true number of conformations. Therefore, we can set bounds on the solution.

The basic lattice model permits folding back on itself. This is certainly physically impossible and should be removed from the set of possibilities. This is addressed by considering a self avoiding chain. Since folding back on the same chain is forbidden in this model, the coordination number (q_o) must necessarily be reduced. We therefore introduce an effective coordination number $\tilde{q}$ for the lattice where $\tilde{q} < q_o$ (Sykes, 1963). See App B for an explanation of how to estimate an effective coordination number.

The upper bound is the GPC model. For the GPC model, the self avoiding walk is often approximated using the lattice model results of (Fisher, 1966), where the exponent (γ) on the volume term of Equation (A12) or the logarithmic term of Equation (25) or (A19) is increases from 1.5 to 1.75 in 3 dimensions. The consequence of a self avoiding walk is that it tends to increase the volume of the polymer.

We begin by assuming that the Kuhn length (ξ) is 1 mer for the GPC. There is no loss of generality in this assumption because the "lattice" can also be set to have the same spacing as the Kuhn length so that the same boundaries apply. For a lattice constant different from the Kuhn length, Equation (30) scales the lattice model accordingly since the Kuhn length tends to freeze out the degrees of freedom of the monomers.

The true solution must lie between these two bounds such that

$$\tilde{q}^{N-1} \le C_N^T \le \tilde{C}_N^\Gamma;\ \tilde{q} < N \qquad (34a)$$

$$C_N^T \sim \tilde{C}_N^\Gamma \le \tilde{q}^{N-1};\ \tilde{q} > N \tag{34b}$$

where $\tilde{C}_N^\Gamma$ is the adjusted Gaussian solution (Equation 23) and C_N^T is the true solution.

According to Equation (22), $q(N)$ should be an increasing function of N. To satisfy Equation (34), we choose

$$q(N) = (\Psi N)^{\nu\delta} \tag{35}$$

where Ψ is a constant, ν is an excluded volume weight, and δ ($0.5 \le \delta \le 2$) is the weight on the exponential function (see App A5 and (Dawson et al., 2006)). For the standard GPC-model $\nu = 1/2$ and $\delta \equiv 2$. When $\delta < 2$, the weight on N decreases, and if the system is globular, $\nu \to 1/3$, this further decreases the weight. Setting $\alpha(N)=\gamma N$, $q(N)=(\Psi N)^{\delta\nu}$, $g(N)=\exp\{(\Psi N)^{1-\delta\nu}/(1-\delta\nu)\Psi\}$, $h(N)=e^N$, $w = \exp(\delta\nu)$, and $\beta = \zeta(\gamma,\delta)$ in Equation (22), the derivative of the result (for $\delta\nu < 1$) [§§] is

$$\frac{\partial\left[\ln(C_N^T)\right]}{\partial N} = \left\{\nu\delta\gamma\ln(\Psi_\nu N) - \zeta(\gamma,\delta)\left(1-1/(\Psi_\nu N)^{\delta\nu}\right)\right\} \tag{36}$$

where $\Psi_\nu = \xi^{-1}(\xi/\lambda)^{1/\nu}$, $\zeta(\gamma,\delta)=[\Gamma(\gamma+3/\delta)/\Gamma(\gamma+1/\delta)]^{\delta/2}$ (from Equation (A14)) and $\xi = 1$ mer. The form of Equation (36) is identical to that given in Equation (A19).

When we apply Equation (29) for $\xi > 1$ mer, $\alpha(N) = \gamma N/\xi$ and $\beta=\zeta(\gamma,\delta)/\xi$ (where $q(N)$, $g(N)$, $h(N)$ and w are unchanged), we obtain the exact expression in Equation (A19)

$$\Delta S = -\frac{k_B}{\xi}\frac{\partial\left[\ln(C_N^T)\right]}{\partial N} = -\frac{k_B}{\xi}\left\{\nu\delta\gamma\ln(\Psi_\nu N) - \zeta(\gamma,\delta)\left(1-1/(\Psi_\nu N)^{\delta\nu}\right)\right\} \tag{37}$$

which indicates that ξ is scaling the conformations (and therefore also the entropy) by the effective mers. Reductions to γ, particularly on the logarithmic term of Equation (A19), would further reduce this number of conformations from the standard GPC-model. This offers a far better description of the actual number of conformations.

The weight δ is a measure of the long range correlation where $\delta = 2$ (Gaussian) reflects localized or independent coupling, $\delta = 1$ (exponential) reflects diffusive coupling and $\delta = 1/2$ (exponential square root) reflects a glassy unstructured coupling. Because the polymer chain requires real physical length considerations in evaluating this coupling, there is certainly a "diffusive" component in the structure in the sense that the correlation extends over a far longer range than would occur if the polymer chain was non-interacting. Consequently, this reduces the number of degrees of freedom and independence of each effective mer. In general, most biopolymers that we have studied so far tend to fall in the range $1 \le \delta \le 2$. The parameter ν (App A) tends to be less than 1/2 in globular proteins (Grosberg and Khokhlov, 1994) suggesting that $\nu\gamma\delta < 1$; i.e., the correlation is glassy. By proper partitioning of this function, one could even introduce a variable δ or γ to this problem. In addition to regions of variable flexibility, some biopolymers are believed to have disordered regions and globular regions as well; hence "squeezing" offers additional options for future exploration.

In this Section, we have shown that we can "squeeze" the correct solution between limits; the lattice model on the one end and the GPC at the other end. The true conformation limits on folding a beta-strand back and forth can be largely accounted for by including a weight δ on the logarithmic term of Equation (36) because the solution is bounded between the two extremes (Equation (34)). "Squeezing" is convenient starting point for developing tractable statistical models that considers the steric effects and long range correlation contributions that are ignored in statistical Markov chain based models (Montroll, 1950; Feller, 1968 and 1971).

Incorporating the Worm Like Chain Model into the CLE Model

As shown in (Dawson et al., 2001a), the logarithmic function in Equation (37) represents the resistance of a polymer to compression and the remaining function is associated with the stretching of a polymer chain. The stretching term is important to comment on.

The function in Equations (36) and (37) is the generalized treatment of the probability based on a Gamma function

[§§] The expression for $g(N)$ is also true for $\delta\nu = 1$. To see that it is so, one can integrate the following inequality $1/x^{1+\varepsilon} \le 1/x^{1-\varepsilon}$ between fixed limits a to b ($a < b$) and then bring $b - a$ arbitrarily close to zero as $\varepsilon \to 0$. It therefore follows that when $\delta\nu = 1$, the assembled components of the expression $g(N)^\beta$ will contain the argument $(\Psi N)^{1-\delta\nu}/(1-\delta\nu)$ which must approach $\ln(\Psi N)$ as $\delta\nu \to 1$ and this results in the Gaussian expression found in Equation (23). It is therefore part of the same family of equations. The equation is also true for $\delta\nu > 1$; however, the result can even exceed the GPC model which already overestimates the true number of conformations. This case may be valid for denatured proteins and RNA where the solvent could become part of the "conformations" (in effect). This case is not considered here.

(and its derivative). The GPC does not properly model changes in the entropy due to stretching the chain to a point approaching the contour length. The solution for the worm like chain model (Marko and Siggia, 1995) – also known as a Porod-Kratky Chain (Flory, 1969) – weights the stretching term ($g^{\beta}(N)/\ h^{\beta}(N)$) with far greater accuracy.

The force response for the worm like chain is shown by (Marko and Siggia, 1995) to be

$$f = -T\left(\frac{\partial S}{\partial r}\right)_T = \frac{k_B T}{A}\left(\frac{r}{L} + \frac{1}{4(1-r/L)^2} - \frac{1}{4}\right) \tag{38}$$

where A is the persistence length ($A \approx \xi b/2$; (Flory, 1969)), L is the contour length ($L=Nb$; Equation (A2)) and we have used the relationship in Equation (8). Neglecting bending and over-stretching issues of DNA (Rouzina and Bloomfield, 2001ab) etc., the entropy can be approximated by integrating the Equation (38) with respect to r, yielding

$$\tilde{S}_{\kappa}(r) = -\frac{k_B}{A}\left(\frac{r^2}{2L} - \frac{r}{4} + \frac{L}{4(1-r/L)}\right) + C \tag{39}$$

where $\tilde{S}_{\kappa}(r)$ emphasizes the "spring like" contribution to the entropy and C is an integration constant.

Equation (38) is scaling the system to N/ξ links with a persistence length A. The CLE model is defined by each mer and there are ξ mers in the Kuhn length (for the usual case where $\xi > 1$ mer). To transform this to a mer-equivalent (r_{ij}) expression, we must scale Equation (39) by a weight $1/\xi$. Therefore, for a single binding pair ij;

$$S_{\kappa}(r) = -\frac{k_B}{\xi}\left(\frac{r^2}{2AL} - \frac{r}{4A} + \frac{L}{4A(1-r/L)} + C\right) \tag{40}$$

Since there is a group of n_g ($=\xi$) binding pairs in a given link of the chain, the entropy is unchanged when we consider the average entropy of the group; $(\tilde{S}_{\kappa}(r_g)/\xi)n_g = \tilde{S}_{\kappa}(r_g)$, where r_g is the effective averaged position of the group. The CLE model averages the contribution from each binding pair of mers ij in the group.

Let

$$A \approx \xi b/2 \text{ and } L \rightarrow L_{ij} = N_{ij} b \tag{41}$$

Then Equation (39) transforms to

$$S_{\kappa}(r) = -\frac{k_B}{\xi}\left\{\frac{r_{ij}^2}{\xi N_{ij} b^2}\left(\frac{(3/2)N_{ij}b - r_{ij}}{\left(N_{ij}b - r_{ij}\right)}\right) - \frac{k_B}{2\xi} + C\right\};$$

$$r_{ij} < N_{ij} b \cdot^{***} \tag{42}$$

From the definition of the reference state in Equation (A17), we obtain

$$\Delta S_{\kappa}(N_{ij}) = S_{\kappa}(\lambda b) - S_{\kappa}\left(\langle r^2\rangle_{ij}^{1/2}\right)$$

$$= -\frac{k_B}{\xi}\left\{\frac{(3/2)N_{ij}b - \lambda b}{\left(N_{ij}b - \lambda b\right)}\left(\frac{(\lambda b)^2}{\xi N_{ij} b^2}\right) - \frac{(3/2)N_{ij}b - \langle r^2\rangle_{ij}^{1/2}}{\left(N_{ij}b - \langle r^2\rangle_{ij}^{1/2}\right)}\left(\frac{\langle r_{ij}^2\rangle}{\xi N_{ij} b^2}\right)\right\} \tag{43}$$

Since for large N_{ij}, both $\left\langle r_{ij}^2\right\rangle = \xi N_{ij} b^2 << L_{ij}^2$ and $\lambda b << L_{ij}$, Equation (43) quickly simplifies to

$$\Delta S_{\kappa}(N_{ij}) \approx \frac{3k_B}{2\xi}\left\{1 - \frac{\lambda^2}{\xi N_{ij}}\right\} \tag{44}$$

which is exactly the second expressed term of Equation (A17).

For structure prediction problems, the stretching contribution can to some extent be neglected. The Jacobson-Stockmayer model is a prime example (Jacobson and Stockmayer, 1950). However, if one were to consider the same situation in which multiple points were pulled apart, we must include the independent contributions of Equation (43). It should be clear that the so-called "Gaussian" contribution does not adequately address this issue because it allows the chain to extend to infinite length. Equation (43) is in far more reasonable agreement with the anticipated behavior of a real polymer when stretched out to length L.

For the case of stretching, we can also use Equation (29). Consider a chain that is stretched from the equilibrium position $r_{[i]} = b\sqrt{\xi N}$ (App A2) to some significant fraction of its contour length $\rho_{[f]} = r_{[f]}/(Nb)$, where [i] refers to the initial and [f] the final state of the system and $r_{[i]} < r(=r_{[f]}) < Nb$.

***In the definition of the entropy in Equation (A11), the GPC model has the limits $0 < r < \infty$. For the worm like chain model, these limits must change to $0 < r < L$.

Using $\rho_{[i]} = r_{[i]} / (Nb) = \sqrt{\xi / N}$, $\rho = r / Nb$ and integrating Equation (43) with respect to r using the states [i] and [f], we obtain

$$\Delta S_{\kappa}(\rho) = S_{\kappa}(r) - S_{\kappa}(r_{[i]})$$
$$= -\frac{k_B}{\xi}\left\{\frac{3-2\rho}{1-\rho}\left(\frac{\rho}{\rho_{[i]}}\right)^2 - \frac{3-2\rho_{[i]}}{1-\rho_{[i]}}(1)\right\} = \frac{k_B}{\xi}\left\{\tau(\sqrt{\xi / N}) - \frac{r^2}{\xi N b^2}\tau(\rho)\right\} \quad (45)$$

where $\tau(\rho)=(3-2\rho) / (1-\rho)$.

We now have the means to seek an equivalent expression for stretching the GPC toward its full extension: $\rho \to 1$. To define $g(N)$ and $h(N)$ in Equation (29), the terms on the right hand side of Equation (45) are integrated. Integrating with respect to N and using the substitution $N=r/(\rho b)$ while holding r and b constant, this yields

$$g(N) = \exp\left\{-\int \frac{r^2}{\xi b^2 \rho}\tau(\rho) d\rho\right\} = \exp\left\{-\frac{r^2}{\xi b^2}\left[3\ln\left(\frac{r}{Nb}\right) - \ln\left(1-\frac{r}{Nb}\right)\right]\right\} \quad (46)$$

and

$$h(N) = \exp\left\{-\int \tau\left(\sqrt{\xi / N}\right) dN\right\} = \exp\left\{-2\xi \ln\left(\sqrt{\frac{N}{\xi}} - 1\right) - 3N - 2\xi\sqrt{\frac{N}{\xi}}\right\} \quad (47)$$

where the remaining terms are $q(N) = r / (b\sqrt{\xi N}) = N / \xi$, $\alpha(N)=-\gamma N$, $w=e$ and $\beta=1$. Equation (29) becomes

$$C_N \approx \left(\frac{r}{e\sqrt{\xi N} b}\right)^{-\gamma N} \left(\frac{\exp\left\{-\frac{r^2}{\xi b^2}\left[3\ln\left(\frac{r}{Nb}\right) - \ln\left(1-\frac{r}{Nb}\right)\right]\right\}}{\exp\left\{-2\xi \ln\left(\sqrt{\frac{N}{\xi}} - 1\right) - 3N - 2\xi\sqrt{\frac{N}{\xi}}\right\}}\right) \quad (48)$$

where $r \to Nb$ is assumed. From Equation (48), the entropic response of a chain stretched out to its contour length from the equilibrium position $r_{[i]}$ is

$$-\frac{k_B}{\xi}\frac{\partial[\ln(C_N)]}{\partial N} = -\frac{k_B}{\xi}\left\{-\gamma \ln\left(\frac{\sqrt{\xi N} b}{r}\right) - \tau(\sqrt{\xi / N}) + \tau(r / Nb)\left(\frac{r^2}{\xi N b^2}\right)\right\} \quad (49)$$

where $0 < b\sqrt{\xi N} < r < Nb$. Equation (49) yields an expression that handles stretching. As $r \to Nb$, the dominant term in Equation (49) is $\tau(r / Nb)$ and the logarithmic term can be basically neglected.

We have shown that we can incorporate the worm like chain model directly into the model in a seamless fashion and therefore drastically improve the stretching domain predictions of the CLE model. This is because the stretching and compression components are decoupled. We have therefore shown that the CLE model is not only universal; it is highly versatile as an entropy estimation scheme for biopolymers. Moreover, this shows that the weight of the stretching term need not be precisely a Gaussian weight; even an alternative constant weight is allowed because the compression and extension components are separable.

The Virial Equation of State and the Contact order Model

In Equation (8), we introduced the virial equation of state. Here we examine the equation of state of an ideal polymer in the context of the CLE model.

To tie this to familiar concepts, we first construct the equation of state for an ideal gas. An ideal gas consists of non-interacting particles. For such a gas, the measured parameters P, V and T represent, respectively, the average values for the pressure, volume and temperature of a gas consisting of N gas particles. There are so many gas particles in a normal volume that we simply cannot measure each one; instead, we measure their average collective properties. We can say effectively that each gas particle in a vessel occupies an average fractional volume V/N and if we leave P free, then P depends on N, V and T. For an ideal gas, the Helmholtz equation is $F = c_v T - N k_B T \ln(V/V_o)$, where c_v is the specific heat at constant volume and V_o is a reference state volume. The virial equation of state for an ideal gas is immediately obtained

$$P = \left(\frac{\partial F}{\partial V}\right)_T = \frac{N k_B T}{V} \quad \text{or} \quad PV = N k_B T.$$

In a similar way, using Equation (8) and referring to Figure 5, we can construct an average $\bar{r}$ and an average $\bar{f}$ such that

$$\bar{f} = \left(\frac{\partial F}{\partial \bar{r}}\right)_T = -T\left(\frac{\partial S}{\partial \bar{r}}\right)_T = N_{bp} 2 k_B T\left(\frac{\gamma}{\bar{r}} - \alpha\bar{r}\right)$$

where N_{bp} is the number of binding pairs (base pairs in this

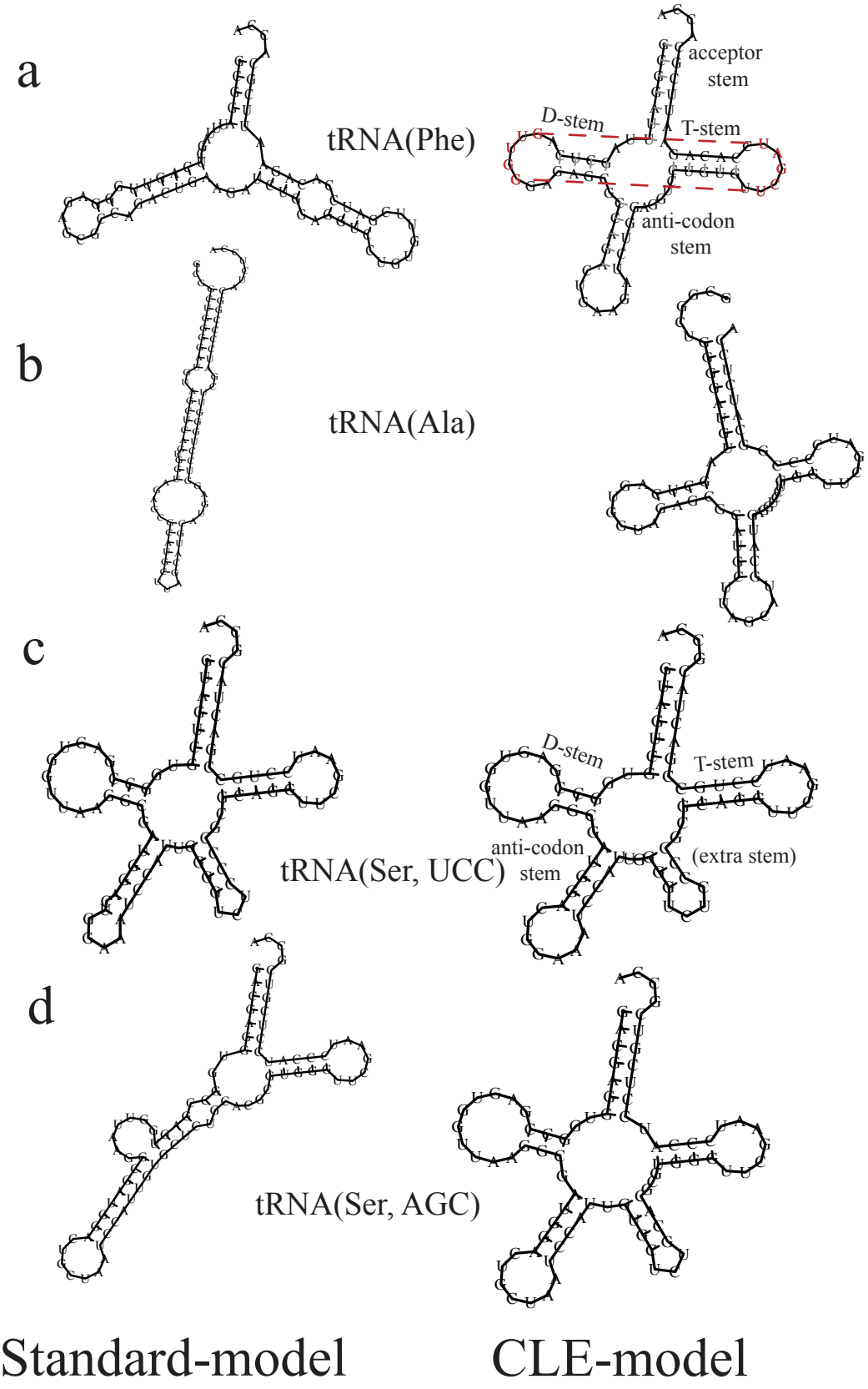

Figure 7: A comparison of the predicted minimum-free-energy secondary-structures of tRNA using the standard-model (left) that neglects global interactions and the CLE-model (right) that incorporates them. The base-pairing thermodynamic parameters are identical for both calculations. (a) Optimal secondary structure predictions of tRNA(phe) for E. coli. (b) Optimal secondary structure predictions of tRNA(ala). (c) Optimal secondary structure predictions of tRNA(Ser) corresponding to codon UCC. (d) Optimal secondary structure predictions of tRNA(Ser), codon AGC.

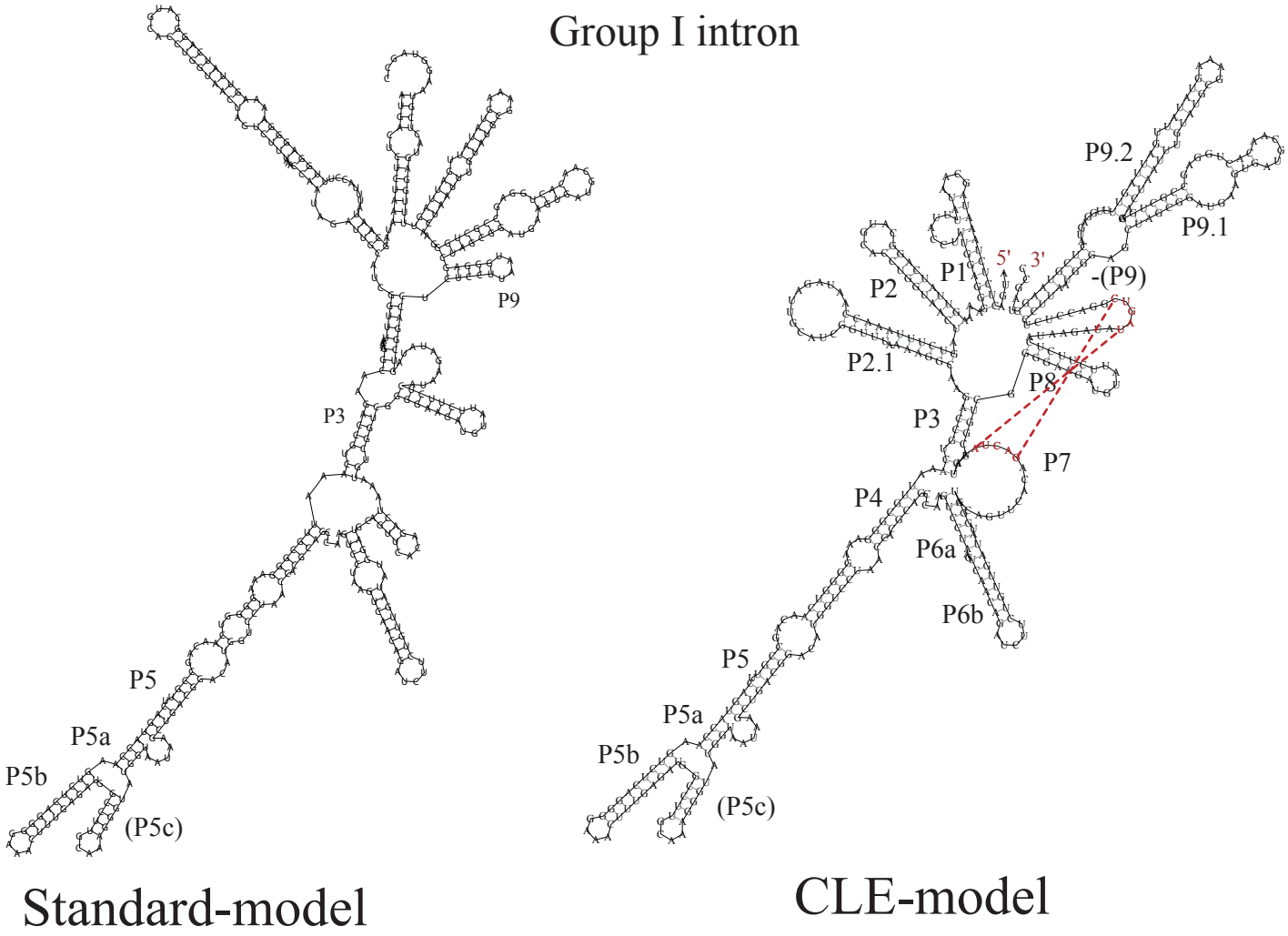

Figure 8: Comparison of the optimal secondary structure of the Tetrahymena thermophila group I intron (the L-21 ScaI ribozyme) using the standard-model (left) and the CLE-model (right).

case) and $\bar{r}$ would be taken as roughly the midpoint of the stem shown in Figure 5 and reflects the collective interaction of all the base pairs forming the single stem of the folded RNA molecule in Figure 5. Extrapolating to more complex structures, a single domain can be defined by $\bar{r}$ and the observed behavior of the system will depend mostly on the largest domain. Hence a biopolymer would have some $\bar{r}$ such that $\bar{r} = (1/2)\max\{r_{ij}\}$.

For both the ideal gas and the ideal polymer discussed here, the contributions to P and $\bar{f}$ are due to the sum of the interactions of all the components in the system. For the ideal gas, this was just added up by multiplication. For the ideal polymer, we have to sum the binding pair contributions individually. Using the average values for $\bar{r}$, $\bar{f}$ and N_{bp}, the ideal polymer equation can also be expressed in the same form as the ideal gas.

The variable $\bar{f}$ is closely connected with the contact order model, where the rate determining folding time is established by $\max\{r_{ij}\}$ (Ivankov et al., 2003). The maximum in the entropy is correlated with the largest value N_{ij}. This means that the folding time of the largest domain will be the rate limiting step. We have shown that the contact order model is a form of the virial equation of state and therefore expresses the average equation of state for the system. Therefore, The CLE model has the contact order model within its interpretation framework.

To Experimentalists

We have derived and discussed — at length — a theory that supports modeling the coarse-grained entropy of biopolymers. We have shown that the existing models are subsumed and extended under the theoretical framework of

the CLE model. Here we explain *why* experimentalists should want to understand the coarse-grained entropy we try to model.

First, the Kuhn length (ξ) is rarely mentioned in most studies of biopolymers, yet flexibility is known to be very important in functional proteins and RNA molecules. For typical protein structures or folded single strand RNA (ssRNA) structures, we can assume that $3<\xi \leq 10$ mers. However, double strand RNA or DNA (dsRNA/dsDNA) can easily show $\xi >200$ mers; the very same linear sequence has two drastically different Kuhn lengths (i.e., flexibilities). Similarly, fibrous proteins (Lehninger, 1975) like collagen (a major component of tissue consisting of a triple helix of amino acids) and α-keratin (found in hair with an alpha helix) have a long Kuhn length. A short fragment of such an amino acid sequence looks similar to many protein fragments or peptides. Why does ξ change?

Second, aggregation is what can happen when you boil or denature any of these biopolymers. We also know of plaques that form in neurodegenerative diseases. It is actually quite easy to produce aggregation in an amino acid sequence: indeed, it seems more difficult to produce amino acid sequences that don't easily aggregate (He et al., 2008). Perhaps natural selection has already filtered out most of these dysfunctional amino acid sequences from the gene pool and what we see is a small subset of the actual possibilities. Why is aggregation so common?

Third, we know that there are domains in folded proteins and RNA. These are typically of the order of 200 to 500 monomers, though some are larger. What process limits this size?

Fourth, what are the coarse-grained differences between protein-protein binding and protein folding?

Equation (30) reveals a large part of where these features come from. First, for dsRNA and fibrous proteins there are no loops. The second term can be neglected in first approximation. In the absence of any well defined tuning from natural selection, the entropy cost of a functional domain is non-linear (Dawson et al., 2001ab)

$$\Delta S(N) \approx -NC(\xi) - \frac{p_{bp} N k_B}{2\xi} \ln(N)$$

where p_{bp} is the fraction of paired monomers in the domain and $C(\xi)$ is the Kuhn length corrections contained in the first term of Equation (30). Like $C(\xi)$, the enthalpy tends to be local and linear in contribution. Moreover, the primary contributors to the enthalpy are the statistical pairing potentials (Zhang et al., 1997; Mintseris and Weng, 2003) that only grow linearly with the presence of pairing interactions. Hence, on the whole, it is typically far less expensive in entropy to combine these biopolymers in fibrils than to form complex folds. Aggregation is far easier than well ordered and expensive structural folds. It is more economical to dock many proteins together than to fold up a single complex functional protein. Indeed, according to Equation (30), it is hardly surprising that amyloid proteins form plaques, rather, it is surprising that they don't.

Yet ignorance abounds. In some *biophysics* meetings, only two or maybe three people even mention persistence length or Kuhn length. Flexibility receives honorable mentioned, but its application to the design and properties of biopolymers is essentially ignored because there is no global concept of entropy. One can see many people who treat the entire domain of a protein or an RNA molecule with the same type of additive statistical pairing potential as if there is no difference between biopolymers that fold, form fibrous structures or dock. Occasionally, there is mention that a global effect may confound the prediction (Zhang et al., 1997), but that is as far as it goes. Hardly anyone seems to find it strange that proteins can so easily aggregate. Lattice models and worm like chains models are used on the same protein yet no one even asks how the same protein can have such different entropies. If we do so fallaciously on the global coarse-grained scale, how in the world can we expect to get the fine grained details right?

Qualitatively, the CLE model can certainly explain these properties. In our previous work, we have also shown that in at least some important cases, the CLE model can quantitative address these issues (Dawson et al., 2007) and provide structures that are predicted at the minimum free energy. Some solutions for RNA folding are quite stable and hardly difficult to hit on with the CLE-model. Figures 7 and 8 show some examples of the predicted minimum free energy structures for tRNA and the group I intron respectively for the standard model that neglects these global contributions and the CLE model that considers these interactions. The local statistical-thermodynamic potentials are identical in these calculations; where we used the the Mfold 3.0 data set (Mathews et al., 1999). The standard model results

are calculated using the Vienna Package RNAfold version 1.4 and the CLE model calculations are done using vsfold5 and vsfold4 (Dawson et al., 2006; 2007). This clearly shows that it is possible to use statistical-thermodynamic pairing potentials and predict a minimum free energy structure that approaches the native state structure for the RNA molecule. For tRNA, we observed 80% success in a complete genome of RNA (Ito N, unpublished data). Preliminary protein calculations also show promise (Dawson et al., 2005).

Success is not guaranteed. For one thing, currently, there is no way to know what the Kuhn length should be for a particular problem, and therefore, we usually have to make an educated guess. There are clear differences in the behavior of the pairing potentials such as the Mfold 2.9 data set (Freier et al., 1986). This shows local interactions are important in these problems too. Likewise, there are indications that the GPC formulation could use different weights for g and h in Equation (29). Hence, what tuning should be applied to the CLE approach is still not completely clear. This suggests that more needs to be done with statistical pairing potentials in the context of the global entropy. Therefore, there is more work to be done. However, the model has consistently offered a fighting chance and has already shown that it can overcome many obstacles and progress onward.

In this work, we have provided a foundation that unifies the lattice model, the worm-like chain model, the Gaussian polymer chain model, and the contact order model under one framework. Clearly, each of these models has hit around the right answer for the coarse-grained entropy of polymers. This work does not solve every aspect of this problem. Nevertheless, the method presented here is a powerful tool for guiding us on how to ask the right questions.

Acknowledgments

This work was supported by a Grant-in-aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT). We thank Elliot Lieb for pointing us to the subadditivity of entropy theorem and Michael Zuker for asking WD "why does the lattice model fail when $q > N$". WD would like to thank Neil A McDougall (theoretical particle physicist) Kenji Yamamoto (International Medical Center of Japan), Greg Rose (business consultant), Craig Stevens (software engineer) and Yucong Zhu (optical engineer) and Bejon Kumar Bhowmick (bioinformatics) for their encouragement and discussions.

References

1. Adzhubei AA, Sternberg MJ (1994) Conservation of polyproline II helices in homologous proteins: implications for structure prediction by model building. Protein Science 3: 2395-2410.

2. Arinstein AE (2005) Uniaxial ordering and rotator phase of ribbonlike polymers. Phys Rev E 72: 051806.

3. Ashcroft NW, Mermin ND (1976) Solid State Physics. Philadelphia, Saunders College.

4. Baldwin RL, Rose GD (1999) Is protein folding hierarchic? I. Local structure and peptide folding. Trends in Biochemical Sciences 24: 26-33.

5. Baldwin RL, Rose GD (1999) Is protein folding hierarchic? II folding intermediates and transition states. Trends in Biochemical Sciences 24: 77-83.

6. Chan SC, Dill KA (1997) Solvation: how to obtain macroscopic energies from partitioning and solvation experiments. Annual Review Biophysics and Biomolecular Structucture 26: 425-59.

7. Chen SJ (2008) RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annu Rev Biophys 37: 197-214.

8. Cohen FE, Sternberg MJ, Taylor WR (1982) Analysis and prediction of the packing of alpha-helices against a beta-sheet in the tertiary structure of globular proteins. Journal of Molecular Biology 156: 821-62.

9. Dawson W, Fujiwara K, Kawai G (2007) Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. PLoS One 2: 905.

10. Dawson W, Fujiwara K, Kawai G, Futamura Y, Yamamoto K (2006) A method for finding optimal RNA secondary structures using a new entropy model (vsfold). Nucleotides, Nucleosides, and Nucleic Acids 25: 171-189.

11. Dawson W, Kawai G, Yamamoto K (2005) Modeling the long range entropy of biopolymers: A focus on protein structure prediction and folding. Recent Research Developments in Experimental & Theoretical Biology 1: 57-92.

12. Dawson W, Suzuki K, Yamamoto K (2001) A physical origin for functional domain structure in nucleic acids as evidenced by cross-linking entropy: part 1. Journal Theoretical Biology 213: 359-86.

13. Dawson W, Suzuki K, Yamamoto K (2001) A physical origin for functional domain structure in nucleic acids as evidenced by cross-linking entropy: part 2. Journal Theoretical Biology 213: 387-412.

14. Day R, Daggett V (2003) All-atom simulations of protein folding and unfolding. Advances in Protein Chemistry 66: 373-403.

15. de Gennes PG (1979) Scaling Concepts in Polymer Physics. Ithaca, Cornell University Press.

16. Dill KA, Stigter D (1995) Modeling protein stability as heteropolymer collapse. Advances in Protein Chemistry 46: 59-104.

17. Ding F, Tsao D, Nie H, Dokholyan NV (2008) Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure 16: 1010-8.

18. Feller W (1968) An Introduction to Probability Theory and its Applications (pt I). New York Wiley.

19. Feller W (1971) An Introduction to Probability: Theory and Its Applications (pt II). New York John Wiley & Sons.

20. Fisher ME (1966) Effect of excluded volume on phase transitions in biopolymers. J Chem Phys 45: 1469-1473.

21. Flory PJ (1953) Principles of Polymer Chemistry. Ithaca, Cornell University Press.

22. Flory PJ (1969) Statistical Mechanics of Chain Molecules. New York Wiley (Regrettably out of print.)

23. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, et al. (1986) Improved free-energy parameters for predictions of RNA duplex stability. Proc Natl Acad Sci USA 83: 9373-7.

24. Friederich MW, Vacano E, Hagerman PJ (1998) Global flexibility of tertiary structure in RNA: yeast tRNA(phe) as a model system. Proceedings of the National Academy of Science (USA) 95: 3572-77.

25. Go N (1999) The consistency principle revisited. In: Old and new views of protein folding. Amsterdam. Elsevier Science pp249-257.

26. Grosberg AY, Khokhlov AR (1994) Statistical Physics of Macromolecules. New York AIP Press.

27. Hagerman PJ (1997) Flexibility of RNA. Annual Review Biophysics and Biomolecular Structucture 26: 139-56.

28. He YN, Chen YH, Alexander P, Bryan PN, Orban J (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. Proc Natl Acad Sci USA 105: 14412-14417.

29. Hnizdo V, Tan J, Killian BJ, Gilson MK (2008) Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. J Comput Chem 29: 1605-14.

30. Honig B, Ray A, Levinthal C (1976) Conformational flexibility and protein folding: rigid structural fragments connected by flexible joints in subtilisin BPN. Proc Natl Acad Sci USA 73: 1974-8.

31. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, et al. (2003) Contact order revisited: influence of protein size on the folding rate. Protein Science 12: 2057-62.

32. Jacobson H, Stockmayer W (1950) Intramolecular reaction in polycondensations. I. the theory of linear systems. Journal of Chemical Physics 18: 1600-1606.

33. Kolinski A, Gront D, Pokarowski P, Skolnick J (2003) A simple lattice model that exhibits a protein-like cooperative all-or-none folding transition. Biopolymers 69: 399-405.

34. Kuhn W (1934) Uber die Gestalt fadenformiger Molekule in Losungen (on the shape of filiform molecules in solution]. Kolloidzeitschrift 68: 2. from citation in I. M ller. 2007 ISBN: 978-3-540-46226-2).

35. Kuhn W (1936) Beziehungen zwischen Molekulgroshe, statistischer Molekulgestalt und elastischen Eigenschaften hochpolymerer Stoffe [Relations between molecular size, statistical molecular shape and elastic properties of high polymers. Kolloidzeitschrift 76: 258. (from citation in I. M ller. 2007 ISBN: 978-3-540-46226-2).

36. Lebedev NN (1965) Special Functions & their Applications. Englewood Cliffs (NJ), Prentice-Hall. Dover reprint.

37. Lehninger AL (1975) Biochemistry. Ed. 2. New York, Worth Publishers, INC. (Out of print: newer editions *may* contain the same subject material. Newer books such as Stryer L Biochemistry, Freeman also contain very similar material on this related topic but not in the same exposition.)

38. Lesk AM (2001) Protein Architecture. Oxford, Oxford University Press.

39. Liu SM, Haynes CA (2005) Energy landscapes for adsorption of a protein-like HP chain as a function of native-state stability. J Colloid Interface Sci 284: 7-13.

40. Ma SK (1973) Introduction to Renormalization Group. Reviews of Modern Physics 45: 589-614.

41. Marko JF, Siggia ED (1995) Stretching DNA. Macromolecules 28: 8759-8770.

42. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology 288: 911-940.

43. McKenzie DS (1976) Polymers and scaling. Physics Reports 27C: 35-88.

44. Mintseris J, Weng Z (2003) Atomic contact vectors in protein-protein recognition. Proteins 53: 629-39.

45. Mirny LA, Abkevich VI, Shakhnovich EI (1998) How evolution makes proteins fold quickly. Proc Natl Acad Sci USA 95: 4976-81.

46. Montroll EW (1950) Markoff chain and excluded volume effect in polymer chains. J Chem Phys 18: 734-743.

47. Murray LJ, Arendall WB 3rd, Richardson DC, Richardson JS (2003) RNA backbone is rotameric. Proc Natl Acad Sci USA 100: 13904-9.

48. Nash LK (1974) Elements of statistical Thermodynamics. Reading, Addison-Wesley. (Kindly reissued recently by Dover Books and well worth the investment.)

49. Onuchic JN, Nymeyer H, Garcia AE, Chahine J, Socci ND (2000) The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. Advances in Protein Chemistry 53: 87-152.

50. Pappu RV, Rose GD (2002) A simple model for polyproline II structure in unfolded starts of alanine-based peptides. Protein Science 11: 2437-2455.

51. Pappu RV, Srinivasan R, Rose GD (2000) The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. Proceedings of the National Academy of Science USA 97: 12565-70.

52. Pokarowski P, Kolinski A, Skolnick J (2003) A minimal physically realistic protein-like lattice model: designing an energy landscape that ensures all-or-none folding to a unique native state. Biophys J 84: 1518-26.

53. Richardson JS (1977) beta-Sheet topology and the relatedness of proteins. Nature 268: 495-500.

54. Richardson JS (1981) The anatomy and taxonomy of protein structure. Advances in Protein Chemistry 34: 167-339.

55. Rouzina I, Bloomfield VA (2001a) Force-induced melting of the DNA double helix 1. Thermodynamic analysis. Biophys J 80: 882-93.

56. Rouzina I, Bloomfield VA (2001b) Force-induced melting of the DNA double helix. 2. Effect of solution conditions. Biophys J 80: 894-900.

57. Swendsen RH (2006) Statistical mechanics of colloids and Boltzmann's definition of the entropy. Am J Phys 74: 187-190.

58. Swendsen RH (2008) Gibbs' Paradox and the Definition of Entropy. Entropy 10**:**15-18.

59. Sykes MF (1963) Self-avoiding walks on the simple cubic lattice. J Chem Phys 39: 410-412.

60. Takasu A, Watanabe K, Kawai G (2002) Analysis of relative positions of ribonucleotide bases in a crystal structure of ribosome. Nucleosides Nucleotides Nucleic Acids 21: 449-62.

61. Taylor WJ (1948) Average length and radius of normal

paraffin hydrocarbon molecules. J Chem Phys 16: 257-267.

62. Tinoco I, Bustamante C (1999) How RNA folds. Journal of Molecular Biology 293: 271-81.

63. Wall FT, Erpenbeck JJ (1959) New method for the statistical computation of polymer dimensions. J Chem Phys 30: 634-637.

64. Wall FT, Hiller LA, Atchison WF (1955) Statistical computation of mean dimensions of macromolecules. II J Chem Phys 23: 913-921.

65. White RP, Funt J, Meirovitch H (2005) Calculation of the Entropy of Lattice Polymer Models from Monte Carlo Trajectories. Chem Phys Lett 410: 430-435.

66. Zhang C, Vasmatzis G, Cornette JL, DeLisi C (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 267: 707-26.