J. theor. Biol. (2001) **213,** 359–386
doi:10.1006/jtbi.2001.2436, available online at http://www.idealibrary.com on IDEAL®

A Physical Origin for Functional Domain Structure in Nucleic Acids as Evidenced by Cross-linking Entropy: I

WAYNE DAWSON*†‡, KAZUO SUZUKI* AND KENJI YAMAMOTO†

**Department of Bioactive Molecules, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjuku-ku, Tokyo* 162-8640, *Japan and* †*Department of Medical Ecology, Tokyo International Medical Center Japan*, 1-21-1 *Toyama, Shinjuku-ku*, 162-8640, *Japan*

(*Received on* 14 *July* 2000, *Accepted in revised form on* 3 *August* 2001)

A global strategy for estimating the entropy of long sequences of RNA is proposed to help improve the predictive capacity of RNA secondary structure dynamic programming algorithm (DPA) free energy (FE) minimization methods. These DPA strategies only consider the effects that occur in the immediate (nearest neighbor) vicinity of a given base pair (bp) in a secondary structure plot. They are therefore defined as nearest-neighbor secondary structure (NNSS) strategies. The new approach utilizes the statistical properties of the Gaussian polymer chain model to introduce both local and global contributions to the entropy of a given secondary structure. These entropic contributions are primarily a function of the persistence length of the RNA. Limits on the domain size are strongly suggested by this model and these limits are a function of both the length and the percentage of bp enclosed within a given domain. The model generalizes the penalties found in the NNSS algorithms. The approach considers the importance of flexibility in the folding and stability of RNA by considering the role of the persistence length in a biopolymer structure. The theory also suggests that molecular machinery may also take advantage of this global entropic effect to bring about catalytic effects. The applications can also be extended to protein structure calculations with some additional considerations.

1. Introduction

For long RNA sequences (more than 1000 nucleotides (nt)), the dynamic programming algorithm (DPA) designed to fit RNA secondary structure using free energy (FE) minimization strategies (Nussinov & Jacobson, 1980) often fails to reflect the experimentally known folding and thermodynamic distribution of RNA secondary structure. The DPA strategy relies on the assumption that when a base pair (BP) is formed, the enthalpic and entropic contributions are restricted to the structural context in the *immediate* vicinity of the BP and when a secondary structure feature such as a loop is encountered, the free segment region (or regions) are evaluated via an entropy look up table that is independent of *where* in that given secondary structure in which such a penalty is applied (McCaskill, 1990; Zuker & Stiegler, 1981). In effect, the strategy considers only the nearest neighbor in the context of the evaluations, and cannot "see" where the nearest neighbor is in the context of the entire secondary structure. We therefore call this approach a nearest-neighbor secondary structure (NNSS) strategy.

‡Author to whom correspondence should be addressed. E-mail: dawson@nih.go.jp

0022–5193/01/230359 + 28 $35.00/0

Significant discrepancies in NNSS structure prediction often begin to emerge when a predicted domain size (defined in Section 2.1) exceeds 100 nt in length. The problem appears to be a function of both the size and the number of BPs that encompass the domain.

Naturally, because RNA folds from the 5′ end toward the 3′ end, some of this discrepancy is probably due to the non-equilibrium conditions under which the sequence is formed (Brion & Westhof, 1997; Tinoco & Bustamante, 1999; Nussinov *et al*., 1982; Mironov *et al*., 1985).

Notwithstanding, predictions of short sequences of RNA yield reasonable results in most cases and appeals to non-equilibrium conditions are rarely sought. Likewise, when the secondary structure has highly conserved regions, short stems, and only yields small hierarchies of secondary structure, the NNSS predictions also come out fairly well. Again, equilibrium thermodynamics is sufficient for the job. Yet, as the sequences become significantly longer than 100 nt, there is an increasingly stronger divergence between the experimentally and theoretically predicted "best guess". Still, if the sequence is windowed in the "right way", one can sometimes coerce a reasonably good secondary structure prediction out of the NNSS strategies. Likewise, if there is existing experimental evidence for the *bona fide* structure (NMR spectroscopy, X-ray crystallography, RNase digestion, comparative sequence alignment, etc.), a reasonable approximation of that structure can usually be ferreted out of the long list of suboptimal structures (often on the order of hundreds). Hence, some aspects of the free energy estimation using equilibrium thermodynamics are clearly correct when the example set, the suboptimal structures, or the proper windowing can be discovered.

However, windowing does not explain *why* there are limits to the complexity or the length of a functional domain, what those limits could be, or how those limits might influence the formation of RNA functional domains. Neither does it explain why so much of the genetic repertoire is so skillfully self-assembled under extreme (and consequently *inefficient*) non-equilibrium conditions. Finally, if our desire is to predict an unknown structure of RNA, searching through a long list of suboptimal structures is of little use.

It is our proposal here that the major problem is not equilibrium thermodynamics. Rather, it is the model that is used for estimating the entropy. As large complex hierarchal secondary structures form, extensive order is introduced. NNSS algorithms do not account for this global effect because (as their name implies), they only take into account effects that occur in the immediate (nearest neighbor) vicinity of a given BP in a secondary structure plot.

This new strategy relies heavily on terminology originally developed in polymer science (Doi & Edwards, 1986; Grosberg & Khokhlov, 1994; Flory, 1953; James & Guth, 1947). However, we have applied these concepts from a different perspective than what is brought out in the literature, including works devoted to the biosciences. This work bears some resemblance to earlier work in loop weighting functions (Scheffler *et al*., 1970) and also parallels a recently developed methodology in protein structure calculations known as the Gaussian polymer network (GPN) model (Lustig *et al*., 1998; Keskin *et al*., 2000; Debe & Goddard, 1999). In addition, Frederic *et al*., 1996) have also developed a related approach on 3-D structures for the ends of the sequence. The term "network" although mathematically correct seems a bit obtuse for our needs. Flory referred to the effect we report here as a kind of "cross-linking" (Flory, 1953, 1956, 1976). However, the term "cross-linking" is usually associated with covalent bonding. Since the bonds that form in RNA folding are of a Van der Waals nature (stacking), this also creates some confusion. Another possibility is "hairpin folding entropy" (E. Westhof, private comm.). This is a powerful visualization of the concept, but any appeal to the structure forces us to speak of individual BPs or "cross-links". Therefore, although we find problems with all terminologies so far rendered, for the time being, we will use the terms "cross-link" (cl) and cross-linking entropy (CLE). As we will show, the CLE is general enough to describe the folding of any polymer into a secondary structure irrespective of the type of chemical bond.

This work is written in two parts. In Part I, we develop the strategy for estimating the global entropy caused by folding a sequence of RNA into its corresponding secondary structure, and

in Part II, we develop and apply these concepts to RNA secondary structure prediction. In the process, we will show that double strand RNA (dsRNA) and single strand RNA (ssRNA) are different particularly if the enclosed RNA sequence is quite long. Overall, this work explores the role and extent that equilibrium thermodynamics plays in the stability and formation of functional domains of RNA secondary structure. We also show that this strategy offers a surprisingly powerful means for predicting the behavior of secondary structure in long RNA sequences.

2. Theory and Calculation Methods

2.1 SECONDARY STRUCTURE DEFINITIONS

General definitions of secondary structure can be found in the literature (Zuker, 1998), and some clear visual descriptions can be found in (Comay *et al*., 1984). These definitions are restricted to *linear* single stranded sequences of RNA (ssRNA) and DNA (ssDNA).

Let the index i identify the position of a nucleotide in a nucleic acid sequence and let the ordered pair (i, j) express a bp (BP) in that sequence (with $i < j$). Then secondary structure is expressed by the following rules:§

- $\{(i', j') | (i' = i \,\&\&\, j' = j) \,\|\, (i' \neq i \,\&\&\, j' \neq j)\}$.
- $\forall (i', j') \in (i, j) \Rightarrow i < i' < j' < j$.
- $\forall (i', j') \notin (i, j) \Rightarrow \begin{cases} i' < i \Rightarrow j' < i, \\ i' > j \Rightarrow j' > j. \end{cases}$

These rules result in a hierarchal ordering of the structures where no pseudoknots or triple helices are allowed. Secondary structure is usually divided into stems ($\mathscr{S}$), loops ($\mathscr{H}$), bulges ($\mathscr{B}$), internal loops ($\mathscr{I}$), branch points ($\mathscr{V}$), and multibranch loops (MBL) (Lyngsø, 1999; Studnicka *et al*., 1978; Zuker *et al*., 1998).

To develop and describe the issues discussed in this work, some additional definitions are required.

§The mathematical symbols used here are defined as follows: $|$ ≡ "such that"; $\|$ ≡ "OR"; && ≡ "AND"; $\forall$ ≡ "for any"; $\Rightarrow$ ≡ "implies"; and $(i', j') \in (i, j)$ indicates the set of ordered pairs (i', j') which are contained in the secondary structure bounded by (i, j) (where $i' < j'$ is implied).

Definition 1 (*Domain boundary*)**.** For a specified secondary structure (**S**) containing the ordered pairs $\{(i', j')\}$, suppose (i, j) denotes a domain boundary of secondary structure. Then for any ordered pair $(i', j') \in \mathbf{S}$ with $i' \neq i$ and $j' \neq j$, the following conditions are satisfied:

$$\forall (i', j') \notin (i, j) \Rightarrow \begin{cases} i' < i \Rightarrow j' < i, \\ i' > j \Rightarrow j' > j, \end{cases}$$

$$\forall (i', j') \in (i, j) \Rightarrow j - i > max\{j' - i'\}.$$

In short, no ordered pair $(i', j') \in \mathbf{S}$ for which $i' < i$ or $i' > j$ can simultaneously satisfy $(i', j') \in (i, j)$ if (i, j) denotes a domain boundary of secondary structure **S**. Further, the BP (i, j) and (k, l) (with $i \,\&\&\, j < k \,\&\&\, l$) form two separate domains if and only if both (i, j) and (k, l) satisfy the definition of a domain boundary individually.

Definition 2 (*Domain*). If (i, j) defines a domain boundary, then a domain consists of all (i', j') that satisfy

$$\{(i', j') | (i', j') \in (i, j)\}.$$

Figure 1 shows a schematic example of some secondary structure in which the domain boundaries are indicated by ∂(domain k) where k (in this example) is an integer between 1 and 4. The domains represent the secondary structure enclosed by the domain boundaries. Hence, a domain represents the collection of $\mathscr{S}$, $\mathscr{B}$, $\mathscr{I}$, $\mathscr{H}$, and $\mathscr{V}$ structures that form a hierarchy of secondary structure above the domain boundary (i, j).

Domains defined by way of Definition 1 may differ from the common biological descriptions of domains that generally consist of large secondary structural features. In the definition given here, we must assign any structure that satisfies Definition 1 the title "domain" irrespective of its size. To distinguish between these definitions, the biological (or more common) designation for "domain" will be numbered using Roman numerals, and domains according to Definition 1 will be numbered using Arabic numerals.

We also must distinguish between different types of multibranch loops (MBLs). These MBLs

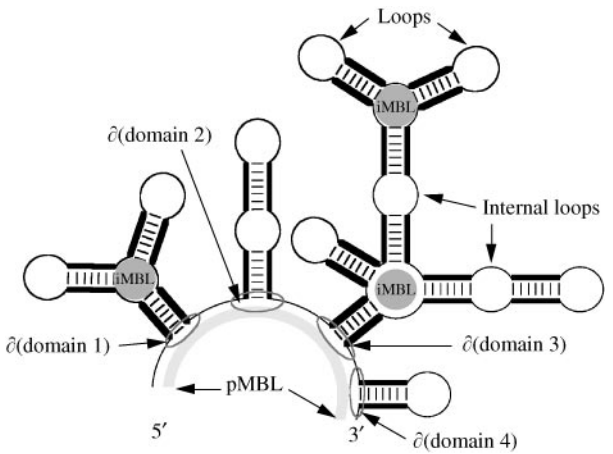

FIG. 1. A schematic example of secondary structure showing the principle multibranch loop (pMBL), internal multibranch loops (iMBL), and domain boundaries [∂(domain k): the semi-circle at the base of the stems along the pMBL, where in this case $k = 1 \cdots 4$]. A domain consists of the secondary structure contained within the regions enclosed by the domain boundaries (∂(domain k)). Subdomains would include such points as the closing BP of the iMBLs, internal loops, etc.

have also been called a "bifurcation loop" in the literature (Zuker & Stiegler, 1981).

Definition 3 (*Internal multibranch loop* (*iMBL*))**.** An iMBL consists of a sequence fragment which exhibits multiple branching points but is closed at its 5′ and 3′ ends by a stem consisting of at least one BP. Figure 1 shows an iMBL in three different locations.

Definition 4 (*Subdomain*)**.** Any subsequence that forms secondary structure and that is contained within a specified domain.

In this work, we will restrict our discussion of a subdomain (k', l') to secondary structure that is bounded at its 5′ and 3′ ends by either a stem or an iMBL.

Definition 5 (*Principal multibranch loop* (*pMBL*))**.** A pMBL runs from the 5′ end to the 3′ end of the sequence and represents the effective length of sequence that results from connecting the gaps formed at the domain boundaries with any free segments that join those domains.

The pMBL in Fig. 1 is represented by the light gray semi-circle and consist of the unhybridized bases from the 5′ to the 3′ end and the gaps formed by the circled regions at ∂(domain k) There are no pMBLs in a circular RNA or DNA sequence and there can only be one pMBL in a given linear sequence. The pMBL resembles the standard definitions for MBLs in (Zuker *et al.*, 1998); however, because the 5′ and 3′ ends are not necessarily closed, this structure is different from an iMBL. It has also been called an "open structure" (Williams & Tinoco, 1986).

Definition 6 (*MBL hierarchal complexity* (*HC*))**.** In a tree diagram of secondary structure, the highest node (indexed with respect to the MBLs) that forms branching points ($\curlyvee$) in a given domain: for a pMBL, HC $\equiv 0$ and for all iMBLs, HC $\geqslant 1$.

The HC index corresponds to the level or recursion required to search a domain of secondary structure. In Fig. 1, domains 1–4 have the following HC: 1, 0, 2, and 0, respectively.

Definition 7 (*cross-link* (*cl*))**.** Any kind of chemical bond that joins two monomers together whether they be within the same polymer chain or between two different polymer chains.

In this work, we are mostly concerned with *intra*-chain cross-links or BPs. *Inter*-chain cross-linking is of interest with respect to dsRNA/dsDNA and RNA/DNA binding proteins.

Definition 8 (*base pair density* (*BPD*))**.** In a sequence of length N, the ratio between the number of observed BPs (cross-links) and $N/2$, which is the maximum conceivable number of BPs allowed for a sequence of length N that only forms secondary structure.

In principle, only sequences like $G_{N/2}C_{N/2}$ ($N/2$ large) can achieve a BPD approaching 1.

2.2. REAL POLYMERS AND THE GAUSSIAN POLYMER CHAIN

2.2.1. *Length Scales in the GPC: the Persistence Length*

A polymer consists of individual segments of monomers or "mers". These monomers have an

average separation b between successive segments and the polymer consists of N such "mers" in total. A sequence of RNA contains different monomers: adenine (A), cytosine (C), guanine (G), and uracil (U). Although these represent different monomers, the structure and properties are similar enough such that, in first approximation, the details of their individual properties can be neglected.

In the Gaussian polymer chain (GPC) approximation of a real polymer, the chemical concept of a monomer shifts to a description in terms of the number of "links" and the distance between each "link": the persistence length (Grosberg & Khokhlov, 1997). Persistence length expresses the rigidness of a polymer chain. For example, a steel pipe will "persist" over a much longer distance than a segment of rubber hose of equivalent size and shape (Hagerman, 1997).

The persistence length reflects the distance between "links" on the GPC and does not necessarily equal the distance between actual chemical monomers (or "mers") on the chain.

Figure 2 shows a polymer in which the monomers are expressed by a sawtooth-like pattern where each "mer" is located at the tip of the sawtooth. Chemical bonds join some of these monomers together (the thin lines) folding the monomer into a hairpin. The GPC approximation of this chemical model is shown by the large gray bars capped with the hemispheric balls at the respective centers of the individual bars. Figure 2 also demonstrates a case where the persistence length (or the distance between "links") is longer than the distance between the individual chemical "mers" in the polymer, the persistence length is about 5.5 "mers" per "link".

In Fig. 3, a very different polymer is shown in which the chemical "mers" (the large circles) are separated by very flexible connecting segments. Chemical bonds also join some of the monomers together folding the sequence into a hairpin (the thick black bars joining the circles). To make such a highly flexible segment as seen in Fig. 3, there must be a very large number of "links" joining these "mers". Hence, in Fig. 3, the persistence length is very short and the number of "links" is very large.

For a polymer chain of length (L), the relationship between the number of "mers" (N) and the number of "links" ($\tilde{N}$) is

$$L = Nb = \tilde{N}\tilde{b}, \tag{1}$$

where b is the distance between individual "mers" in a real polymer, and $\tilde{b}$ is the persistence length or the distance between the "links" in the GPC approximation of the polymer chain. We define the parameter ξ as the persistence ratio (Grosberg & Khokhlov, 1994) such that

$$\tilde{N} = N/\xi \quad \text{and} \quad \tilde{b} = \xi b. \tag{2}$$

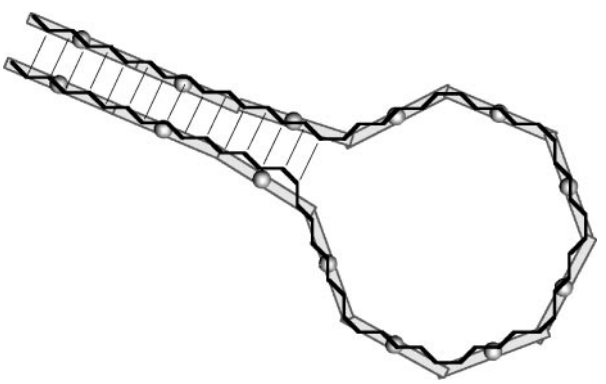

FIG. 2. An example of a polymer where the persistence length is longer than the distance between the individual monomers. The "mers" are located at the joints of the sawtooth-like pattern in the figure. The "links" of the GPC approximation consist of the gray bars with the hemisphere at their respective centers. The sequence is folded into a hairpin in which the chemical bonds between the monomers are indicated by the thin black lines. The final structure appears quite inflexible to further folding.

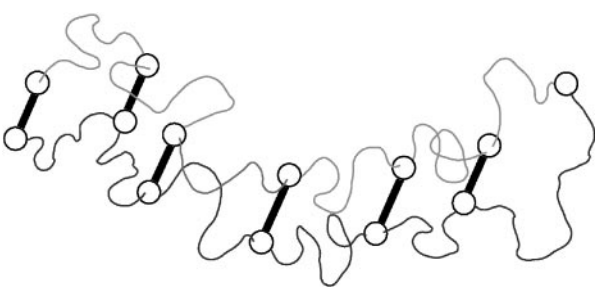

FIG. 3. An example of a polymer where the persistence length is shorter than the distance between the individual monomers. The "mers" are indicated by the large white circles and the connections between the "mers" are indicated by the thin gray flexible lines that connect the monomers together. This sequence is also folded into a hairpin where the chemical bonds are indicated by the thick black bars. Although this polymer is folded into a hairpin, the structure appears quite flexible to further folding compared to Fig. 2.

Figure 2 expresses the case where $\tilde{b} > b$ ($\xi > 1$) and Fig. 3 shows the case where $\tilde{b} \ll b$ ($\xi \ll 1$).

In the greater literature of statistical mechanics, the persistence ratio is analogous to such parameters as the correlation length in the Ising model and the coherence length in superconductivity (Plischke & Bergersen, 1994). Hence, ξ measures the distance between different units in a given system where the interaction between such units can be treated as though the units behave independently. In polymer science, these units are the "links".

Two important points must be stressed here. First, it can be seen in Fig. 2 that the bonding of the polymer into a hairpin makes that structure rigid and very little further folding is conceivable in this structure. On the other hand, the structure in Fig. 3 is still very flexible and can be folded with only a small amount of resistance. Hence, there is an intuitive higher level of persistence length that can be envisioned in these folded structures. The structure in Fig. 2 is extremely rigid and has a persistence length that is roughly a function of L, whereas the structure in Fig. 3 remains flexible with a persistence length that is roughly a function of b (the monomer separation distance).

Second, the value of ξ need not be assumed to be a constant (Hagerman, 1997). The *context* of the structure is likely to influence the value of ξ: the type of monomer, the type and arrangement of the chemical bonds in the structure, the temperature, etc. Figure 4 shows a case of a mixed polymer where the loop of the hairpin has a shorter persistence length than the stem (cf., Fig. 2). The persistence ratio in the coil state (ξ_c) and the folded state (ξ_f) are also likely to be different due o the presence of helical stacking (AU, GC, etc.), Mg^{2+} receptor binding, electrostatic effects and hydrogen bonding. However, to evaluate the model in Parts I and II, we will assume that ξ is constant under all experimental conditions. Presumably, there are at least some structures of RNA where this approximation is reasonable.

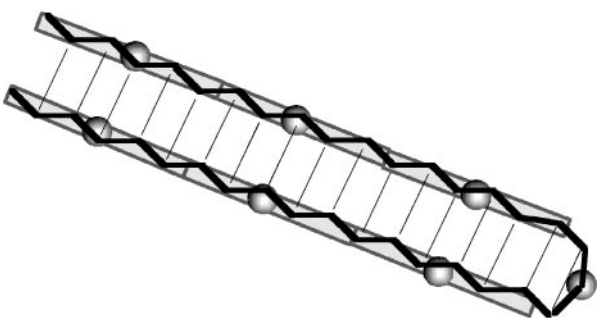

FIG. 4. An example of a polymer which has a mixed persistence length that combines the features of Figs 2 and 3. The stem of the polymer exhibits the same persistence length as in Fig. 2. However, at the point of the hairpin, the sequence is tightly folded such that only one "mer" occupies the hairpin resembling the structure in Fig. 3. Hence, at the base of the hairpin, the molecule shows far more flexibility than in the stem region.

2.2.2. *Statistical Root Mean Separation Distance of a GPC: the Coil State*

For any polymer chain composed of a single monomer, the observed elastic effect is almost entirely due to entropy. Changes in the entropy of a polymer chain are caused by a loss of conformational degrees of freedom. A polymer chain stretched out to its full extent has only one configuration available to it, whereas a free chain can assume many configurations (Grosberg & Khokhlov, 1997). Hence, the folded chain has a higher entropy than the extended chain. Similarly, if the chain is compressed, less configurations are available to the chain and again, the entropy must decrease. Since entropy always tends toward a maximum, the randomly folded configuration of the polymer chain will be the preferred state in thermodynamic equilibrium unless external forces are applied to the structure. This is the origin of the elastic effect in a polymer chain and this equilibrium state is known as the "coil state" which we will denote by "c".

From polymer science theory (Appendices B and C), the equilibrium separation between the ends of an ideal polymer chain is

$$r_c = R = \tilde{b}\left(\frac{2\gamma\tilde{N}}{3}\right)^{1/2}, \tag{3}$$

where γ is an excluded volume correction for the GPC (Appendix C), $\tilde{b}$ is the persistence length and $\tilde{N}$ is the number of "links" in the GPC (see Section 2.2.1). The value of R represents the maximum entropy for an ideal polymer chain where there are no chemical bonding interactions between "mers" except for the bonds that link the polymer chain together.

By the central limit theorem (Feller, 1971) that yields eqn (3), it follows that *for any* two bases ($\tilde{i}$ and $\tilde{j}$) at any position on the GPC, the root mean equilibrium separation distance ($R_{\tilde{i},\tilde{j}}$) will be

$$R_{\tilde{i},\tilde{j}} = \tilde{b}\left(\frac{2\gamma\tilde{N}_{\tilde{i},\tilde{j}}}{3}\right)^{1/2} \quad \text{(coil state)}, \qquad (4)$$

where $\tilde{N}_{\tilde{i},\tilde{j}} = (\tilde{j} - \tilde{i} + \tilde{1})$ (Grosberg & Khokhlov, 1997) and the tilde notation (although cumbersome), is meant to emphasize that we are indexing the "links" and not the "mers". Hence, eqn (3) applies both to the chain as a whole (where $\tilde{i} = \tilde{1}$ and $\tilde{j} = \tilde{N}$) and to a particular segment within the chain (from $\tilde{i}$ and $\tilde{j}$).

2.2.3. *Bound Nucleic Acid Chains and the Stacking Gap: the Folded State*

In a real nucleic acid, there are stacking effects¶ due to AU, GC, and GU binding, as well as some other non-Watson Crick pairs (Burkard *et al*., 1999). Stacking causes the separation between nucleotides to change from the equilibrium value of R to a value r_f which represents the distance between the BPs that form in a folded RNA or DNA sequence. The distance between the BPs (AU, GC, and GU) is in the same order as the bond lengths between adjacent nucleic acids on the chain. Therefore, for convenience, we define r_f (the "folded state") in terms of the monomer bond length

$$r_f = \lambda b \quad \text{(stacking gap)}, \qquad (5)$$

where λ is the *stacking parameter* (Section 3.1.1).

2.3. THE ENTROPY OF FOLDING FOR A GAUSSIAN POLYMER CHAIN

A brief description of the equilibrium thermodynamics of reversible reactions in polymer chains and a derivation of the equations related to the Gaussian polymer chain (GPC) model can be found in Appendices A–D. In this section, we develop the conceptual framework of the GPC model and the method for calculating the CLE for a domain in terms of a given secondary structure.

2.3.1. *A Conceptual Model of Cross-linking*

As a hypothetical experiment, let us imagine that the monomer pairing interactions can be turned off at our command (with all other parameters such as $\tilde{b}$, $\tilde{N}$, etc. held constant).

Suppose we start with a polymer of $\tilde{N}$ links where all the bonds are turned off [no cross-linking present: Fig. 5(a)]. Such a sequence would resemble a neutral polymer and tend to behave like a GPC with an equilibrium separation distance R_{AB} [eqn (3)]. Moreover, at any part of the sequence, the equilibrium separation distance between "link" $\tilde{i}$ and "link" $\tilde{j}$ will tend to be $R_{\tilde{i},\tilde{j}}$ [eqn (4)].

Now suppose that we turn on a "cross-link" at position AB and allow the system to achieve thermodynamic equilibrium [Fig. 5(b)]. A single "cross-link" closes part of the sequence forming a loop and forcing a change of state: $R_{AB} \rightarrow r_{AB}$ [Fig. 5(b)]. A "price" must be paid for this (call it ΔS_{AB}).

Suppose further that we turn on a second "cross-link" at position CD and again allow the system to equilibrate [Fig. 5(c)]. Position AB (and r_{AB}) has no direct influence on position CD (or r_{CD}) (Poland & Scheraga, 1966). Note that we have added *specific* correlations between C and D that were not present before and have introduced greater "order" to the structure as a consequence. Therefore, we must pay another "price" to form this bond: ΔS_{CD}. The cross-linking entropy contribution will now be the sum of the entropy closing the loop at AB and the entropy closing the loop at CD.

It follows that if we turn on yet another "cross-link" at position EF [Fig. 5(c)], we must also "pay" for that with a cost ΔS_{EF}. The total "cost" required to maintain the GPC in the configuration shown in Fig. 5(d) will be $\Delta S_{net} = \Delta S_{AB} + \Delta S_{CD} + \Delta S_{EF}$ (neglecting minor inter-link correlation effects which will be examined in Part II).

At this point, some comment is needed on the additive properties of the entropy.

¶ *Note*: it is not actually the hydrogen bonding per se, but the *stacking interaction* (involving a group of consecutive H-bonded BPs) that is the major contribution to the stability and formation of BPs in nucleic acid sequences (Searle & Williams, 1993).

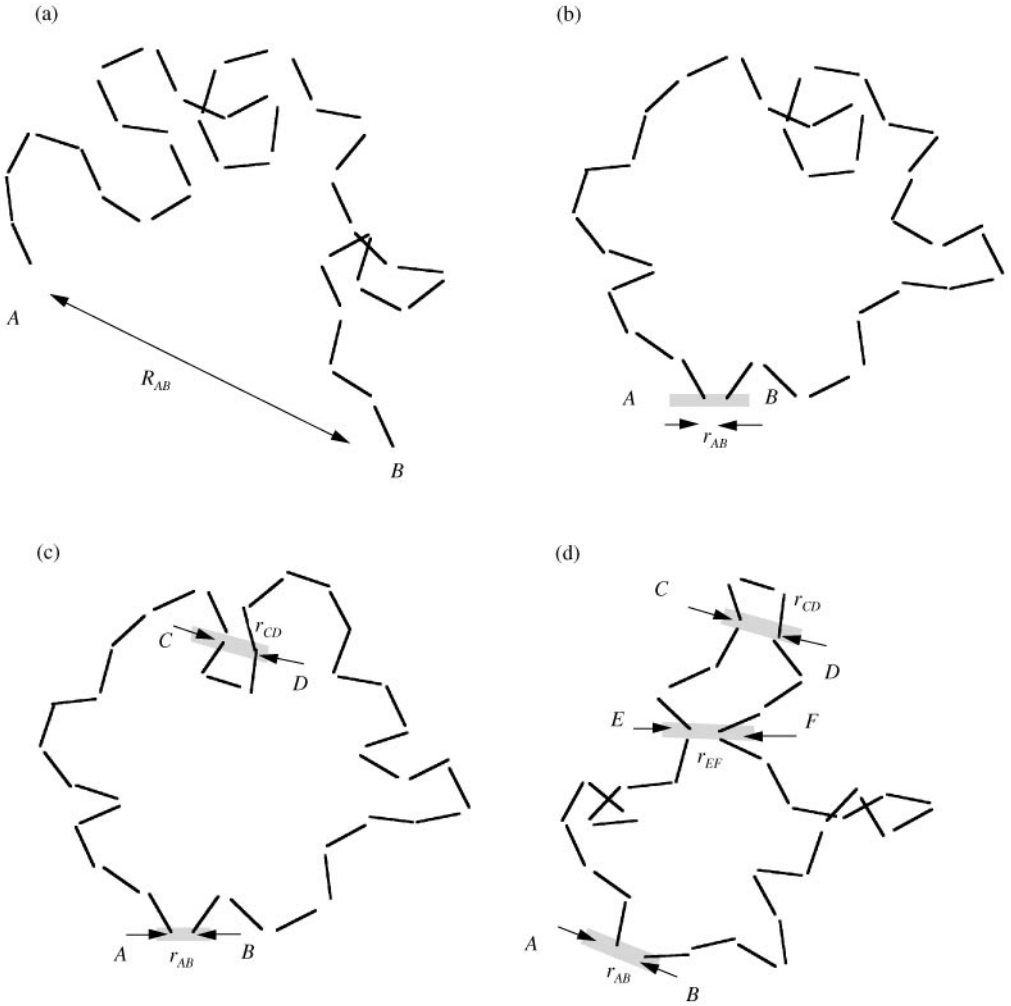

FIG. 5. The origin of the cross-linking entropy. (a) A model polymer chain of length $\tilde{N}\tilde{b}$ (see text) and end-to-end separation (AB) is shown with an equilibrium (end-to-end) separation distance R_{AB}. The arrow is drawn to emphasize that the measurement is the *displacement* between points AB. (b) The polymer chain is now bound at position AB on the chain (via a chemical bond, for example). Upon binding the polymer chain at AB, the chain loses $\tilde{N}_{AB}$ degrees of freedom. Therefore (using (47)—Appendix B), a force $f(r_{AB})$ must be applied at AB to maintain the structure in a closed loop. (c) A second position on the polymer chain is now bound at position CD. The chain loses an additional $\tilde{N}_{CD}$ degrees of freedom, and a force $f(r_{CD})$ must be applied at CD to maintain the structure in a closed loop. (d) A third bond is formed at position EF. The chain loses yet an additional $\tilde{N}_{EF}$ degrees of freedom, and a force $f(r_{EF})$ must be applied to keep the loop closed at EF. An integration of these forces (using eqn (A.3)—Appendix A) will yield the entropy. (A total of $\tilde{N}!$ degrees of freedom exist in an $\tilde{N}$ independent particle system.)

At one extreme is the "floppy" chain (Fig. 3) where many "links" comprise a single "mer". In other words, the distance between the chemical bonds (formed by the "mers") is far compared with the distance between the "links". Since correlation effects vanish exponentially as a function of the persistence length, and these correlation effects are a function of $1/\tilde{N}$ (Grosberg & Khokhlov, 1994) (note the "link" notation), it should be clear that the influence between nearest neighboring "mers" is negligible.

At the other extreme is the "stiff" chain (Fig. 2) where the "links" are composed of many "mers". Again, the functional unit is the "link" which is equal to the persistence length. The "mers" respond as a single unit (forming a particular "link") and the entropy must be evaluated in terms of the "links" rather than the "mers". Therefore, at the level of the nearest-neighboring chemical monomers on the chain the correlations are *already* accounted for collectively via the persistence length. Since the links are long and

"stiff", the influence of chemical bonding on such large structures is again small, and we can treat the nearest-neighboring "links" as effectively independent in first approximation.

Moreover, in considering the influence of a cross-link at (E, F) on the formation of the cross-link at (C, D), one should not ignore the rest of the chain which will try to maximize its entropy in the context of the distortions introduced by the cross-links. The transition between Fig. 5(c) and (d) emphasizes the fact that a multitude of configurations are possible for a GPC: none of which need to have any influence on the formation of a neighboring state. It is noted that the binding of a GPC is not restricted by the natural physical constraints usually found in real materials. However, extreme contortions are highly improbable and do not represent any significant fraction of the macrostate. Finally, this is an *approximation* which is likely to have limits due to the sub-additivity of entropy theorem (E. Lieb, private comm.). Nevertheless, nearest-neighbor secondary structure algorithms already rely on this *assumed* additivity with *no* corrections for correlation. The success of such methods already indicates that the correlation effects are second-order corrections in nucleic acid systems. Indeed, we will even examine this matter in Part II of this work and show that inter-link correlation effects only introduces modest changes in the results.

If the structure in Fig. 5(d) also represents the most thermodynamically probable state of the system, then when all the bonds are turned off the sequence will unfold to its equilibrium separation distance $R_{i,j}$. Likewise, if all the bonds are turned on again (at the same time) and the system is allowed to equilibrate, then the polymer will fold back to its original native structure [Fig. 5(d)]: assuming the process is entirely reversible. An approximation of this effect in a real polymer can be achieved by denaturing and renaturing nucleic acid sequences (Brion *et al.*, 1997; Pan & Woodson, 1999). However, a variety of Van der Waals interactions still remain even in highly denaturing conditions which result in deviations from this idealized model. The current description can be considered the theoretical state of a nucleic acid in an "ideal denaturing solvent".

The diagrams in Fig. 5 represent a conceptual construct to help in understanding the cross-linking entropy effect and should not be used to understand the entropy of the reaction intermediates. The hypothetical experiment resembles more of a process akin to Maxwell's daemon because we have manipulated which bonds are turned on and when. To properly express the reaction intermediates, we need detailed configuration information about the particular paths in Fig. 5. In thermodynamic equilibrium, the entropy of a reversible reaction is path independent; hence we only need to know the initial and final states of the system and then sum up the individual contributions with some consideration taken for the degree of correlation.

2.3.2. *The Entropy of Folding for a Single Cross-link*

In the GPC model (Appendices B through D), the entropy is expressed by the root mean separation distance $r_{i,j}$ between the BP i and j, where $r_{i,j}$ represents a state of thermodynamic equilibrium but need *not* be the equilibrium separation distance $R_{i,j}$. (*Note*: here we are using the "mer" indices i and j.) The change in the entropy of the GPC as a function of $r_{i,j}$ is

$$\Delta S(r_{i,j}) = (S(r_{i,j}) - S_0)$$

$$= k_B\left\{\ln(C_{i,j}^{\gamma}) + 2\gamma\ln\left(\frac{r_{i,j}}{\tilde{b}}\right) - \frac{3}{2\tilde{N}_{i,j}}\left(\frac{r_{i,j}}{\tilde{b}}\right)^2\right\}, \tag{6}$$

where S_0 is a reference entropy (Appendix B), $S(r_{i,j})$ is the entropy of the GPC for an end-to-end displacement $r_{i,j}$ between base i and base j, k_B is the Boltzmann constant (about $0.002\ \text{kcal mol}^{-1}\ \text{K}^{-1}$) and $C_{i,j}^{\gamma}$ is a constant associated with normalization of the probability density function (Appendix C).

It is noted that full thermodynamic reversibility (Appendix A) has only been shown for *some* nucleic acid sequences such as group I introns (Pan & Woodson, 1999) or special conditions of tRNA (Brion *et al.*, 1997). In Part II, a kinetic model is developed which is able to address this issue. Here we must *assume* that the reactions are reversible in this system. Given so, the reference

state entropy (S_0) in eqn (6) is cancelled out by measuring two states of the system. For these two states, we choose the statistical mean for the end-to-end separation of the Gaussian polymer chain corrected for the excluded volume ($R_{i,j}$) as one state (Appendix C), and a hybridized or folded structure (r_f) for the other (Section 2.2.3). The state r_f is well defined because the "stacking" locks the chain into a fixed end-to-end separation that is maintained over a sufficiently long time scale that thermodynamic averaging is permitted.

Equation (6) predicts that the change in entropy due to unfolding (e.g. breaking the hydrogen bonds between a BP) consists of the transition $r_f \to R_{i,j}$:

$$\Delta S_{i,j}^{f\to c} = \Delta S(R_{i,j}) - \Delta S(r_f) = 2\gamma k_B \ln\left(\frac{R_{i,j}}{r_f}\right) - \frac{3k_B}{2\tilde{N}_{i,j}\tilde{b}^2}(R_{i,j}^2 - r_f^2), \tag{7}$$

where $f \to c$ denotes a transition from the "folded state" to the "coil state" (Section 2.2).

It must be recalled at this point that the "links" in the GPC are not necessarily the same as the separation between the "mers" (recall that i and j represent the monomer positions, *not* the "link" positions). If the environment resembles Fig. 3 ($\tilde{b} < b$), then eqn (7) can be utilized without any adjustments because each bond functions as a *discrete* unit. However, if the environment resembles Fig. 2 ($\tilde{b} > b$), then the bonds are acting collectively and we must weight the contribution from each bond accordingly to avoid overestimating the number of degrees of freedom in the system. Figure 2 suggests that a sensible approximation would be to average the entropic contribution from the "mers" over the persistence length $\tilde{b}$ (Fig. 2: where the overlapping bonding contributions from the "mers" are indicated by the thin lines that intersect the gray bars at right angles and the weighted contribution from these monomer bonds is represented by the hemispheric balls at the center of the gray bars). For $\xi > 1$, the weighted entropic contribution of (i, j) is the sum of the bond formed at (i, j) plus the neighboring bonds within the same "link" of the sequence averaged over the persistence length. Hence, the weight function ($\Theta(\xi)$) for the entropy of a single bond at (i, j) is roughly

$$\Theta(\xi) = \begin{cases} 1, & \xi < 1, \\ 1/\xi, & \xi \geqslant 1 \end{cases}. \tag{8}$$

Equation (8) is discontinuous at $\xi = 1$ which is (at best) true for a genuine GPC where the joints are connected by infinitely small springs whose flexibility even permits contortions that are impossible for real molecular bonds. The essential properties of the weight function are $\xi < 1 \Rightarrow \Theta \sim 1$ and $\xi > 1 \Rightarrow \Theta \propto 1/\xi$. The function $\Theta(\xi) = 1/(1+\xi)$ roughly satisfies this and is continuous; however, it has the greatest error in the range surrounding $\xi \sim 1$.

Substituting eqn (5) into eqn (7) and including corrections for the persistence length [eqn (8)], the estimated entropic contribution of "mer" i (associated with a "link" encompassing point i), and "mer" j (associated with a "link" encompassing point j) will be

$$\langle \Delta S_{i,j}^{f\to c} \rangle = \gamma k_B \Theta(\xi)\left(\ln(\psi N_{i,j}) - 1 + \frac{1}{\psi N_{i,j}}\right), \qquad \psi = \frac{2\gamma\xi}{3\lambda^2}. \tag{9}$$

This represents the weighted contribution of (i, j) to the cross-linking entropy.

If $\xi \gg 1$, then the primary *extensive* effect on the entropy appears in the logarithmic term which increases as a function of the sequence length and, for $\xi \ll N_{i,j}$, the dominant contribution to eqn (9) will be the logarithmic term

$$\langle \Delta S_{i,j}^{f\to c} \rangle \approx k_B \gamma \Theta(\xi)\{\ln(N_{i,j}) + \ln(2\gamma\xi/3\lambda^2)\}, \tag{10}$$

which closely resembles the Jacobson & Stockmayer's (1950) equation. Moreover, the cross-linking entropy becomes negligible for large ξ ($\lim_{\xi\to\infty}\Delta S = 0$). The R^2 contribution in eqn (9) reduces to a constant which is independent of N and the contribution to the entropy due to r_f vanishes at the rate of $1/\xi N_{i,j}$.

On the other hand, if $\xi \ll 1$, the dominant contribution will be from the $1/\psi N_{i,j}$ term

$$\langle \Delta S_{i,j}^{f\to c} \rangle \sim k_B \left\{\frac{3\lambda^2}{2\xi N_{i,j}}\right\} \tag{11}$$

because $\lim_{\xi \to +0} (|\ln(\xi)|/(1/\xi)) = 0$. Since λ is proportional to the end-to-end distance of the enclosed polymer chain, the result is exactly the stretching term in a GPC (Grosberg & Khokhlov, 1994). The entropy is positive because this represents a situation where the polymer is actually *stretched* between the stacked monomers across the stacking gap [$\lambda b > R_{i,j}$: where (i,j) closes the loop in Fig. 4].

Equation (9) also has a minimum where $\Delta S \to 0$ when $\lambda b \to R_{i,j}$ (i.e. $N_{i,j} = 1/\psi$). This has little bearing on the current work except that the binding of a protein to the major grove or loop regions of the A-RNA helix (Draper, 1999) and the intercalation of a metal ion or a water molecule into a loop or a bulge of the RNA structure (Hermann & Patel, 1999; Holbrook & Kim, 1997) might take advantage of this feature of polymer physics to enhance its binding properties.

This decomposition of the "link" into a "mer" by "mer" average does imply that a particular stem length (L_{stem}) formed by a given "link" could be either longer ($L_{stem} > \xi$) or shorter ($L_{stem} < \xi$) than an integral "link" distance (for $\xi > 1$). In a final model, some accounting is of considerable relevance. However, for $L_{stem} < \xi$, the stems would turn out to be quite "flimsy" since the corresponding enthalpic contribution is also lost. Likewise, for $L_{stem} > \xi$, the stems would be quite "stiff" due to the presence of additional enthalpic terms. Hence, if ξ becomes too large, short loops will be eliminated quickly from consideration due to the "lack of funds" (i.e. enthalpy) needed to maintain them. Averaging is merely used to *estimate* the CLE contribution. We also reiterate that ξ is probably *not* a constant over an entire sequence and independent of the choice of state (coil/native) as we assume in this simple model.

2.3.3. *Cross-linking Entropy for Domains in a GPC*

In the previous sections, we introduced four undefined parameters: the monomer separation distance (b), the persistence ratio (ξ), the excluded volume (γ), and the stacking gap ($\lambda b (= r_f)$). Since the theoretical concepts should allow for flexibility in these variables, we will refrain from introducing any experimentally known values for b, ξ, γ or λb until Section 3.1.

In general, the entropy contribution of a domain k will be

$$\langle \Delta S_k^{f \to c} \rangle = \sum_{b_{ij}}^{m_k} \langle \Delta S_{b_{ij}k}^{f \to c} \rangle, \tag{12}$$

where k is the index of the domain, m_k is the number of BPs in domain k, b_{ij} is an index that specifies a particular BP (i,j) in domain k, and $\langle \Delta S_{b_{ij}k}^{f \to c} \rangle$ is the entropy change caused by *unfolding* the nucleic acid from a structure in which b_{ij} is present to its equilibrium thermodynamic displacement (R_{ij}): the ideal denatured state. For any given (i,j) in a GPC, $R_{ij}^2 \propto j - i + 1$. The "cost" is estimated from eqn (9) to be the change in entropy resulting from *un*folding the chain (at position (i,j)) from its hydrogen bonded value r_f to its equilibrium thermodynamic displacement (R_{ij}):

$$\langle \Delta S_k^{f \to c} \rangle = \gamma k_B \Theta(\xi) \sum_{b_{ij}}^{m_k} \left\{ \ln(\psi N_{b_{ij}k}) - \left(1 - \frac{1}{\psi N_{b_{ij}k}}\right) \right\}, \tag{13}$$

where $N_{b_{ij}k} (= j - i + 1)$ is the segment length enclosed by a given BP (i,j) in domain k. This function is positive for any reasonable value of $N (= N_{b_{ij}k})$ (see Fig. 6).

If the persistence length can be treated as a constant over the entire region of domain k, then the total entropy of domain k can be divided into three separate terms which correspond to the sum of the cross-linking contributions in domain k for a particular secondary structure

$$\langle \Delta S_k^{f \to c} \rangle = \gamma k_B \Theta(\xi) \times \left\{ \sum_{b_{ij}}^{m_k} \ln(\psi N_{b_{ij}k}) + \sum_{b_{ij}}^{m_k} \frac{1}{\psi N_{b_{ij}k}} - m_k \right\}. \tag{14}$$

When $m_k = 1$, this expression reduces to eqn (9).

Since we have been *assured* that reactions are reversible (Section 2.3.2), $\langle \Delta S_k^{c \to f} \rangle = - \langle \Delta S_k^{f \to c} \rangle$, where the superscript $c \to f$ denotes a transition

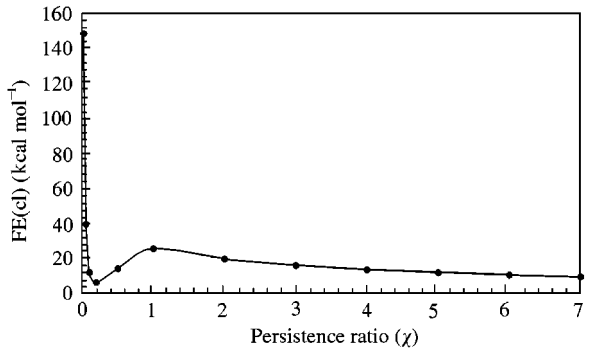

FIG. 6. A description of the free energy due to cross-linking entropy—FE(cl)—as a function of the persistence ratio (ξ) for tRNA^{Phe} based on the best secondary structure prediction of tRNA found by MFOLD. For $\xi \gg 1$, the CLE reflects the cost to *squeeze* the sequence together into a shape conforming to tRNA with the corresponding stacking gap separation between all the BP (eqn (10)), and for $\xi \ll 1$, the CLE reflects the cost to *stretch* a sequence apart to the same state (eqn (11)). Clearly, one can immediately presume that a very small ($\xi \ll 1$) is unlikely for tRNA.

from the "coil state" to the "folded state": the reverse process of $f \to c$. The total entropy to *fold* the sequence into the hybridized state is the sum of all the cross-linking entropies from the individual domains of the sequence (note opposite sign!)

$$\langle \Delta S_{cl}^{c \to f} \rangle = - \sum_{all\ k} \langle \Delta S_k^{f \to c} \rangle, \qquad (15)$$

where the cross-linking entropic contribution is denoted by the subscript "*cl*". The cross-linking entropy $\langle \Delta S_{cl}^{c \to f} \rangle$ is inserted into the Gibbs equation to find the total FE

$$\begin{aligned} \Delta \mathcal{G} &= (\Delta \mathcal{G}_{NN} + \langle \Delta \mathcal{G}_{cl} \rangle) \\ &= \Delta \mathcal{H}_{NN} - T(\Delta S_{NN} + \langle \Delta S_{cl}^{c \to f} \rangle), \end{aligned} \qquad (16)$$

where $\Delta \mathcal{G}$ defines the total Gibbs FE, $\langle \Delta \mathcal{G}_{cl}^{c \to f} \rangle$ is the Gibbs FE due to cross-linking and the subscript "NN" denotes local or "nearest-neighbor" regime where $\Delta \mathcal{H}_{NN}$ is the NNSS result for the enthalpy and ΔS_{NN} is the nearest-neighbor contributions to the entropy. Then

$$\langle \Delta \mathcal{G}_{cl} \rangle = - T \langle \Delta S_{cl}^{c \to f} \rangle = T \sum_{all\ k} \langle \Delta S_k^{f \to c} \rangle. \qquad (17)$$

More will be discussed on the evaluation of $\Delta \mathcal{G}_{NN}$ in Part II of this work.

Fig. 6 shows the tendencies of the CLE as a function of ξ for a calculation of tRNA^{phe}. For $\xi \to \infty$, we obtain an NNSS-like treatment: $\langle \Delta \mathcal{G}_{cl} \rangle = 0$. For $\xi \sim 1$, there is a local maximum. For $N = 1/(\psi) \propto 1/\xi$, there is an entropic minimum (Section 2.3.2), and for $\xi \to 0$, there is again a very large increase due to stretching of the chain across the gap of the helix. The latter case is rather odd, but would characterize a polymer with a very short persistence length relative to its cross-linking. From Fig. 6, it is clear that the CLE dominates the FE at very small ξ and to a lesser extent in the range of $1 < \xi < 10$. Likewise, an FE minimum is found for $\xi \sim 0.2$ (for the current parameters: $\lambda = 2$ and $\gamma = 1.75$) and $\xi \to \infty$ contributes almost nothing to the overall CLE of the polymer.

3. Results

A test of this CLE strategy with respect to specific examples of RNA secondary structure will be reported in Part II of this work along with further development of the basic theory presented here.

Here, we discuss the known experimental values for nucleic acids and will use these parameters (within their discussed limits) in the remainder of this work as the basis for predictions using the CLE theory and for making comparisons with NNSS strategies. We later show why the entropy estimation used in traditional NNSS strategies appears to work under certain specified conditions (i.e. short domain sizes). The latter condition is shown by demonstrating that, the "sensible" parameterizations yield the same entropic penalties currently used in NNSS approaches (to within an additive constant).

3.1. EXPERIMENTAL PARAMETERS: b, λ, γ AND ξ

In Section 2.3, we side stepped any discussion about appropriate values for b, λb, γ and ξ. Here, we report sensible values for these parameters in the context of RNA secondary structure calculations.

3.1.1. *The Bond Distances* (b *and* λ)

The distance (b) between the nearest-neighboring ribofuranosyl sugars (on the *same* chain) of

a hybridized A-RNA structure is roughly 5.9 Å (Wyatt & Tinoco, 1993), and the distance between the C'_1 bonds‖ of the hybridized pair is roughly 10.5 Å for A-RNA (obtained from Insight II, Molecular Biosym Corp.). Hence, a reasonable estimate for r_f is $r_f \sim 2b$, or $\lambda \sim 2$. We will assume that the stacking interaction reduces the distance between monomers i and j such that (i,j) has a stacking gap (separation distance) of $r_f = 2b$.

3.1.2. *The Excluded Volume* (γ)

The value for the excluded volume has been reported in the literature to be $\gamma = 1.75$ in 3-D (Fisher, 1966; Zuker *et al.*, 1998). There is no reason to assume that this parameter should be changed in this work. In the course of this work, we have experimented with different values of excluded volume ($1 \leqslant \gamma \leqslant 2$: Appendix C); however, we think that $\gamma = 1.75$ is the most justifiable parameterization based on the literature (Fisher, 1966), and will use this value exclusively in this work.

3.1.3. *The Persistence Ratio* (ξ)

Currently, there is very little experimental information available on the persistence length. A list of known values is compiled in Table 1.

There are five potentially flexible bonds between each nucleotide in an RNA sequence. An additional degree of flexibility comes from the small amount of pucker (most commonly the 2′ endo and 3′ endo; (Gautheret & Cedergren, 1993; Gelbin *et al.*, 1996)). Hence, the six bonds that join a single nucleic acid between the 5′ and 3′ ends (with an average separation of about 1.5 Å) should allow a fair degree of flexibility in such a sequence. However, in the case of nucleic acids of single strand RNA and DNA, the persistence length appears to be similar to Fig. 2 where the stems are rather inflexible. The experimentally determined values for the persistence ratio of single strand RNA and DNA are shown in Table 1. After correcting for the excluded volume γ, the persistence ratio (ξ) at 37°C is approximately 3 (or $\tilde{b} \sim 17.7$ Å) in most of the data listed in Table 1.

The value of ξ also shows temperature dependence in the persistence length (Fig. 7). In poly-A (Eisenberg & Felsenfeld, 1967), there is a rapid decrease in the persistence length as the temperature increases (Fig. 7). This is most likely due to the gradual break up of non-Watson–Crick pairing between adenines in the sequence (Burkard *et al.*, 1999). As the temperature increases, the stems formed by non-Watson–Crick-type BPs gradually shorten in length. For poly-U (Inners *et al.*, 1970) the temperature dependence of $\tilde{b}$ above 15°C appeared to be flat or increases slightly. It is not clear as to why $\tilde{b}$ should increase with increasing temperature; however, this may reflect swelling (Grosberg & Khokhlov, 1994) of poly-U: another phenomena of polymers which we have neglected to mention in part because we assume that our miraculous "ideal solvent" is capable of compensating for all these uncertainties. The increase is small in these measurements. The sequence of poly(AUGC) (Achter & Felsenfeld, 1971) was measured with various bases randomly removed from the sequence. No temperature dependence is indicated explicitly in Achter *et al.* (1971), but the authors state that the properties are similar to poly-U. The remaining measurements were done at only one temperature; hence, no temperature dependence can be extrapolated. The measurements on mRNA (β actin) were carried out *in vivo* on cells from a rat kidney (Femino *et al.*, 1998). This would suggest that the global structure of mRNA resembles a self-avoiding polymer chain under biologically active conditions. Nevertheless, there still is likely to be local structure since the β actin must be transported to the cytosol.**

The variability of the persistence length is particularly important in considerations about the hairpins and stems in folded nucleic acid sequences. The minimum hairpin loop size closing a typical nucleic acid sequence is assumed to be 3 nt in NNSS calculations. This is exactly the persistence length of the examples listed in

‖ The bond joining the ribofuranosyl sugar to the nucleotide base: A, C, G, or U.

** The measurement was done with the cells fixed. By "fixed", the terminology implies "fixed for good"; hence, it is still arguable that the properties of the mRNA of a dead cell are quite different from those of a living cell.

TABLE 1

*A list of known experimentally obtained persistence ratios for single strand RNA and DNA (ssRNA, ssDNA). An additional row indicates the value for double strand DNA (dsDNA) for comparison**

Polymer sequence	Experimental technique	Persistence ratio $\langle R^2\rangle/Nb^2$	Persistence ratio ξ'	Persistence ratio ζ	Experimental conditions	Source
poly(rA)	LS, Sd	23.0	5.9	3.4	26°C, 1.0 M NaCl, pH ~7	a
ssRNA(†)	LS, Sd	18.2	5.2	3.0	18°C, 1.0 M NaCl, pH ~7	b
poly(rU)	LS, Sd	17.6	5.1	2.9	18°C, 1.0 M NaCl, pH ~7	c
ssDNA	AFM		4.7	2.7	Ambient temp., 0.15 M NaCl, pH 8	d
ssDNA	AFM		4.4	2.5	Ambient temp., 0.15 M NaCl	e
ssRNA (mRNA)	FISH	25.5	6.2	3.5	*In vivo*	f
dsDNA	AFM			~150	Ambient temp.	g

*Column ξ' is the persistence ratio without any corrections for the excluded volume ($\xi' = \gamma\xi$). Column ξ shows the persistence ratio after correcting for the excluded volume using $\gamma = 1.75$. Both column ξ' and ξ are increased by a factor of 1.22 to reflect the $\sqrt{3/2}$ in eqn (3).The persistence length is $\tilde{b} = \xi b$, where $b = 5.9$ Å. The cited measurement techniques include light scattering (LS), sedimentation (Sd), atomic force spectroscopy (AFM), and fluorescence *in situ* hybridization (FISH).The following references were used: (a) (Achter & Felsenfeld, 1971); (b) (Eisenberg & Felsenfeld, 1967); (c) (Inners & Felsenfeld, 1970); (d) (Rief *et al*., 1999); (e) (Smith *et al*., 1996); (f) (Femino *et al*., 1998) (g) (Smith *et al*., 1992). Ref (d) and (e) are estimated from the quoted persistence lengths: 16 and 15 Å, respectively using the nominal "mer" separation distance 3.4 Å along the linear double stand chain of DNA. Likewise, ref (g) is estimated in the same way as ref (d/e) for $\tilde{b} = 500$ Å. Ref (f) is estimated from the measured end-to-end separation (6×494 Å, where 494 Å is the radius of gyration) and the number of nucleotides in rat kidney β-actin mRNA sequence (1648 nt). (†) Ref (b) involves a composite material in which some of the bases are missing in the polymer chain, hence these values are expected to be only a remote approximation.

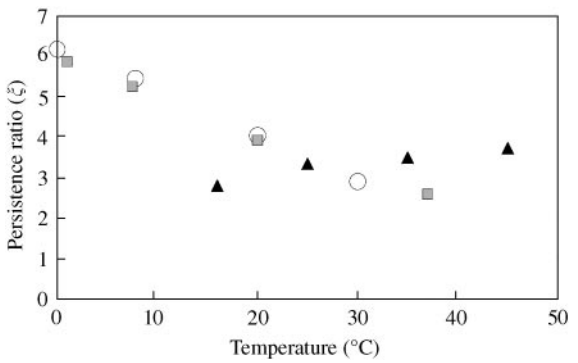

FIG. 7. Temperature dependence of the persistence length for poly-A (○, 1740 nt; ■, 1462 nt: Eisenberg *et al*., 1967) and poly-U (▲, 880 nt: Inners *et al*., 1970). The decrease in the poly-A reflects the reduction in the length of stems resulting from non-Watson Crick base pairing. Above 30°C, both poly-A and poly-U have roughly the same persistence length.

Table 1 and Fig. 7, and suggest that "rigidity" is a major component in determining the minimum size of a loop. However, it should be kept in mind that the loop region could have a different ξ than the stem because the context is different. The stems are also likely to have context dependent persistence lengths due to the subtle base pairing differences of AU, GC, and GU bases. These matters will be considered elsewhere.

3.2. CONNECTION BETWEEN CLE APPROACH AND TRADITIONAL NNSS APPROACHES

Now that we have developed the CLE model and have established an order of magnitude for the persistence length in Section 3.1, we can now proceed to show that the CLE reduces to the traditional NNSS penalties when conditions are properly specified. In this way, the NNSS penalties should be understood to be a specialized subset of parameters which will work successfully for certain types of calculations and will yield results similar to the CLE. The CLE is an attempt at generalizing the rules used in traditional NNSS algorithms. In addition, we use these generalizations to help develop a theoretical prediction about folded ssRNA and ssDNA sequences for the regime where $\xi > 1$.

3.2.1. *The Estimated CLE Free Energy of a Stem*

The cross-linking entropy tends to weight the size of the domain according to the following lemma.

Lemma 1. *For a specified domain k of length ΔN_k, suppose that only half the BPs are stacked.*

Suppose further that this stacking follows a regular pattern such that it forms a single loop which is closed at $N_{ij} = 6$ and only the next-nearest-neighboring BPs are cross-linked (Fig. 8). Then for large ΔN_k, the total entropy of domain k (with $\xi \geqslant 1$) will be approximately equal to

$$\Delta S_k = (1/4)(\kappa/T)\Delta N_k \ln(\psi \Delta N_k), \qquad (18)$$

where $\kappa = k_B T\gamma/\xi$, for $\xi > 1$.

Proof. The base pairing renders a combinatorial series where $j - i = 6, 10, 14 \ldots$. From eqn (10), the maximum entropic contribution on $\Delta S_{b_{ij}k}$ (Section 2.3) will come from the total length of the enclosed subsequence. The lengths of the separate enclosed subsequences become a product

$$\ln(6) + \ln(10) + \ln(14) \cdots + \ln(\Delta N_k)$$

$$= \ln(6 \cdot 10 \cdots \Delta N_k).$$

Therefore, the combinatorial pattern will be

$$\frac{1}{5!} \frac{2^{(\Delta N_k/2)}(\Delta N_k/2)!}{4^{((\Delta N_k - 2)/4)}((\Delta N_k - 2)/4)!}. \qquad (19)$$

Applying Sterling's formula and preserving only the leading terms, we obtain eqn (18). □

The same procedure can be used to show that if all the BPs are cross-linked in Fig. 8, then

$$\Delta S = (1/2)(\kappa/T)\Delta N_k \ln(\psi \Delta N_k). \qquad (20)$$

In an actual subsequence, the cross-linking entropy of that domain cannot be determined without calculating the specific configuration. However, if we imagine the set of all shuffled subsequences of the same length and base composition, then it would be reasonable to propose that a grid-like structure having equal spacing between adjacent BPs (as in Fig. 8) is the most reasonable "average structure" for an arbitrarily selected sequence in which only half the positions are likely to be cross-linked.

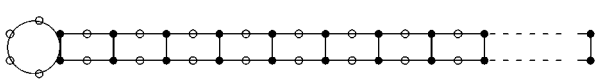

FIG. 8. A model of a hairpin loop (enclosing 4 nt) in which the stacking interactions only occur between every other BP in the sequence. Open circles refer to the unpaired bases and filled circles refer to the cross-linked nucleotides.

We now generalize this result to a sequence of fractional percentages ACGU (p_A, p_C, p_G and p_U). The pre-factor in both eqns (18) and (20) is related to the probability of cross-linking. It is shown in (Dawson & Yamamoto, 1999b) that the mean free energy of a set of shuffled sequences is a function of half the maximum number of BPs that are possible in an ideally arranged sequence weighted by the FE of the type of BPs. For a sequence of arbitrary percentages of ACGU, the *maximum* number of BPs that can form will be roughly proportional to $min(p_A, p_U) + min(p_G, p_C)$ (where we neglect the cross-correlation between AU and GU pairs and assume homogeneity of the general sequence). For example, $p_C = p_G = 0.5$ can potentially form BPs with the entire sequence, hence the pre-factor 1/2 that appears in [eqn (20)]. However, in any given shuffled sequence of $p_C = p_G = 0.5$, only *half* the maximum number of possible BPs are likely to be found; hence, the *mean* cross-linking entropy (mCLE) for this shuffled sequence will be [eqn (18)]. A rough estimate of the mCLE to domain k is

$$\langle \Delta \bar{S}_k \rangle \approx (p/2)(\kappa/T)\Delta N_k \ln(\psi \Delta N_k), \qquad (21)$$

where $p = \{min(p_A, p_U) + min(p_C, p_G)\}$.

From eqn (21), it is clear that the domain size is a function of both the number of cross-links and the persistence length of the given RNA sequence. If ξ is very large, the cross-linking entropy becomes vanishingly small because $\Delta S \propto 1/\xi$.

3.2.2. *Correspondence Between NNSS and CLE Models of Entropy*

Postulate 1. *The FE of an internal loop region is related to the difference between the subdomain formed by (i′, j′) and the subdomain closed by (i, j) in Fig. 9(a).*

Evidence. We start by showing that the enclosed region forming an internal loop can be approximated by a linear penalty when $N \gg \delta$ in Fig. 9(a). If the general form for the internal

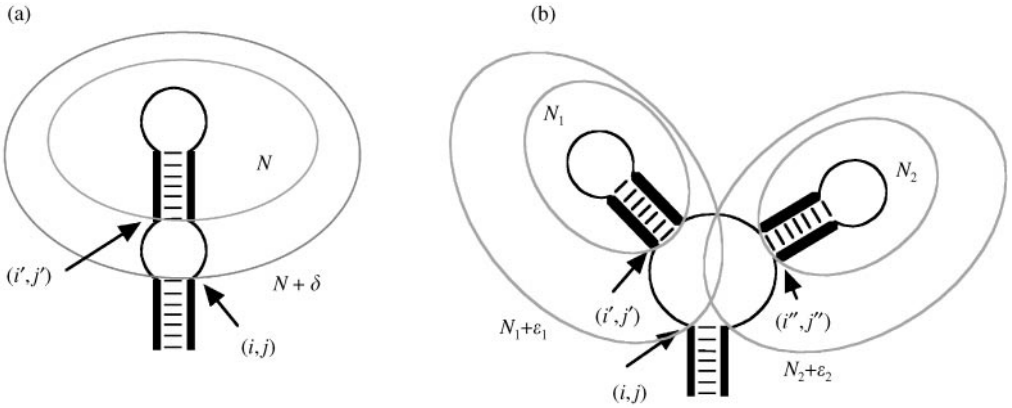

FIG. 9. A conceptual model for understanding Predictions 1 and 2: (a) a symmetric internal loop, where (i',j') encloses a subdomain of N nt (inner circle) and (i,j) encloses the internal loop of $N+\delta$ nt (outer circle). In (b), the subdomain enclosed by (i',j') contains N_1 nt and the larger circle encloses $N_1+\varepsilon_1$ nt. Likewise, the subdomain enclosed by (i'',j'') has N_2 nt and the larger circle encloses $N_2+\varepsilon_2$ nt. The outer circle encloses both (i',j') and (i'',j'') with a total of $(N_1+N_2)+(\varepsilon_1+\varepsilon_2)=N+\varepsilon$ nt.

loop penalty is obtained, then we have shown consistency with Postulate 1. We neglect all asymmetric penalties in these considerations (Mathews *et al.*, 1999) which are important in more precise evaluations but should not affect these general observations.

From Fig. 9(a), let N be the enclosed segment between (i',j'). Let δ represent the difference in the length of the sequence that encloses the region between (i',j') and the enclosed region at (i,j), where $\delta \ll N$. Neglecting all pre-factors: the cross-linking entropy enclosing segment (i',j') is approximately $N\ln(N)$ [using eqn (10)]. Likewise, the cross-linking entropy at (i,j) is $(N+1)\ln(N+\delta)$, where the $(N+1)$ reflects the addition of one more bond at position (i,j) than at (i',j'). Then the difference in the cross-linking entropy contribution in the segment between (i',j') and (i,j) is approximately

$$\frac{\langle \Delta \bar{\mathscr{G}}^{\mathscr{I}}_{cl} \rangle}{(\kappa p/2)} = (N+1)\ln[\psi(N+\delta)] - N\ln(\psi N), \tag{22}$$

where $\kappa = k_B T\gamma/\xi$ (because we assumed $\xi > 1$) and $\langle \Delta \bar{\mathscr{G}}^{\mathscr{I}}_{cl} \rangle$ is the *mean* FE difference of the internal loop due to cross-linking [eqn (21)] in a *shuffled* sequence of ACGU (fractional percentages: p_A, p_C, p_G, and p_U) (see also Section 2.3.3).

Neglecting all higher-order powers of δ/N, we have $N\ln[\psi(N+\delta)] - N\ln(\psi N) \approx \delta$. Hence,

$$\langle \Delta \bar{\mathscr{G}}^{\mathscr{I}}_{cl} \rangle \approx \frac{\kappa p}{2}(\delta + \ln(\psi N)). \tag{23}$$

Using the fact that $\delta > \ln(\delta)$ and making the crude approximation that $\ln N \sim$ "Const" for a fixed set of measurements on sequences of similar length, we arrive at an expression that resembles the penalties found in NNSS algorithms

$$\langle \Delta \bar{\mathscr{G}}^{\mathscr{I}}_{cl} \rangle \approx \left(\frac{\kappa p}{2}\right)\ln(\delta) + \text{"Const"}. \tag{24}$$

For $\xi \sim 3$, $p \sim 0.5$, and assuming *all* cross-links are formed on the stem [i.e. starting from eqn (20) instead of eqn (21)], the pre-factor in eqn (24) becomes $\kappa p \sim 0.2$ kcal mol at 37°C. This is almost in quantitative agreement with the *slope* found in the entropy increase per nucleotide for small n in the $\mathscr{I}$ penalty look up tables and is also consistent with the experimental conditions (Table 1; Freier *et al.*, 1986). The "Const" on the right-hand side of eqn (24) is *not* constant [proportional to $\ln(\psi N)$] and reflects the real cost of closing $\mathscr{I}$ as the size of the enclosed subdomain region increases.

Postulate 2. *The linear contribution to the penalty in iMBL structures calculated in NNSS*

algorithms is related to the difference between the average cross-linking entropy contribution from the separate branch point subdomains and the cross-link that closes the iMBL subdomain [*Fig.* 9(*b*)].

Evidence. Here, we will show that the solution is bounded for the limiting cases and that the solution has the same form as Postulate 1 in its lower bound.

In Fig. 9(b), region (i', j') and (i'', j'') enclose N_1 and N_2, respectively. The region enclosed by (i, j) is $(N_1 + N_2) + (\varepsilon_1 + \varepsilon_2) = (N + \varepsilon)$, where $N = N_1 + N_2$ and $\varepsilon = \varepsilon_1 + \varepsilon_2$. For the structure in Fig. 9(b), the difference in the cross-linking FE is

$$\frac{\langle \Delta \bar{\mathscr{G}}_{cl}^{\mathscr{M}} \rangle}{(\kappa p/2)} = (N + 1)\ln[\psi(N + \varepsilon)] - N_1 \ln(\psi N_1) - N_2 \ln(\psi N_2), \tag{25}$$

where $\langle \Delta \bar{\mathscr{G}}_{cl}^{\mathscr{M}} \rangle$ is the *mean* FE difference for the iMBL region due to CLE.

First we examine the case where $N_1 = N_2 = N/2$. Equation (25) becomes

$$\langle \Delta \bar{\mathscr{G}}_{cl}^{\mathscr{M}} \rangle \approx (\kappa p/2)\{\varepsilon + \ln(\psi N) + N \ln(2)\}.$$

To obtain the limits for $N_1 \gg N_2$, we divide both sides of eqn (25) by $N_1 N_2$ and use the fact that $\ln[\psi(N + \varepsilon)] - \ln(\psi N_1) > \ln[\psi(N_1 + \varepsilon)] - \ln(\psi N_1)$ to obtain

$$\frac{\langle \Delta \bar{\mathscr{G}}_{cl}^{\mathscr{M}} \rangle}{(\kappa p/2)} > N_1\{\ln[\psi(N_1 + \varepsilon)] - \ln(\psi N_1)\} + \ln[\psi(N + \varepsilon)] \approx \varepsilon + \ln(\psi N), \tag{26}$$

which yields the inequality

$$\varepsilon + \ln(\psi N) < \frac{\langle \Delta \bar{\mathscr{G}}_{cl} \rangle}{(\kappa p/2)} \leqslant \varepsilon + \ln(\psi N) + N \ln(2).$$

For an iMBL of k branching points

$$\langle \Delta \bar{\mathscr{G}}_{cl}^{\mathscr{M}} \rangle \approx \left(\frac{\kappa p}{2}\right)\left\{\varepsilon + \ln(\psi N) + \sum_{i=1}^{k} N_i \ln\left(\frac{N}{N_i}\right)\right\}$$

which is bounded between the limits

$$\varepsilon + \ln(\psi N) < \frac{\langle \Delta \bar{\mathscr{G}}_{cl}^{\mathscr{M}} \rangle}{(\kappa p/2)} \leqslant \varepsilon + \ln(\psi N) + N \ln(k). \tag{27}$$

Hence, in the limit, $N_1 \gg N_2$, $N_3 \cdots N_k$, the $\mathscr{I}$ solution prevails, and for $N_1 = N_2 \cdots N_k = N/k$, the FE (i.e. the entropy decrease) tends toward a maximum.

These observations suggest that the entropic penalties of the $\mathscr{H}$s, $\mathscr{B}$s, $\mathscr{I}$s and iMBLs currently evaluated as discrete entities, are more likely to be closely connected with the cross-linking entropy. For fully cross-linked stems [i.e. starting from eqn (20)], the pre-factor $p\kappa$ is 0.2 kcal mol^{-1} (at 37°C, $p \sim 0.5$ and $\xi \approx 3$) which is consistent with the values used in NNSS algorithms: between 0.2 (Williams & Tinoco, 1986) and 0.4 kcal mol^{-1} at 37°C (Zuker *et al.*, 1998). The main point of approaching the problem by way of eqns (24) and (27) is to *illustrate* that the cross-linking entropy and traditional NNSS approaches have some common ground in which they both are likely to produce similar results.

Equation (27) also shows that a symmetrically shaped distribution of branch point stems extending off of iMBLs (such as shown in Fig. 9) has the maximum FE and an asymmetric configuration of branching stems has the minimum FE.

Postulate 3. *The loop penalty reflects the mean cross-linking entropy cost required to close a free segment of length n into a loop with m BPs.*

Evidence. A simple consideration of how cross-linking entropy is calculated is sufficient for Postulate 3. A comparison of the penalties used in the Turner energy rules, and a sequence containing 4 bp and a loop size of n (total length: $N = 4 \times 2 + n$) is shown in Table 2, where $\xi = 3.0$ for the stem and the loop regions, respectively. The result matches the Turner energy rules (Turner *et al.*, 1988) to within a constant [the "difference" column in Table 2, cf., eqn (24)]. Hence, the tendency is exactly the same and a persistence ratio of 3.0 is already quite reasonable as shown in Section 3.1. This further

TABLE 2
Comparison between the cross-linking entropic penalty, and the entropic penalties used in the Turner energy rules for a loop which is closed by a stem composed of 4 *bp**

Loop length n (nt)	$\mathscr{G}_{cl}$ (kcal mol^{-1})	$-T\Delta S_{ss}$ (kcal mol^{-1})	Difference (kcal mol^{-1})
3	1.533	4.100	2.567
4	1.688	4.900	3.212
5	1.828	4.400	2.572
6	1.955	4.700	2.745
7	2.072	5.000	2.928
8	2.180	5.100	2.920
9	2.280	5.200	2.920
10	2.374	5.300	2.926
15	2.769	5.800	3.031
20	3.078	6.100	3.022
25	3.332	6.300	2.968
30	3.548	6.500	2.952

* The first column indicates the length of the loop in terms of the tabulated loop penalties. The second column indicates the CLE. The third column corresponds to the Turner energy rules. The last column shows the difference between the two values (which is nearly constant for all $n > 6$). The effect strongly resembles eqn (24). Values used in this calculation were $\xi = 3.0$, $\gamma = 1.75$. Other stem lengths are also amenable to this approach.

suggests that the minimum size of the loop is partly a function of the persistence length in the GPC.

4. Discussion

A theoretical model has been proposed to account for the folding behavior of RNA. The application of the model to various aspects of folding will be developed in Part II. However, enough has already been explored to draw generalizations about how this entropy behaves.

Using *known* parameters for RNA, the results already suggest that the current entropic penalties used in NNSS algorithms are essentially the average entropic contribution for a generic sequence of RNA with adjacent nearest-neighbor cross-links (complete cross-linking of the bonds shown in Fig. 9). This point will be further substantiated in Part II where the method is applied to secondary structure prediction.

The current treatment ignores any strain due to the asymmetry of the structures that are formed. Additional effects like the angular dependence of allowed monomer orientations (Flory *et al.*, 1966; Flory & Semlyen, 1966; Grosberg & Khokhlov, 1994) are likely to have an influence on the entropy, but are neglected in the current model. The local value of ξ (and the proximate neighboring values of ξ), will also lead to corrections in the predictions. None of these factors should be routinely ignored in a more precise treatment. However, these *local* effects do not detract from the central point of this series: large domains of high MBL hierarchal complexity (HC) formed into a thicket of long double helices of high base pair density (BPD) are unlikely in most biologically significant RNA because the entropy is a logarithmic function of the enclosed sequence length and the BPD.

The persistence ratio (ξ) is the most likely parameter to influence the entropy in any given sequence of RNA. This parameter is essentially a measure of the flexibility of the RNA (Hagerman, 1997). Although other parameters like the excluded volume (γ) could be important in certain unusual conditions, γ is unlikely to vary significantly from one sample to the next. The role of persistence length can be seen in the minimum loop size which has generally been thought to be about 3 nt (Scheffler *et al.*, 1970; Delisi & Crothers, 1971a) because the persistence length is also typically in the range of 3 nt. Since a piece of RNA cannot fold tighter than its own persistence length, this experimental evidence is another reflection of the flexibility of RNA. NNSS strategies cannot say anything about the flexibility of a given piece of RNA. Indeed, they even *assume* that the persistence length is less than 1 nt because they evaluate the enthalpy and entropy only in terms of the *nearest*-neighboring BPs. This is clearly contrary to what was found in Section 3.1.

The theory predicts from eqn (27) that balanced domain sizes [Fig. 9(b)] in the branching points $(N_1 = N_2 \cdots = N_k = N/k)$ will have a higher entropic cost because highly symmetric structures maximize the FE. Highly unbalanced structures where $N_1 \gg N_2, N_3, \cdots, N_k$ minimize the FE (due to the reduced symmetry). Hence, the CLE also has an influence on the *shape* of a functional domain. For example, a lever action due to an iMBL moving from a balanced subdomain

distribution to a more unbalanced subdomain distribution can now be envisioned as a possible heat engine for a molecular machine. This has been suggested for the T-loop of tRNA (Yamamoto *et al*., 1984). This effect is also suggested in the collapse of the P5abc stem of the self-splicing intron *T. thermophila* (Wu & Tinoco, 1998), although other explanations have also been proposed (Thirumalai, 1998). Likewise, one can envision a kind of ATP-driven pump action in the suboptimal structure where one branch grows longer while the other one becomes shorter. These locomotive properties of iMBLs are also likely to influence the type of conserved sequences which are found in ribozymes (Fontana & Schuster, 1998; Youhei & Yamamoto, 1994) because certain mutations may be lethal to this locomotion even when two given secondary structures look identical in their respective base pairing patterns. There does appear to be a predominance of unbalanced iMBLs in natural RNA.

Predictions using the cross-linking entropy model offer a thermodynamic mechanism that anticipates the break up of a loop from the 5′ to 3′ end rather than at the base (or closing point) of the loop. The cross-linking entropic effect will be much stronger at the 5′ to 3′ end of the chain because of the larger force that is required to hold the hairpin folded polymer chain closed [eqn (C.3), Appendix C]. As the temperature is raised, the probability of loss of structure will be highest where the structure is the weakest. Hence, a clear physical explanation is provided by the CLE to explain why the chain will unzip from the 5′ to 3′ end rather than from the base (closing point) of a hairpin loop. Likewise, the nucleation point is most likely to occur at the head of the hairpin rather than the tail because the lower resistance to folding and proximity of the adjacent links increases the probability of contact. (The evidence for this will be shown explicitly in Part II.) This property cannot be predicted from NNSS algorithms which say nothing about the global entropic contributions. Likewise, this folding property was only an assumption in the loop weight function models as no directionality can be decided from a constant weight (Scheffler *et al*., 1970; Delisi & Crothers, 1971b; Delisi, 1973b). In principle, the loop weight function approach could be modified by adopting the weight relations derived in eqn (9), where the degree of cooperativity is incorporated into the problem via the persistence length (the "link"). The Gaussian polymer network model (Keskin *et al*., 2000; Lustig *et al*., 1998; Debe & Goddard, 1999) also uses linear weights but could be modified accordingly to show this effect.

The concept of CLE is also applicable to the evaluation of the FE in protein folding problems. However, a clear distinction between the properties of α-helices and β-hairpins or β-sheets is required.

The α-helix forms a staircase which roughly makes one rotation with every four peptides (Poland & Scheraga, 1965; Gibbs & DiMarzio, 1958), forming linkages between every fourth peptide in a repetitive pattern. This cross-link separation distance is *approximately* the same for most peptides in the α-helix. Hence, the cross-linking entropy will be an additive constant along the chain of the α-helix. Indeed, the persistence length is also about three peptides in length or about 10 Å (Mueller *et al*., 1999; Reif *et al*., 1998). The α-helix can attain infinite length in this context (in principle) with only a linear increase in the entropy with each peptide cross-link in the chain. The α-helices are more likely to exhibit cooperative transitions due to the additive constant cost of collapsing the helix. This cooperativity is accounted for in the size of ξ which is likely to be quite variable in protein structure.

It should be clear that an infinite α-helix carries little functional utility. (The triple helix collagen serves a very important "function" and is effectively "infinite"; however, this is stretching the definition of "functionality" from the role of an active enzyme or protein, to a passive role.) An *active* biological protein must interplay the electro-mechanical properties of interacting parts to make an effective enzyme. This interplay is provided by the β-sheets, hairpins, etc.

In contrast to the α-helix, the β-hairpins involve folding the peptide chain into a loop of variable length (Fersht, 1999). These loops *will* exhibit cross-linking entropic effects which depend on the length of the β-hairpin that is formed. The cross-linking entropy in the various types of β-sheets will depend upon how the loops are closed. Again, the CLE must be averaged

over the persistence length and an exact model for λ (the stacking gap) as a function of the peptide is needed for the sheets.

The cross-linking entropy is likely to govern the length relationships between the structures of the α-helices and β-sheets in proteins. These can be coupled to other regions that may involve no structure in the isolated state and only take on order when attached to another biopolymer. These regions would resemble random coils or disordered regions in a protein (Garner *et al*., 1999; Li *et al*., 1999). Proteins often have peptide sequences shorter than 400 peptide. Perhaps proteins use a strategy of subunits to address functional domain size limits.

This key difference between α-helices and β-sheets or β-hairpins also applies to differences between dsRNA and ssRNA (likewise: dsDNA and ssDNA). Since the entropy of folding for dsRNA (or dsDNA) is only attributed to the freezing out of molecular bonds (loss of degrees of freedom), the entropic penalty is essentially constant for each BP formed. Consequently, the unzipping process in dsRNA (or dsDNA) can be modeled (Poland & Scheraga, 1966) without accounting for the logarithmic contribution because this penalty is essentially constant for a double helix (Poland & Scheraga, 1966) as it is for the α-helix (Gibbs & DiMarzio, 1958; Poland & Scheraga, 1965). Long chains of dsDNA will form with only a linear increase in "cost" and the dsDNA will behave like a GPC with a very long persistence length ~ 150 bp. Hence, building a human genome in dsDNA introduces no conflict with respect to the cross-linking entropy. On the other hand, in folding a segment of ssRNA (or ssDNA), a logarithmic weight will dominate the FE at very large domain sizes and high BPD.

Questionable predictions from NNSS algorithms are especially prevalent when there are long segments of somewhat similar repetitive sequence features present that extend over the entire length of a given sequence. For example, NNSS strategies often predict that the long spliceosome-type introns found in some mammalian mRNA form extremely long contiguous stems and, in some cases, secondary structures where HC > 12 (Dawson & Yamamoto, 1998, 1999a, c). In such cases, the HC often grows with added sequence length. Yet all of the known catalytic RNA structures including ribosomal RNA (rRNA) have comparatively short stems with *max*(HC) ~ 6 (Glotz & Brimacombe, 1980; Mueller *et al*., 2000; Wimberly *et al*., 2000). Unlike the ribozymes, spliceosome introns need "help" from the spliceosome to be extracted. Do examples where HC > 12 actually exist in nature? Likewise, an RNA sequence such as $(A_{100}U_{100})_{100}$ is predicted to form a single contiguous stem of 10^4 BPs. Yet is there any evidence for long stems of any comparable length existing either in nature or as artificial constructs? The CLE suggest that the answer to both questions is "probably not".

5. Conclusions

A Gaussian polymer chain model is used to describe the effects of intramolecular cross-linking (stacking) that occurs in biopolymers like single strand RNA. General approximations of this cross-linking entropy model imply that the standard entropic penalties used in secondary structure calculation algorithms are actually the default penalties for an average structure of RNA whose domain size does not exceed 100 nt and also introduces important concepts such as flexibility and persistence length to RNA secondary structure considerations.

In multibranch loop structures, the cross-linking entropy predicts several ways in which entropy could be utilized by ATP engines to run the molecular machinery of some RNA biomolecular complexes. It makes the correct predictions about the direction of folding in a biopolymer. The model is also applicable to protein folding calculations.

We do not expect everyone to agree with us on all the issues presented in this work. Nevertheless, we have benefited from the discussions we have had with a number of people and wish to acknowledge them with all due respect. We graciously thank Prof. M. Doi (Nagoya University) for his generous assistance in affirming the correctness of eqns (B.14) and (B.15) (Appendix B) and Prof. Schuster for contributing to our deeper understanding of the concept of persistence length. Research was supported in part by a fellowship from the Japan Society for the Promotion of Science (JSPS), the Japan International Science & Technology Exchange Center (JISTEC), the Soyou Medical Foundation, and MTB. we also extend our gratitude to

Dr Y. Fujitani who kindly reviewed an early version of this manuscript. We also thank Prof. T. Takagi of the University of Tokyo Institute of Medical Science, and Prof. S. Morishita of the University of Tokyo, Department of Complexity Science and Engineering, Graduate School of Frontier Sciences for their support and encouragement in this research. One of the authors (wkd) would like to express his appreciation for his advisors at University of Tokyo and at San Jose State University for patiently reminding him of the importance of thermodynamics in his work, and a host of other advice he never listened to.

REFERENCES

ACHTER, E. K. & FELSENFELD, G. (1971). The conformation of single-strand polynucleotides in solution: sedimentation studies of apurinic acid. *Biopolymers* **10,** 1625–1634.

BASKARAN, S., STADLER, P. F. & SCHUSTER, P. (1996). Approximate scaling properties of RNA free energy landscapes. *J. theor. Biol.* **181,** 299–310.

BRION, P. & WESTHOF, E. (1997). Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **26,** 113–137.

BURKARD, M. E., TURNER, D. H. & TINOCO, I. (1999). Structure of base pairing involving at least two hydrogen bonds. In: *The RNA World* (Gesteland, R. E., Cech, T. R. & Atkins, J. F., eds), 2nd Edn. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

COMAY, E., NUSSINOV, R. & COMAY, O. (1984). An accelerated algorithm for calculating the secondary structure of single stranded RNAs. *Nucl. Acids Res.* **12,** 53–66.

DAWSON, W. K. & YAMAMOTO, K. (1998). Evidence of structural order in globin intron sequences of messenger RNA. In: *Genome Informatics Series No.* 9 (Miyano, S. & Takagi, T., eds), pp. 49–61. Tokyo: Universal Academy Press, Inc.

DAWSON, W. K. & YAMAMOTO, K. (1999a). Toward full 3D structural investigation of intron splicing in human cytochrome P450 2D6 pre-mRNA. In: *Genome Informatics Series No.* 10 (Asai, K., Miyano, S. & Takagi, T., eds), pp. 336–337. Tokyo: Universal Academy Press.

DAWSON, W. K. & YAMAMOTO, K. (1999b). Mean free energy topology for nucleotide sequences of varying composition based on secondary structure calculations. *J. theor. Biol.* **201,** 113–140.

DAWSON, W. K. & YAMAMOTO, K. (1999c). Evidence of structural information in cytochrome P450 family intron sequences of messenger RNA. In: *RECOMB* 99, *Abstracts.* (Istrail, S., Pevzner, P. & Waterman, M., eds). MA: ACM, Inc.

DEBE, D. A. & GODDARD III, W. A. (1999). First principles prediction of protein folding rates. *J. Mol. Biol.* **294,** 619–625.

DE GENNES, P. G. (1979). *Scaling Concepts in Polymer Physics.* Ithaca: Cornell University Press.

DELISI, C. & CROTHERS, D. M. (1971a). Electrostatic contribution to oligonucleotide transitions. *Biopolymers* **10,** 2323–2343.

DELISI, C. & CROTHERS, M. (1971b). Theory of the influence of oligonucleotide chain conformation on double helix stability. *Biopolymers* **10,** 1809–1827.

DELISI, C. (1973b). Conformational changes in transfer RNA: I. Equilibrium theory. *Biopolymers* **12,** 1713–1728.

DOI, M. & EDWARDS, S. F. (1986). *The Theory of Polymer Dynamics.* Oxford: Clarendon Press.

DRAPER, D. E. (1999). Themes in RNA-protein recognition. *J. Mol. Biol.* **293,** 255–270.

EISENBERG, H. & FELSENFELD, G. (1967). Studies of the temperature-dependent conformation and phase separation of polyriboadenylic acid solutions at neutral pH. *J. Mol. Biol.* **30,** 17–37.

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*, 3rd Edn., Vol. 1. New York: John Wiley & Sons, Inc.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications.* 2nd edn., Vol. 2. New York: John Wiley & Sons, Inc.

FEMINO, A. M., FAY, F. S., FOGARTY, K. & SINGER, R. H. (1998). Visualization of single RNA transcripts *in situ. Science* **280,** 585–590.

FERSHT, A. (1999). *Structure and Mechanism in Protein Science*: *A Guide to Enzyme Catalysis and Protein Folding.* New York: Freeman & Co.

FISHER, M. E. (1966). Effect of excluded volume on phase transitions in biopolymers. *J. Chem. Phys.* **45,** 1469–1473.

FLORY, P. J. (1949). The configuration of real polymer chains. *J. Chem. Phys.* **17,** 303–310.

FLORY, P. J. (1953). *Principles of Polymer Chemistry*, chapter 11. Ithaca: Cornell University Press.

FLORY, P. J. (1956). Theory of elastic mechanisms in fibrous proteins. *Phys. Rev.* **78,** 5222–5235.

FLORY, P. J., MARK, J. E. & ABE, A. (1966). Random-coil configurations of vinyl polymer chains. The influence of stereoregularity on the average dimensions. *J. Amer. Chem. Soc.* **88,** 639–650.

FLORY, P. J. & SEMLYEN, J. A. (1966). Macrocyclization equilibrium constants and the statistical configuration of poly(dimethylsiloxane) chains. *J. Amer. Chem. Soc.* **88,** 3209–3212.

FLORY, P. J. (1976). Statistical thermodynamics of random networks. *Proc. R. Soc. Lond. A* **351,** 351–380.

FONTANA, W. & SCHUSTER, P. (1998). The possible and the attainable in RNA genotype–phenotype mapping. *J. theor. Biol.* **194,** 491–515.

FONTANA, W., STADLER, P. F., BORNBERG-BAUER, E. G., GRIEAMACHER, T., HOFACKER, I. L., TACKER, M., TARAZONA, P., WEINBERGER, E. D. & SCHUSTER, P. (1993). RNA folding and combinatory landscapes. *Phys. Rev. E* **47,** 2083–2099.

FREDERIC, T., RAKEFET, R., CANTOR, C. R. & DELISI, C. (1996). RNA loop structure prediction via bond scaling and relaxation. *Biopolymers* **38,** 769–779.

FREIER, S. M., RYSZARD, K., JAEGER, J. A., NAOKI, S., CARUTHERS, M. H., NEILSON, T. & TURNER, D. H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. U.S.A.* **83,** 9373–9377.

GARNER, E., ROMERO, P., DUNKER, A. K., BROWN, C. & OBRADOVIC, Z. (1999). Predicting binding regions within disordered proteins. In: *Genome Informatics Series No.* 10 (Asai, K., Miyano, S. & Takagi, T., eds), pp. 41–50. Tokyo: Universal Academy Press.

GAUTHERET, D. & CEDERGREN, R. (1993). Modeling the three-dimensional structure of RNA. *FESEB* **7,** 97–105.

GELBIN, A., SCHNEIDER, B., CLOWNEY, L., HSIEH, S.-H., OLSON, W. K. & BERMAN, H. M. (1996). Geometric parameters in nucleic acids: sugar and phosphate constituents. *J. Amer. Chem. Soc.* **118,** 519–529.

GIBBS, J. H. & DIMARZIO, E. A. (1958). Statistical mechanics of helix-coil transitions in biological macromolecules. *J. Chem. Phys.* **30,** 271–282.

GLOTZ, C. & BRIMACOMBE, R. (1980). An experimentally-derived model for the secondary structure of the 16S ribosomal RNA from Escherichia coli. *Nucl. Acids Res.* **8,** 2377–2395.

GROSBERG, A. YU. & KHOKHLOV, A. R. (1994). *Statistical Physics of Macromolecules.* New York: American Institute of Physics (AIP) Press.

GROSBERG, A. YU. & KHOKHLOV, A. R. (1997). *Giant Molecules.* New York: Academic Press.

HAGERMAN, P. J. (1997). Flexibility of RNA. *Annu. Rev. Biophys. Biomol. Struct.* **26,** 139–156.

HERMANN, T. & PATEL, J. D. (1999). Stitching together RNA tertiary architectures. *J. Mol. Biol.* **294,** 829–849.

HOLBROOK, S. R. & KIM, S.-H. (1997). RNA crystallography. *Biopolymers* **44,** 3–21.

INNERS, L. D. & FELSENFELD, G. (1970). Conformation of polyribouridylic acid in solution. *J. Mol. Biol.* **50,** 373–389.

JACOBSON, H. & STOCKMAYER, W. (1950). Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* **18,** 1600–1606.

JAMES, H. M. & GUTH, E. (1947). Theory of increase in rigidity of rubber during cure. *J. Chem. Phys.* **15,** 669–683.

KESKIN, O., JERNIGAN, R. L. & BAHAR, I. (2000). Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophys. J.* **78,** 2093–2106.

LI, XIAOHONG, ROMERO, P., RANI, M., DUNKER, A. K. & OBRADOVIC, Z. (1999). Predicting protein disorder for N-, C- and internal regions. In: *Genome Informatics Series No.* 10 (Asai, K., Miyano, S. & Takagi, T., eds), pp. 30–40. Tokyo: Universal Academy Press.

LUSTIG, B., BAHAR, I. & JERNIGAN, R. L. (1998). RNA bulge entropies in the unbound state correlate with peptide binding strengths for HIV-1 and BIV TAR RNA because of improved conformational access. *Nucl. Acids Res.* **26,** 5212–5217.

LYNGSØ, R. B. (1999). Computational Biology. Dissertation, University of Aarhus.

MATHEWS, D. H., SABINA, J., ZUKER, M. & TURNER, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288,** 911–940.

MIRONOV, A. A., DYAKONOVA, L. P. & KISTER, A. E. (1985). A kinetic approach to the prediction of RNA secondary structures. *J. Biomol. Struct. Dyn.* **2,** 953–962.

MCCASKILL, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29,** 1105–1119.

MUELLER, H., BUTT, H.-J. & BAMBERG, E. (1999). Force measurements on myelin basic protein adsorbed to mica and lipid bilayer surfaces done with the atomic force microscope. *Biophys. J.* **76,** 1072–1079.

MUELLER, F., SOMMER, I., BARANOV, P., MATADEEN, R., STOLDT, M., WOEHNERT, J., GOERLACH, M., VAN HEEL, M. & BRIMACOMBE, R. (2000). The 3D arrangement of the 23 S and 5 S rRNA in the *Escherichia coli* 50 S Ribosomal Subunit based on a cryo-electron microscopic reconstruction at 7. 5 Å resolution. *J. Mol. Biol.* **298,** 35–59.

NUSSINOV, R. & JACOBSON, A. B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl Acad. Sci. U.S.A.* **77,** 6309–6313.

NUSSINOV, R., TINOCO JR., I. & JACOBSON, A. B. (1982). Secondary structure for the complete simian virus 40 late precursor mRNA. *Nucl. Acids. Res.* **10,** 351–363.

ORR, W. J. (1947). *Trans. Faraday Soc.* **43,** 12 (Cited in deGennes, 1979 p. 31.)

PAN, J. & WOODSON, S. A. (1999). The effect of long-range loop-loop interactions on folding of the tetrahymena self-splicing RNA. *J. Mol. Biol.* **294,** 955–965.

PLISCHKE, M. & BERGERSEN, B. (1994). In: *Equilibrium Statistical Physics*, 2nd Edn. Englewood Cliffs, NJ: Prentice Hall.

POLAND, D. C. & SCHERAGA, H. A. (1965). Comparison of theories of the helix-coil transition in polypeptides. *J. Chem. Phys.* **43,** 2071–2074.

POLAND, D. C. & SCHERAGA, H. A. (1966). Occurrence of a phase transition in nucleic acid models. *J. Chem. Phys.* **45,** 1464–1469.

REIF, M., CLAUSEN-SCHAUMANN, H. & GAUB, H. E. (1999). Sequence-dependent mechanics of single DNA molecules. *Nat. Struct. Biol.* **6,** 346–349.

REIF, M., FERNANDEZ, J. M. & GAUB, H. E. (1998). Elastically coupled two-level systems as a model for biopolymer extensibility. *Phys. Rev. Lett.* **81,** 4764–4767.

SCHEFFLER, I. E., ELSON, I. L. & BALDWIN, R. L. (1970). Helix formation by d(TA) oligomers. II. Analysis of the helix-coil transitions of linear and circular oligomers. *J. Mol. Biol.* **48,** 145–171.

SCHNEIDER, T. D. (1990). Theory of molecular machines. I. Channel capacity of molecular machines. *J. theor. Biol.* **148,** 83–123.

SEARLE, M. S. & WILLIAMS, D. H. (1993). On the stability of nucleic acid structures in solution: enthalpy–entropy compensations, internal rotations and reversibility. *Nucl. Acids Res.* **21,** 2051–2056.

SEARS, F. W. & SALINGER, G. L. (1986). *Thermodynamics Kinetic Theory, and Statistical Thermodynamics*, 3rd Edn. Menlo Park: Addison-Wesley.

SMITH, S. B., CUI, Y. & BUSTAMANTE, C. (1996). Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science* **271,** 795–799.

SMITH, S. B., FINZI, L. & BUSTAMANTE, C. (1992). Direct mechanical measurement of the elasticity of single DNA molecules by using magnetic beads. *Science* **258,** 1122–1126.

STUDNICKA, G. M., RAHN, G. M., CUMMMINGS, I. W. & SALSER, W. A. (1978). Computer methods for predicting the secondary structure of single-stranded RNA. *Nucl. Acids Res.* **5,** 3365–3387.

THIRUMALAI, D. (1998). Native secondary structure formation in RNA may be a slave to tertiary folding. *Proc. Natl Acad. Sci.* U. S. A. **95,** 11506–11508.

TINOCO JR., I. & BUSTAMANTE, C. (1999). How RNA folds. *J. Mol. Biol.* **293,** 271–281.

TURNER, D. H., SUGIMOTO, N. & FREIER, S. M. (1988). RNA structure prediction. *Annu. Rev. Biophys. Biophys. Chem.* **17,** 167–192.

WILLIAMS JR., A. L., TINOCO JR., I. (1986). A dynamic programming algorithm for finding alternate RNA secondary structures. *Nucl. Acids Res.* **14,** 299–315.

WIMBERLY, B. T., BRODERSEN, D. E., CLEMONS, W. M., MORGAN-WARREN, R. J., CARTER, A. P., VONREIN, C., HARTSCH, T. & RAMAKRISHNAN, V. (2000). Structure of the 30S ribosomal subunit. *Nature* **407,** 327–339.

WU, M. & TINOCO JR., I. (1998). RNA folding causes secondary structure rearrangement. *Proc. Natl Acad. Sci.* U. S. A. **95,** 11555–11560.

WYATT, J. R. & TINOCO JR., I. (1993). RNA structure elements and RNA function. In: *The RNA World*, (Gesteland, R. F., and Atkins, J. F. eds) pp. 465–596. Cold Springs Harbor: Cold Springs Harbor Laboratory Press.

YAMAMOTO, K., KITAMURA, Y. & YOSHIKURA, H. (1984). Computation of statistical secondary structure of nucleic acids. *Nucl. Acids Res.* **12,** 335–346.

YOUHEI, F. & YAMAMOTO, K. (1994). Influence of slight sequence changes on the free energy of the single stranded ribonucleic acid molecule. *J. theor. Biol.* **171,** 151–161.

ZUKER, M. & STIEGLER, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.* **9,** 133–148.

ZUKER, A. M., MATHEWS, D. H. & TURNER, D. H. (1998). Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: *RNA Biochemistry and Biotechnology*, (Barciszewski, J. & Clark, B. F. C., eds). NATO ASI Series, Dordrecht: Kluwer Academic Publishers. (Available at web site "http://bioinfo.math.rpi.edu/~zukerm/seqanal/".)

APPENDIX A

Fundamental Thermodynamic Definitions for the GPC

For a system obeying the properties of a Gaussian polymer chain (GPC) (Doi & Edwards, 1986; Grosberg & Khokhlov, 1997; Grosberg & Khokhlov, 1994), we define the following set of state variables: N the number of "mers" (or links in the polymer chain), T the temperature, f the effective force acting between the ends of the GPC, and r the root mean square (rms) displacement between the ends of the GPC [see Fig. 5(a)]. In direct correspondence to the kinetic theory of the ideal gas (Sears & Salinger, 1986), r relates to the volume (V) and f is analogous to pressure (p).

The interactions of a GPC assembly can be considered a *reversible* process because there are no dissipative interactions in the idealized model. Consequently,

$$T\,\mathrm{d}S = \mathrm{d}U + f\,\mathrm{d}r, \tag{A.1}$$

where U is the internal energy, S is the entropy, and we follow the convention where the work done *by* the system is positive (Sears & Salinger, 1986). The Helmholtz free energy ($\mathscr{F}$) is

$$\mathrm{d}\mathscr{F} = \mathrm{d}U - \mathrm{d}(TS) = -f\,\mathrm{d}r - S\,\mathrm{d}T. \tag{A.2}$$

Moreover, because the process is reversible in the model, the equations are exact

$$-\frac{\partial^2 \mathscr{F}}{\partial r \partial T} = \left(\frac{\partial S}{\partial r}\right)_T = \left(\frac{\partial f}{\partial T}\right)_r, \tag{A.3}$$

where $(\,)_r$ and $(\,)_T$ refer to the variable held constant in the measurement process.

By definition, a system of distinguishable particles (i.e. obeying Maxwell–Boltzmann statistics) must show a chemical potential of the form (Sears & Salinger, 1986)

$$\mu \equiv -k_B T \ln Z = -T\left(\frac{\partial S}{\partial N}\right)_{U,r}, \tag{A.4}$$

where μ is the chemical potential and Z is the partition function. Accordingly, the Helmholtz free energy is

$$\mathscr{F} = -Nk_B T \ln Z = N\mu. \tag{A.5}$$

For any true Maxwell–Boltzmann (MB) particle, the chemical potential is an intrinsic variable which is independent of N. Hence, linear increases in the number of segments (N) of the GPC must change the free energy ($\mathscr{F}$) in a *linear* fashion.

APPENDIX B

Derivation of the GPC Model

The first solution to the GPC problem was originally attributed to Orr (1947) in de Gennes' (1979) classic work. The GPC can be understood, studied and tested by a host of techniques from a simple random walk model (Feller, 1968, 1971), to a random flight or flexible chain model (Doi & Edwards, 1986; Plischke & Bergersen, 1994). These same conclusions can also be arrived at from an evaluation of the vibrational modes of an N-dimensional system (Schneider, 1991). It is a consequence of the central limit theorem (Feller, 1968, 1971; Grosberg & Khokhlov, 1994; Baskaran *et al.*, 1996; Fontana *et al.*, 1993). Our

approach follows primarily from Flory (1949) and James & Guth (1947) with emphasis on the assumptions underlying the GPC model.

In all cases, the statistical average of the mean square displacement between the ends of the GPC (which we define as R) is shown to be $R = b\sqrt{2N/3}$ (in 3-D space, Doi & Edwards, 1986), where N is the number of segments in a polymer chain and b is the length of each segment. Each segment along the GPC represents the position of an individual "link", where the total number of "links"†† is equal to N, and the total (stretched out) length of the polymer is Nb. The parameter b is also known as the "persistence length" (Doi, 1996; Doi & Edwards, 1986; Grosberg & Khokhlov, 1997; Grosberg & Khokhlov, 1994; Plischke & Bergersen, 1994).

The entropy of a group of MB-particles in a specified macrostate l is estimated from the statistical mechanics relationship (Sears & Salinger, 1986)

$$S_l = k_B \ln(\Omega_l), \tag{B.1}$$

where k_B is the Boltzmann constant, and Ω_l specifies the thermodynamic probability of the ensemble for a given temperature, number of "links" (assemblies) in the ensemble, and the distribution of energy microstates.

Since the "links" in the GPC are distinguishable, the GPC must follow MB statistics. The thermodynamic probability of a macrostate obeying MB statistics is defined as (Flory, 1949; Sears & Salinger, 1986)

$$\Omega \equiv N! \prod_i \frac{(g_i)^{N_i}}{N_i!}, \tag{B.2}$$

where the product occurs over all possible microstates (i) of degeneracy (g_i), N_i is the number configurations with a displacement r_i (the rms displacement of *each segment* or "link") which occupy microstate i at a specified temperature T.

For each segment on the GPC, the energy distribution of a segment (i) is expressed by an end to-end displacement (ρ_i) which connects segments i to its nearest-neighboring links (on each side) (Flory, 1949). Since each segment is assumed to be non-interacting, there are no restrictions on the orientation of a given segment or the configuration of the segment's nearest-neighboring links. The statistical meaning of (ρ_i) amounts to counting of all the links (N_i) with such a displacement (ρ_i) and considering how such links could be arranged with respect to each other (i.e. $g_i^{N_i}/N_i!$).

For a flexible chain (or freely joined chain) (Grosberg & Khokhlov, 1997), the displacement distribution for a given segment is expressed by the following probability density distribution function (deGennes, 1979; Doi & Edwards, 1986; Flory, 1949)

$$p_D(\rho) = \left(\frac{\beta}{\pi}\right)^{3/2} \exp(-\beta\rho^2), \tag{B.3}$$

where $\rho = \sqrt{x^2 + y^2 + z^2}$, $\beta = 3/(2b^2)$, and $0 \leqslant \rho < +\infty$ (with variance of order $\rho^2 \sim b^2$).

The probability density function (B.3) is evaluated over a differential volume $\Delta V = 4\pi\rho^2\Delta\rho$, (where V is the volume) (Feller, 1968, 1971). Further, we assume a constant temperature; hence, $\Delta U = 0$, $Z \propto p$, and $\Omega(\rho) \propto p^N(\rho)$ (which follows from the definitions; Sears & Salinger, 1986). The probability of a given microstate configuration ρ_i is (Flory, 1949)

$$\begin{aligned} p(\rho_i) &= p_D(\rho_i) 4\pi\rho_i^2 \Delta\rho_i \\ &= \left(\frac{\beta}{\pi}\right)^{3/2} 4\pi\rho_i^2 \exp(-\beta\rho_i^2)\Delta\rho_i; \\ & \qquad 0 \leqslant \rho_i < +\infty. \end{aligned} \tag{B.4}$$

Equation (34) expresses a macrostate "snapshot" in which each set of configurations (microstates) is represented by the index i (with degeneracy g_i). Let ω_0 be the total number of possible configurations available to a segment

†† A "link" is not necessarily equivalent to the monomer (or "mer") of a real-polymer chain. Indeed, there can be far more "links" than "mers" and vice versa. For classical polymers such as vinyl and its derivatives (Flory *et al*., 1966a), the links are always longer than the spacing between the "mers". However, this does not always have to be the case. For example, an *intra*-polymer chain cross-link formed in such products as rubber will show behavior resembling Fig. 3 of Section 2.2. The assembly of "links" is also called an ensemble.

and let Ω_0 define a reference state where $\Omega_0 = \prod_i \omega_0^{N_i} = \omega_0^N$. Then, a particular microstate is approximated by making the following substitution $g_i \approx \omega_0 p(\rho_i)$. After substitution into eqn (B.2)

$$\Omega \approx N! \prod_i \frac{(p(\rho_i)\omega_0)^{N_i}}{N_i!} \tag{B.5}$$

and, applying Sterling's formula,‡‡ we obtain

$$\ln\left(\frac{\Omega}{\Omega_0}\right) = N \ln N - \sum_i N_i \ln(N_i) + \sum_i N_i \ln(p(\rho_i)). \tag{B.6}$$

We might predict that the largest number of configurations occur at $\rho_i = b$ and that very few configurations occur at $\rho_i = 0$ or $\rho_i \to \infty$. Since we are interested in finding the rms response *by* the GPC for a given displacement ρ, this corresponds to a macrostate in which *all* the microstates of the GPC exist with the same end-to-end displacement ρ. Therefore, we fix $\rho_i = \rho$, and $N_i = N$.

Two points are important here. First, although eqn (B.4) is a continuous distribution, it is actually an approximation of a multinomial expansion (Feller, 1968, 1971). Hence, $\Delta\rho$ [eqn (B.4)] is actually a discrete value. Second, we assume that $\rho_i = b$, because the function [eqn (B.4)] is sharply peaked at $\rho = b$ and (under the conditions of thermodynamic equilibrium), the majority of "snapshots" of the system will tend to reveal a chain in which $\rho = b$.

In other words, since MB statistics do not place any restrictions on the number of states that are occupied (N_i), we have simply *asked* "what is the thermodynamic probability that all ρ_i will be in a macrostate expressed by b"? The maximum for that thermodynamic state is then $N \ln(p(\rho))$. For states $\rho \neq b$, we can establish an upper bound in which $\forall\rho$, $N \ln(p(\rho)) \geqslant N p(\rho) \ln(p(\rho))$. In general, these values of $\rho \neq b$ are *very* improbable (collectively speaking) unless the polymer chain is forced into such a configuration by some interaction such as stacking interactions (for nucleotides) or hydrophobic interactions (for peptides). Undaunted, we proceed forward justifying ourselves via appeals to the upper bound and noting that (for $\rho \neq b$), chemical bonds must be introduced to restrain the value of ρ

$$\ln\left(\frac{\Omega}{\Omega_0}\right) = N \ln\{C'(\rho/b)^2 \exp(-\beta\rho^2)\}, \tag{B.7}$$

where $C' = (4/\sqrt{\pi})\,(3/2)^{3/2}\,(\Delta\rho/b)$.

This shows that $\Omega(\rho) \propto p^N(\rho)$. The N appears because the individual links of the GPC each contribute independently to the FE.

From eqns (B.4) and (B.7), the entropy for a specified ρ is

$$\Delta S = S - S_0 = k_B \ln\left(\frac{\Omega}{\Omega_0}\right) = N k_B \{\ln(C') + 2\ln(\rho/b) - \beta\rho^2\}, \tag{B.8}$$

where $S_0 = k_B \ln \Omega_0 = N k_B \ln \omega_0$. The entropy [eqn (B.8)] has a global maximum at $\rho = 1/\sqrt{\beta} = b\sqrt{2/3}$. Furthermore, the response of the system (due to a given rms state ρ) is

$$f = T\left(\frac{\partial S}{\partial \rho}\right)_T = 2N k_B T\left(\frac{1}{\rho} - \beta\rho\right) \tag{B.9}$$

which shows the correct response (by the system) to fluctuations in ρ and a restoring force centered around $\rho = b\sqrt{2/3}$. Moreover, both the entropy and the response have the expected relationship that is materially dependent on the number of "links" in the GPC where the harmonic oscillator component of the effective spring constant is $2Nk_BT\beta$. It can also be seen that both eqns (B.8) and (B.9) satisfy eqns (A.4) and (A.5) for a system obeying MB statistics. (The partition function $Z(=p)$ is independent of N.)

However, eqns (B.8) and (B.9) are not in a usable form because ρ refers to the length of an *individual* "link". We must project this N-dimensional system into 3-D (Schneider, 1990). To do this, we observe (as a consequence of the central limit theorem (Feller, 1968, 1971)) that the

‡‡ A more accurate estimation is $\ln(N!) = \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln(N) + N\ln(N) - N$, where we have assumed that $N \ln N \gg N$.

partition function (Z) scales with the number of links

$$Z_1 = \left\{\frac{4}{\sqrt{\pi}}\left(\frac{3}{2(1)}\right)^{3/2}\left(\frac{\rho}{b}\right)^2 \exp(\beta\rho^2)\left(\frac{\Delta\rho}{b}\right)\right\}. \tag{B.10}$$

If we group the system into sets of dimers, then we have $N/2$ dimers and

$$Z_2 = \left\{\frac{4}{\sqrt{\pi}}\left(\frac{3}{2(2)}\right)^{3/2}\left(\frac{\sqrt{2}\rho}{b}\right)^2 \exp\left(\frac{\beta}{2}2\rho^2\right) \times\left(\frac{\sqrt{2}\Delta\rho}{b}\right)\right\}^{1/2}. \tag{B.11}$$

Finally, if we group all the links together, then we obtain

$$Z_N = \left\{\frac{4}{\sqrt{\pi}}\left(\frac{3}{2(N)}\right)^{3/2}\left(\frac{\sqrt{N}\rho}{b}\right)^2 \exp\left(\frac{\beta}{N}N\rho^2\right) \times\left(\frac{\sqrt{N}\Delta\rho}{b}\right)\right\}^{1/N}. \tag{B.12}$$

Hence, $Z_N^N = Z_1$ and an individual polymer chain is also understood to be a single MB-particle when viewed from this angle.

We now make the following substitutions in eqn (B.12): $r = \sqrt{N}\rho$, $\alpha = \beta/N$ and $\Delta r = \sqrt{N}\Delta\rho = b$. We use $\Delta r \sim b$ because the shortest possible discrete length Δr must be of the order of the length of an individual segment of the GPC. Moreover, this must be introduced as a semiclassical approximation of the quantum mechanical effects of the GPC (although the MB distribution assumes a truly classical system). Finally, the GPC is based on a lattice model, which has discrete length scales of order b. We now obtain

$$Z_N = \left\{\frac{4}{\sqrt{\pi}}\left(\frac{3}{2N}\right)^{3/2}\left(\frac{r}{b}\right)^2 \mathrm{e}^{-\alpha r^2}\right\}^{1/N} \tag{B.13}$$

and substituting into eqn (B.8), we obtain

$$\Delta S = k_B\{\ln(C_N) + 2\ln(r/b) - \alpha r^2\}, \tag{B.14}$$

where $C_N = (4/\sqrt{\pi})(3/2N)^{3/2}$, and

$$f = 2k_BT\left(\frac{1}{r} - \alpha r\right) \tag{B.15}$$

which is what we wanted to show.

In this final form, it is somewhat pertinent to consider that because the GPC model assumes independent—non-interacting—links, some rather amusing structures can be imagined. In particular, if the number of links in the chain is odd, the chain could fold back and forth on itself forming an rms displacement of 0, and thus occupying the same space $(N-1)/2$ times or $(N+1)/2$ times (depending on the end point). Likewise, if N is even, then the same folding back and forth will yield a chain with displacement b and occupation of the same lattice site $N/2$ times. Although such unphysical structures are extremely improbable, they remain possible in the GPC formulation. This is the primary reason as to why considering the excluded volume is important in improving such calculations (Fisher, 1966).

APPENDIX C

Excluded Volume Considerations

In the derivation of eqn (B.14), we ignored the dimensionality of the GPC in the interest of expediency. In the literature, the logarithmic term in eqn (B.14) has usually been derived from the volume dependence (Flory, 1956, 1976; Jacobson & Stockmayer, 1950). This change in dimensionality of the system introduces a pre-factor (γ) on the logarithmic term such that $\gamma \sim D/2$, where D reflects the dimensions of the system (Fisher, 1966; Poland & Scheraga, 1966). Hence, for the ideal GPC in eqns (B.14) and (B.15), $\gamma = 1$ which is only true for two dimensions.

This dimensionality is further complicated because the GPC is a fictional construct which permits the simultaneous occupation of two or more segments in the same space (although such configurations are extremely rare statistically speaking). A self-avoiding random walk (Feller, 1971) is a more accurate description of a polymer chain. Real polymers occupy space and different monomers cannot simultaneously occupy the same exact place. Hence, in a random folding of a real polymer, not all possible configurations

available to a GPC are allowed, and part of the real polymer must be "excluded" by other parts of the sequence. The net result is that the equilibrium separation distance in eqn (B.14) becomes larger. The excluded volume sets upper and lower bounds on γ such that $D/2 \leqslant \gamma < (D+1)/2$ (Fisher, 1966).

To approximate the excluded volume in the GPC formulation, we introduce the following probability density function (p.d.f.):

$$p(\rho) = C_N^{\gamma}\left(\frac{\rho}{b}\right)^{2\gamma}\exp(-\beta\rho^2)\left(\frac{\Delta\rho}{b}\right),$$

$$0 \leqslant \rho < +\infty, \qquad \text{(C.1)}$$

where $\gamma > 0$, $C_N^{\gamma}/b^{2\gamma+1}$ (with $N = 1$) is the normalization constant for the p.d.f. (analogous to C' in eqn (B.7))

$$C_N^{\gamma} = \frac{2}{\Gamma(\gamma + 1/2)}\left(\frac{3}{2N}\right)^{\gamma+1/2}$$

and $\Gamma(\zeta)$ is the gamma function with argument $\zeta = \gamma + 1/2$. For $\gamma = 1$, we obtain the same value for C_N^{γ} as eqn (B.14).

The central limit theorem applies to the evaluation of the p.d.f. Inserting eqn (C.1) into eqn (B.5) and solving as before using the steps in eqns (B.6)–(B.13), we obtain

$$\Delta S = k_B\{\ln(C_N^{\gamma}) + 2\gamma\ln(r/b) - \alpha r^2\} \qquad \text{(C.2)}$$

$$f = T\left(\frac{\partial S}{\partial r}\right)_T = 2k_BT\left(\frac{\gamma}{r} - \alpha r\right) \qquad \text{(C.3)}$$

hence, the equilibrium separation is $R = \sqrt{\gamma/\alpha}$.

In the literature, γ is known to be approximately 1.4 in 2-D and 1.75 in 3-D dimensions (Fisher, 1966). Therefore, the root mean square displacement will be increased, and the logarithmic contribution to the cross-linking entropic effect will be larger when the excluded volume conditions are taken into account.

APPENDIX D

The GPC in Traditional Polymer Studies

Cross-linking in traditional polymer physics involves the linking *between* individual polymer chains in a random network of chemical bonds.

In the case of vulcanized rubber, the elasticity of the rubber is also related to the basic properties of the single polymer chains; however, because of the cross-linking, the magnitude of the chain interactions are proportional to the number of cross-links in the chain. The greater the number of the cross-links, the greater the responsive force of the rubber (Flory, 1953). (We assume that the density of cross-links is small compared with the segment lengths of the polymer chains that separate each cross-link.)

Ever since (James & Guth, 1947) ignored the volume dependence in their calculation of the elasticity of rubber, the logarithmic contribution to the entropy has typically been neglected in calculations pertinent to traditional polymer cross-linking. To understand as to why this can be done, three main points need to be considered: (1) the studies are only concerned with the uniaxial stretching of a polymer, (2) typical models of polymers have expressed the cross-linking between different polymers as a grid (sometimes referred to as a "phantom network" (Flory, 1976)), and (3) in cross-linked polymers, the interactions occur between *different* polymer chains (interchain), not the same polymer chain as is considered in this work (intrachain).

Since the relationship between cross-linked polymer chains resembles a grid and the only interaction considered is uniaxial stretching, the volume (V) of the polymer can be expressed as $V = l_x l_y l_z$. In a phantom network, the interchain cross-links can be approximated as springs which join the points on the grid (with an equal average spacing). If the phantom network is stretched along l_z, then a new volume (V') is reached with $V' = l_x' l_y' l_z'$. The volume-related entropic change will be $\ln(V'/V)$. However, in a grid like structure, there is no change in the volume ($V = V'$) because when l_z is stretched to l_z', the cross-sectional area will almost exactly compensate for this change by a proportional decrease in $l_x \times l_y$ such that the $\Delta V \sim 0$. As a result, the logarithmic term will show no significant contribution to the stretching of a polymer chain.

There are some rather special cases where this approximation scheme can break down. The most notable example is the "superball": a solid

rubber ball with a high elastic coefficient that serves as a toy for children (particularly because of its impressive bounce). The bounce of a superball does involve a change in volume: due to the compression on the face of a spherically symmetric elastic surface. The uncompressed parts of the rubber superball remain spherical while the face where compression is applied (the floor, the wall, etc.) compresses along a flat planar surface, $V \neq V'$. Hence, the cross-linked rubber "superball" shows a volume-dependent effect.

Another example is when this problem is applied to the folding of RNA. The elasticity of rubber stems from cross-linking between *different* polymer chains, whereas the entropic interactions of the nucleic acids discussed here occur within the *same* chain. This latter effect is usually ignored in traditional theories of rubber elasticity because it makes no significant contribution to the elasticity (Flory, 1953). In addition, because the chains of a nucleic acid are more intricately connected, the specific arrangement of the sequences is more critical to the estimation of the entropy than is the case of a complex network of cross-linked poly-isoprene units.