J. theor. Biol. (2001) **213,** 387–412
doi:10.1006/jtbi.2001.2437, 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: II

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, International Medical Center Japan, 1-21-1 Toyama, Shinjuku-ku, Tokyo* 162-8640, *Japan*

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

In Part I, cross-linking entropy (CLE) was proposed as a mechanism that limits the size of functional domains of RNA. To test this hypothesis, the theory is developed into an RNA secondary structure prediction filter which is applied to nearest-neighbor secondary structure (NNSS) algorithms that utilize a free energy (FE) minimization strategy. (The NNSS strategies are also referred to as the dynamic programming algorithm in the literature.) The cross-linking entropy for RNA is derived from a generalized Gaussian polymer chain model where the entropic contributions caused by the formation of base pairs (stacking) in RNA are analysed globally. Local entropic contributions are associated with the freezing out of degrees of freedom in the links. Both global and local entropic effects are strongly influenced by the persistence length. The cross-linking entropy provides a physical origin for the size of functional domains in long nucleic acid sequences and may go further to explain as to why the majority of the domain regions in typical sequences tend to be less than 600 nucleotides in length. In addition, improvements were observed in the "best guess" predictive capacity over NNSS prediction strategies. The thermodynamic distribution is more representative of the expected structures and is strongly governed by such physical parameters as the persistence length and the excluded volume. The CLE appears to generalize the tabulated penalties used in NNSS algorithms. The principal parameter influencing this entropy is the persistence length. The model is shown to accomodate a variable persistence length and is capable of describing the folding dynamics of RNA. A two-state kinetic model based on the CLE principle is used to help elucidate the folding kinetics of functional domains in the group I introns.

1. Introduction

The final goal of most theoretical biology devoted to structure and function is to gain enough understanding to be able to ask "how reasonable are the results I have just computed or measured". Science is not so much about proclaiming *the* correct answer, rather, it is about having a strong intuition as to whether claims purported by some "great machine" even make sense. After all, a 5-year old could have played with the settings on that "great machine" yesterday. What means are at my disposal to tip me off that something is wrong? Often, it is not easy to find that knowledge and, for us at least, we are grateful when we stumble up on it. In this second part, we present what we think provides some of that intuitive framework on understanding the folding of RNA. It is assumed that the reader is familiar

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

0022–5193/01/230387 + 26 $35.00/0

with the basic content and definitions presented in Part I of this series (Dawson *et al.*, 2001).

Currently, with the possible exception of short sequences such as tRNA, our actual knowledge of the correct RNA secondary structure has come exclusively from experimental techniques. Long nucleic acid structures such as 16S and 23S ribosomal RNA (rRNA) serve an essential function in the biological repertoire of molecular machinery. Most of the knowledge about the structure of rRNA has come from using RNase techniques, comparative sequence analysis, and mutational analysis (Glotz & Brimacombe, 1980; Woese *et al.*, 1980). Recently, other techniques have been brought to bear on the 3D structural problems of these molecules [for recent developments in experimental techniques in this area, see (Mueller *et al.*, 2000; Wimberly *et al.*, 2000)]. All of these advanced approaches have relied heavily on the foundation of experimentally obtained secondary structure information.

RNA secondary structure prediction using free energy (FE) minimization strategies have been a helpful aid in studying short sequences. Some of the earliest attempts began with Tinoco's group (Tinoco *et al.*, 1971) and Pipas & McMohen (1975). A number of efficient secondary structure algorithms were introduced in the late 1970s and early 1980s (Zuker & Stiegler, 1981; Studnicka *et al.*, 1978; Nussinov & Jacobson, 1980) along with analysis strategies (Yamamoto *et al.*, 1984, 1986). From this approach, two major programs have evolved: one developed by the Tinoco group (Nussinov & Jacobson 1980; Williams & Tinoco, 1986), and the other by the Zuker group (Jaeger *et al.*, 1989, 1990; Zuker, 1989). Some more recent strategies have involved the application of parallel computing techniques (Nakaya *et al.*, 1996), genetic algorithm approaches (Chen *et al.*, 2000; Gultyaev *et al.*, 1995; Notredame *et al.*, 1997; von Batenburg *et al.*, 1995), and partition function evaluation schemes in which the thermodynamic probability of a given secondary structure is obtained (Hofacker *et al.*, 1994a, b; McCaskill, 1990). Further developments are now progressing in the application of kinetics (see Hofacker, 1998 and references therein for details).

In all these strategies, the common feature is that they assume that the FE can be completely evaluated based only upon the information in the immediate vicinity of a particular base pair (BP). Thus local context is considered, but a global context is ignored. We have therefore defined these methods as nearest-neighbor secondary structure (NNSS) strategies.

Early versions of the RNA NNSS algorithms lacked information about the stacking free energy parameters for nucleic acid base pairing and entropic free energy contributions of such sequences. As these algorithms advanced, thermodynamic studies of the stacking free energy for AU-, GC- and GU-hybridized pairs were gradually tabulated. A certain amount of work was also done to evaluate the entropy of closed loops and internal loops in the late 1960s and early 1970s (Gralla & Crothers, 1973a, b; Scheffler *et al.*, 1970). One of the first comprehensive lists of FE parameters appeared in the work of Salser (1977). Further work by Freier and Turner have added considerable improvements and refinements to the nearest-neighbor interactions model and the role of entropy (Freier *et al.*, 1986; Turner *et al.*, 1988). This large body of work has ultimately become an exhaustive set of lookup tables of stacking free energies for all conceivable nearest-neighboring nucleic acid combinations for both RNA and DNA structures. (To learn more about recent developments in this area, refer to the reference section of Mathews *et al.*, 1999.) To measure these thermodynamic parameters, all such approaches have utilized mostly short double-stranded oligonucleotide sequences (for a recent discussion, see SantaLucia & Turner, 1998).

Currently, MFOLD (Zuker, 1989) and the Vienna package (Hofacker *et al.*, 1994b; McCaskill, 1990) are the two most frequently cited NNSS strategies. Both methods rely on thermodynamics to discern the optimal structure of a given sequence and use some version of the well-known Turner rules (Turner *et al.*, 1988; Zuker, 1998). The major difference between MFOLD and the Vienna package is that MFOLD evaluates the majority of possible foldings of a sequence, and these structures are later sorted in a post processing approach to establish the minimum FE and prioritize the list of suboptimal structures, whereas the Vienna package starts by evaluating the partition function, and then determines the most

thermodynamically probable structure. If all the structures from the first part of the MFOLD calculation are requested, a reasonable (although perhaps incomplete (Wuchty *et al*., 1999) partition function can also be reconstructed; however, because the objectives are different in these packages, the calculation time is clearly more expensive in the case of MFOLD (for partition function evaluation). Both approaches appear to yield rather similar secondary structure predictions, as would be expected if the rules are essentially the same.

As already pointed out in Part I, these NNSS strategies have been very successful at tackling relatively short sequences or longer sequences composed of short domains; however, when the domain size becomes very long, the *likelihood* of grossly erroneous predictions increases rapidly.

It has often been argued that the reason for this failure is because of the non-equilibrium conditions in which RNA polymerase builds a sequence from the 5′ to the 3′ ends (Tinoco & Bustamante, 1999; Nussinov, *et al*., 1982; Brion *et al*., 1997; Wu *et al*., 1998; Thirumalai, 1998). This may bias the hierarchy and stability of the resulting structure due to the kinetics of the RNA during the folding process (Boyle *et al*., 1980; Brion & Westhof, 1997; Nussinov *et al*., 1982; Gultyaev *et al*., 1995; von Batenburg *et al*., 1995; Williams *et al*., 1986). Such questions have emerged a number of times in relation to tRNA folding (Boyle *et al*., 1980; Fresco *et al*., 1966; Mironov *et al*., 1985). However, for group I catalytic introns, where there is some indication of mis-folding due to mutations, the native state is eventually attained in many cases (at least when sufficient Mg^{2+} is used, Pan & Woodson, 1999). A similar case can be made for tRNA (Brion & Westhof, 1997). This suggests that the final structure is thermodynamically stable or very close to a state of thermodynamic equilibrium. Moreover, whereas non-equilibrium conditions are likely to explain *some* of the differences in the predicted secondary structure, it would still be desirable to have some way to *show* that this is so.

At the heart of this issue is understanding how functional domains may have evolved and, perhaps even more important, how they currently achieve and maintain their function. The RNA world models of evolution typically involve the accretion of separate sequences such as combinations of tRNA-like structures (Tomizawa, 1993; Turner & Bevilacqua, 1993; Volkenstein, 1994; Wyatt & Tinoco, 1993; Noller, 1999). However, in the FE-based NNSS prediction strategies, there are no intrinsic limits on the size of the domains that can form in RNA sequences. Domain sizes for the "optimal structures" often start at the 5′-3′ limits of the sequence length that is calculated. Adding more sequence tends to extend that domain size accordingly, sometimes completely obliterating any evidence of the previously predicted structure in the process. A model that produces inconsistent predictions with respect to input sequence length seriously hampers our understanding of RNA design and function. Nature often has limits, and it is desirable to understand what those limits are. The lack of any sequence length dependence on the domain size found in RNA secondary structure predictions already suggests that something fundamental has been missed.

In this work, we advance the cross-linking entropy (CLE) model and apply it in the form of a "filter" for RNA secondary structure predictions. We show that the appearance and limits on the size of functional domains can be explained primarily by entropic effects that are not so far away from equilibrium thermodynamic conditions, and in many cases, this approach can even improve the "best guess" predictions for the optimal secondary structure and the thermodynamic distribution of neighboring suboptimal structures. Finally, the approach has sufficient simplicity to permit us to develop some basic intuition about RNA structure.

2. Theory and Calculation Methods

The Gaussian polymer chain (GPC) model and its connection to RNA secondary structure along with a host of terminology and concepts have already been discussed at length in Part I and will not be repeated here. Here and in the sections that follow, we will expand upon this model assuming that the reader is already acquainted with Part I or has it available for reference. To indicate references to equations that were presented in Part I, the following shorthand is used: eqn (I-V) where V denotes the corresponding equation.

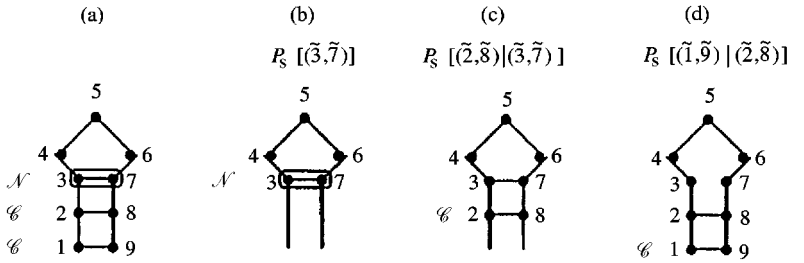

FIG. 1. A diagram showing how the matrix elements of P_S are assigned in a short hairpin loop (Section 2.1): (a) The assignment of the nucleation site ($\mathcal{N}$) and correlation sites ($\mathscr{C}$) for the hairpin as a whole. (a different BP formation order is also allowed: this figure is only meant to *illustrate* the process), (b) Construction of the nucleation matrix elements ($P_S[(\tilde{3}, \tilde{7})]$). (c) Construction of the first conditional matrix with correlation caused by the formation of the cross-link at $(\tilde{3}, \tilde{7})$. (d) Construction of the second condition matrix with correlation caused by the cross-link at $(\tilde{2}, \tilde{8})$. Note that the Markov chain does not call further back than the previous step. This is why the cross-link at $(\tilde{3}, \tilde{7})$ is ignored in Fig. 1(d). Neither can a Markov system "see ahead". This is why some sites are unmarked in Fig. 1(b) and (c).

For section and figure references to Part I, similar shorthand is used: e.g. Section I-V and Fig. I-V.§

2.1. MODELING STRUCTURAL TRANSITIONS VIA CLE

The fundamental assumption behind the GPC model is the concept of the random walk (Feller, 1968; Grosberg & Khokhlov, 1994). Here, the steps of the random walker mark out the positions of the "links" and each step (link position) depends on its relationship to the previous link and has no memory of what is further behind or what is to come next.¶ Some additional "memory" is worked into this model by employing the excluded volume (γ) to the random walker permitting us to *approximate* a self-avoiding random walk [eqn (I-C.2)] (Fisher, 1966). In this way, the polymer can avoid occupying its own space. This is more characteristic of how real polymers behave, and is the model we will use throughout Part II. The Gamma probability function [eqn I-C.2)] reduces to the historic GPC for $\gamma = 1$.

These assumptions are equivalent to a Markov chain model. The Markov model allows us to write the probability of an *uncorrelated* polymer chain (of structure **S**) measured from $\tilde{1}$ to $\tilde{N}$ as follows:

$$p_S(\tilde{1}, \tilde{N}) = p(\tilde{N}|\tilde{N} - \tilde{1})p(\tilde{N} - \tilde{1}|\tilde{N} - \tilde{2}) \cdots p(\tilde{1}), \quad (1)$$

where $p(\tilde{j}|\tilde{j} - \tilde{1})$ expresses the conditional probability of link $\tilde{j}$ given that link $\tilde{j} - \tilde{1}$ is known. The tilde over the indices is meant to emphasize the fact that the index (e.g. $\tilde{j}$) is referencing a "link" that need not have any one to one correspondence with an individual "mer" (see Part I: Sections I-2.2, I-2.3 and Appendix I-B). Indeed, typically, a "link" will involve a cluster of "mers " (Flory *et al.*, 1966a; Flory & Semlyen, 1966b).

Due to a requirement of self-consistency and symmetry considerations (Appendix I-B), $p(\tilde{j}|\tilde{j} - \tilde{1}) = p(\tilde{1})$ and eqn (1) simplifies to

$$p_S(\tilde{1}, \tilde{N}) = \prod_{\tilde{i}=\tilde{1}}^{\tilde{N}} p(\tilde{1}) = p^{\tilde{N}} \quad (2)$$

which puts the expression in the form of eqn (I-B.6). This can also be expressed as a multidimensional Gaussian function (Feller, 1971), whereupon eqn (2) represents the ratio of the thermodynamic probabilities for a given macrostate $\Omega_S(\tilde{1}, \tilde{N})$ with respect to a reference macrostate $\Omega_{S_o}(\tilde{1}, \tilde{N})$

$$\frac{\Omega_S(\tilde{1}, \tilde{N})}{\Omega_{S_o}(\tilde{1}, \tilde{N})} = Det(P_S[(\tilde{1}, \tilde{N})]) = \prod_{\tilde{i}=\tilde{1}}^{\tilde{N}} P_S[(\tilde{1}, \tilde{N})]_{\tilde{i},\tilde{i}}, \quad (3)$$

§The following are some important definitions used in Part I and where they are located. Section I-2.1: Base pair density (BPD), MBL hierarchical complexity (HC). Section I-2.2: monomer separation distance (b), persistence length (ξ: coil state (ξ_c) and folded state (ξ_f)), excluded volume (γ), "link" and "mer".

¶Time reversal symmetry is preserved because the tape can be played either backwards or forwards.

where $P_S[(\tilde{1}, \tilde{N})]$ is a diagonal probability matrix for the *uncorrelated* structure **S** whose non-zero elements correspond to $P_S(\tilde{1}, \tilde{N})]_{\tilde{i},\tilde{i}}$ and whose length is $\tilde{N}$ links.

When cross-links are formed on the GPC, this very simple picture needs to be modified slightly. In adding cross-links to the GPC, we can assume that the formation of an initial cross-link at $(\tilde{i}, \tilde{j})$ has the same entropic cost as folding a free GPC of length $\tilde{j} - \tilde{i} + \tilde{1}$ (where $\tilde{j} > \tilde{i}$ is assumed). However, each additional link that forms a contiguous stem with the nucleation site *might* have some additional correlation effects (σ) associated with it. When the stem is broken by a bulge ($\mathscr{B}$), a loop ($\mathscr{H}$), an internal loop ($\mathscr{I}$) or an iMBL, a new nucleation site is formed and the process begins again at the new junction.

Figure 1(a) shows an example in which one cross-link (labeled $\mathscr{N}$) represents the location of the nucleation site and the other two cross-links (labeled $\mathscr{C}$) represent the correlated cross-link sites. Figure 1 represents one of several possible pathways. For nucleation that proceeds sequentially from the head in the direction of the tail $((\tilde{3}, \tilde{7}) \rightarrow (\tilde{2}, \tilde{8}) \rightarrow (\tilde{1}, \tilde{9}))$, the Markov probability becomes

$$p_S = p_S[(\tilde{1}, \tilde{9})|(\tilde{2}, \tilde{8})]\ p_S[(\tilde{2}, \tilde{8})|(\tilde{3}, \tilde{7})]\ p_S[(\tilde{3}, \tilde{7})], \tag{4}$$

where $p_S = det(P_S)$.

We have already assumed that a Markov rule is sufficient to describe this behavior. The rationale for continuing with a Markov rule is simply that we started with that model to describe the behavior of an *uncorrelated* polymer chain (which has plenty of experimental evidence to support it: Grosberg & Khokhlov, 1994). Likewise, an appropriate choice of the link size (ξb) already accounts for most of the correlation effects. Hence, at this point, there is no reason to discard the model in an intramolecular cross-linked polymer chain. The extent to which the model conforms to reality is much the subject of this series.

We introduce a chain creation operator $\hat{p}_{\tilde{N}}(\tilde{i}, \tilde{j})$ whose properties are to create a subchain from $\tilde{i}$ to $\tilde{j}$ on an $\tilde{N} \times \tilde{N}$ identity matrix ($I_{\tilde{N}}$). For example, if $\tilde{N} = \tilde{5}$ and $(\tilde{i}, \tilde{j}) \Rightarrow (\tilde{2}, \tilde{4})$, then

$$\hat{p}_{\tilde{N}}(\tilde{2}, \tilde{4})\, I_{\tilde{N}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 0 & 0 \\ 0 & 0 & p & 0 & 0 \\ 0 & 0 & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

A corresponding annihilation operator $\hat{p}^{\dagger}_{\tilde{N}}(\tilde{i}, \tilde{j})$ removes these chain elements.

In the interest of considering the issue, we also introduce a correlation operator $\hat{\sigma}_{\tilde{N}}(\tilde{k}, \tilde{l})$ $(\tilde{k} \neq \tilde{l})$ whose function is to add correlation at the off-diagonal positions $(\tilde{k}, \tilde{l})$ and $(\tilde{l}, \tilde{k})$. Applied to the same example in eqn (5), the operator $\hat{\sigma}_{\tilde{N}}(\tilde{2}, \tilde{4})$ yields

$$\hat{\sigma}_{\tilde{N}}(\tilde{2}, \tilde{4})\, \hat{p}_{\tilde{N}}(\tilde{2}, \tilde{4})\, I_{\tilde{N}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & \sigma & 0 \\ 0 & 0 & p & 0 & 0 \\ 0 & \sigma & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{6}$$

This also has a corresponding annihilation operator $\hat{\sigma}^{\dagger}_{\tilde{N}}(\tilde{k}, \tilde{l})$. For $\tilde{i} \leqslant \tilde{k}, \tilde{l} \leqslant \tilde{j}$ and $\tilde{k} \neq \tilde{l}$, the operations $\hat{p}$ and $\hat{\sigma}$ commute $\hat{\sigma}_{\tilde{N}}(\tilde{k}, \tilde{l})\hat{p}_{\tilde{N}}(\tilde{i}, \tilde{j})I_{\tilde{N}} = \hat{p}_{\tilde{N}}(\tilde{i}, \tilde{j})$ $\hat{\sigma}_{\tilde{N}}(\tilde{k}, \tilde{l})I_{\tilde{N}}$ where $\{(\tilde{i}, \tilde{j})|(\tilde{i}, \tilde{j}) \in \mathbf{S}\ \&\&\ (\tilde{i}, \tilde{j})$ one to one$\}$.

This notation system affords us a compact representation of the matrices in eqn (4). In the first step

$$P_S[(\tilde{3}, \tilde{7})] = \hat{p}_{\tilde{9}}(\tilde{3}, \tilde{7})\, I_{\tilde{N}}. \tag{7}$$

For contiguous cross-links formed after $\mathscr{N}$ in Fig. 1, two equally possible cross-links are allowed in the next step: $(\tilde{2}, \tilde{8})$ and $(\tilde{1}, \tilde{9})$. The condition of path independence requires that $P_S[(\tilde{2}, \tilde{8})|(\tilde{3}, \tilde{7})] = P_S[(\tilde{2}, \tilde{8})|(\tilde{1}, \tilde{9})]$. If we insist on incorporating correlation, then the only way to preserve the invariance in eqn (4) is to write the conditional matrix in Fig. 1(c) as follows:

$$P_S[(\tilde{2}, \tilde{8})|(\tilde{3}, \tilde{7})] = \hat{p}_{\tilde{9}}(\tilde{2}, \tilde{8})\hat{p}_{\tilde{9}}(\tilde{2}, \tilde{8})\, I_{\tilde{N}}, \tag{8}$$

where the matrix elements are

$$P_S[(\tilde{2},\tilde{8})|(\tilde{3},\tilde{7})]_{\tilde{i},\tilde{j}} = P_S[(\tilde{2},\tilde{8})|(\tilde{3},\tilde{7})]_{\tilde{j},\tilde{i}}$$

$$= \begin{cases} p, & \tilde{i} = \tilde{j}, \\ \sigma & (\tilde{i} = \tilde{2} \;\&\&\; \tilde{j} = \tilde{8}), \\ 0 & \text{elsewhere.} \end{cases} \quad (9)$$

In essence, the Markov rule is saying that each time a cross-link is introduce, the random walker must travel the circuit of the polymer chain again. The polymer chain has $\tilde{N}!$ states available to it. From Lemma I-1 and eqn (I-20), the *maximum* number of configurations that can be frozen out by the CLE process is $\sqrt{\tilde{N}!}$ where correlation would reduce that number slightly. These states should not be ignored when the circuit is closed because $\tilde{2}$ and $\tilde{8}$ are coupled to the large mass in between them.

In the last step (eqn (4) and Fig. 1(d)), $P_S[(\tilde{1},\tilde{9})|(\tilde{2},\tilde{8})] = \hat{\sigma}_{\tilde{9}}(\tilde{1},\tilde{9})\,\hat{p}_{\tilde{9}}(\tilde{1},\tilde{9})I_{\tilde{N}}$ which yields

$$P_S[(\tilde{1},\tilde{9})|(\tilde{2},\tilde{8})]_{\tilde{i},\tilde{j}} = P_S[(\tilde{1},\tilde{9})|(\tilde{2},\tilde{8})]_{\tilde{j},\tilde{i}}$$

$$= \begin{cases} p, & \tilde{i} = \tilde{j}, \\ \sigma & (\tilde{i} = \tilde{1} \;\&\&\; \tilde{j} = \tilde{9}), \\ 0 & \text{elsewhere.} \end{cases} \quad (10)$$

The determinant of eqn (10) corresponding to Fig. 1(d) becomes

$$p_S[(\tilde{1},\tilde{9})|(\tilde{2},\tilde{8})] = p^7(p^2 - \sigma^2) = p^9(1 - q^2), \quad (11)$$

where $q = \sigma/p$.

From eqns (I-4) and (I-B.12), we write $p^9 = p(r_{\tilde{1},\tilde{9}})$ which leads us to eqn (I-C.2). In general, this becomes

$$p^{\tilde{j}-\tilde{i}+\tilde{1}} = p(r_{\tilde{i},\tilde{j}}), \quad (12)$$

where $\tilde{i} < \tilde{j}$ is assumed. The CLE for $(\tilde{i},\tilde{j})$ becomes

$$\Delta S_{\tilde{i},\tilde{j}} = k_B \ln(p_{\tilde{i},\tilde{j}})$$

$$= \begin{cases} k_B \ln(p(r_{\tilde{i},\tilde{j}})), & \tilde{i},\tilde{j} \rightsquigarrow \mathcal{N}, \\ k_B \ln(p(r_{\tilde{i},\tilde{j}})) + k_B s_\sigma(q), & \tilde{i},\tilde{j} \rightsquigarrow C, \end{cases} \quad (13)$$

where $s_\sigma(q) = \ln(1 - q^2)$. This is the same as eqn (I-13) less the correlation contribution ($k_B s_\sigma(q)$).

This model is easily extended to any form of secondary structure. For example, two adjacent hairpin loops ($\mathscr{H}_a$ and $\mathscr{H}_b$) are independent and represented by a matrix P_{S_a} and P_{S_b} which forms a diagonal

$$\frac{\Omega}{\Omega_0} = \begin{vmatrix} P_{S_a}^{\mathscr{H}} & 0 \\ 0 & P_{S_b}^{\mathscr{H}} \end{vmatrix}, \quad (14)$$

where Ω/Ω_0 is a highly abbreviated shorthand for the conditional thermodynamic probability of the macrostate B given the initial macrostate A (i.e. $p_S[(B)|(A)]$). Likewise, an internal loop $\mathscr{I}$ is represented as

$$\frac{\Omega}{\Omega_0} = \begin{vmatrix} P_S^{\mathscr{I}/2} & 0 & \sigma_S^{\mathscr{I}/2} \\ 0 & P_S^{\mathscr{H}} & 0 \\ \sigma_S^{\mathscr{I}/2} & 0 & P_S^{\mathscr{I}/2} \end{vmatrix}, \quad (15)$$

where $P_S^{\mathscr{I}/2}$ is a shorthand for the diagonal probability elements associated with half of $\mathscr{I}$, and $\sigma_S^{\mathscr{I}/2}$ expresses the corresponding off diagonal components. In all cases, the correlation contributions are independent of the order in which the secondary structure is constructed.

At this point, there are several peculiarities that raise questions about the incorporation of correlation into this problem. As long as eqn (4) is processed as determinants, the path invariance is guaranteed. However, suppose the order of cross-linking for eqn (4) and Fig. 1(a) follows the path $(\tilde{3},\tilde{7}) \rightarrow (\tilde{1},\tilde{9}) \rightarrow (\tilde{2},\tilde{8})$. By a true Markov model, $(\tilde{3},\tilde{7})$ and $(\tilde{1},\tilde{9})$ are *not* nearest neighbors and there is no reason to assign correlation to $(\tilde{1},\tilde{9})$ in this step except to satisfy the path invariance requirements. (We could demand correlation at *all* cross-link sites to remove this discrepancy.) For short ranges as in Fig. 1(a), this may be a minor issue. However, if the stem forms 100 contiguous BPs, such rules would require that $\mathscr{N}$ at the head of the stem should cause correlation at all other contiguous sites along the stem. If the next BP forms 100 nt downstream from the head of the loop, it seems out of reason to presume that this is helped by correlation; particularly, when two non-contiguous stems must find their corresponding cross-links without the

aid of correlation even when that distance only corresponds to the next nearest neighbor. If such contributions are to be taken seriously, then their role resembles something akin to enthalpy which has already been specified in these problems.

Therefore, whereas we have introduced the concept of correlation to the Markov model, in examining its properties in the broader context, the results do not as yet appear to be convincingly Markovian in character and we are still inclined to ignore them at present. Our recommendation currently is to handle changes in ξ by way of renormalization of P_S in the regions where changes in correlation (ξ) occur (Plischke & Bergersen, 1994; de Gennes, 1979). Finally, even if this correlation is present, its effect is to introduce an additive constant [eqn (13)] per link which has a *maximum* variation of about 0.4 kcal mol^{-1} [compare column 4 in Table I(2) against the mean value 3.0 kcal mol^{-1}].

In Part I, we derived eqn (13) through appeals to physical arguments. These were also supported to some extent with experimental evidence and will be further supported in this work. Here, we have explicitly shown the mathematical formalism and what assumptions were made to arrive at the same results found in Part I. This formalism provides the foundation for advances into non-Markov models (to be addressed in future work), it shows us how to move away from our monolithic persistence ratio (ξ), and it helps advance the theory to a kinetic model (Section 2.3).

In a transition from secondary structure $\mathbf{S}_1$ to $\mathbf{S}_2$, operations using $\hat{p}^\dagger$ ($\hat{\sigma}^\dagger\hat{p}^\dagger$) and $\hat{p}$ ($\hat{\sigma}\hat{p}$) are required. In this respect, the operators simulate the dynamics of the folding process. This separability is an important feature of the Markov process.

2.2. MODELS FOR THE LOCAL CONTRIBUTIONS TO THE CLE

In Part I, it was shown that postulate 3 yields the same entropic penalties as the Turner energy rules to within an additive constant for a loop of size n and a stem of 4 bp. This contribution was attributed to local CLE effects. To integrate the CLE formalism into a secondary structure filter of NNSS algorithms, we now consider the origin of that constant.

The NNSS approaches all assume that transitions of RNA secondary structure from the coil state to the folded state ($c \to f$) cause a penalty at sequence locations (i, j) where $\mathscr{B}$, $\mathscr{H}$, $\mathscr{I}$ and iMBL structures are formed (Section I-2.3.2). The estimated entropy loss (due to cross-linking) is

$$\Delta S_{\mathscr{X}}^{c\to f} = \Delta S_{\mathscr{X}_o}^{c\to f} - \gamma k_B \ln(n), \tag{16}$$

where $\Delta S_{\mathscr{X}_o}^{c\to f}(\mathscr{X} \equiv \mathscr{B}, \mathscr{H}, \mathscr{I})$ is a negative constant, $\gamma k_B \ln(n)$ conforms to the theory of Jacobson & Stockmayer (1950), and n is the length of the free segment closing the secondary structure. This also applies to iMBLs, but the formula is often approximated by a linear equation in n and it also depends on the number of branching points ($\mathscr{V}$). Hence, NNSS strategies also carry with them a constant that has never been explained.

At present, we attribute the constant ($\Delta S_{\mathscr{X}_o}^{c\to f}$) to one of three possibilities: (1) a nucleation (or formation) entropic contribution at the secondary structure junctions (Scheffler *et al*., 1970), (2) a freezing out of the local degrees of freedom available to the six bonds that link the 5′ end of one sugar to the 5′ end of the adjacent sugar on a nucleic acid chain as well as the loss of free rotation for the base that is attached to the sugar (Searle & Williams, 1993), or (3) a combination of these effects.

The NNSS penalty is assigned as though the source of $\Delta S_{\mathscr{X}_o}^{c\to f}$ is nucleation. The CLE can adopt a similar strategy; however, it is not expected to fit the same baseline ($\Delta\mathscr{G}^0 = -T\Delta S_{\mathscr{X}_o}^{c\to f}$). To standardize the local CLE interactions (for $\xi = 3$ nt), the NNSS results of tRNA^{phe} were used to establish the baseline for the local contributions.‖ The following procedures are used.

(1) If freezing out is the source, then a constant local cross-linking contribution ($\Delta\overline{\mathscr{G}}_{frz}$) is added to the cross-linking entropy (the bar over $\Delta\mathscr{G}$ indicates an "average value").

(2) If nucleation is the source, then a constant $\Delta\overline{\mathscr{G}}_{ncl}^{\mathscr{X}}(\mathscr{X} \equiv \mathscr{B}, \mathscr{I}, \mathscr{H}$ or iMBL) is added for each stem closing a secondary structure feature

‖The NNSS entropic penalties have also come about largely as a result of "tuning".

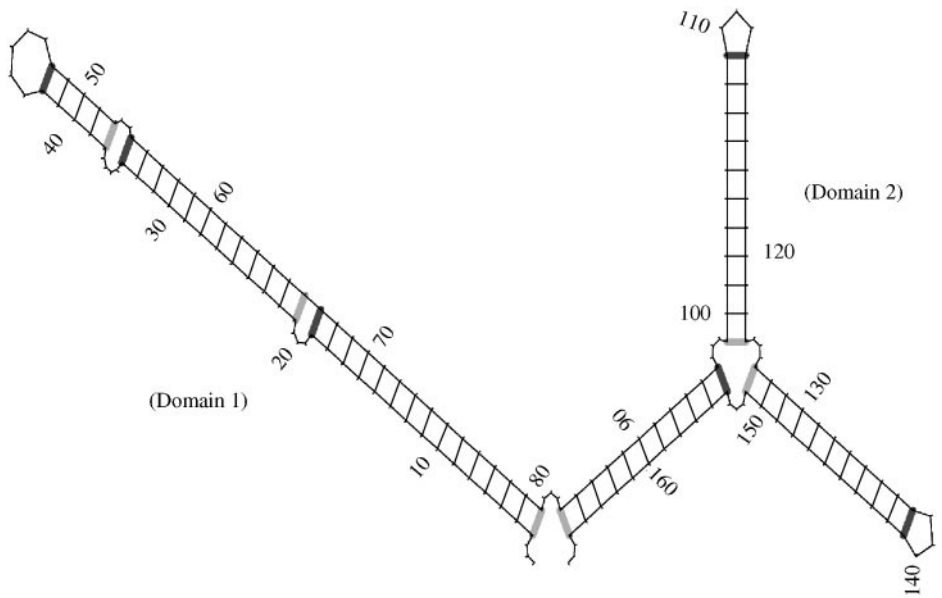

FIG. 2. A model for understanding the assignment of nucleation and correlation effects discussed in Section (2.2). Nucleation that begins at the head of the loop and progresses toward the 5′-3′ end of the domain is marked by the dark gray bars. Nucleation that proceeds from the 5′-3′ end of the domain toward the head of the respective loop (or loops) is indicated by the light gray bars. Irrespective of where nucleation begins on a contiguous stem (loop side, middle, 5′-3′ end or the folding direction), assignment of correlation goes to the remaining contiguous bonds forming a single stem. (Correlation can also be assigned after accounting for the first full link by excluding the additional bonds along the link up to one full persistence length ξb, where b is the monomer separation distance. Likewise, correlation can be assigned to all sites.)

(Fig. 2). For simplicity, it is assume here that all secondary structures have the same nucleation energies. As shown in Section 2.3, nucleation is most likely to proceed from the head of a loop (the dark gray bars: Fig. 2) than from the closing end of the stem (the light gray bars: Fig. 2) (Scheffler *et al.*, 1970). Nevertheless, regardless of where the nucleation point forms on a contiguous stem, the FE is path independent (Section 2.1).

(3) If correlation is used, then $\Delta\overline{\mathscr{G}}_{corr} = k_{\mathrm{B}}Ts_{\sigma}(q)$ is added to $\Delta\overline{\mathscr{G}}^{\mathscr{X}}_{ncl}$ and $\Delta\overline{\mathscr{G}}_{frz}$ through a three parameter relationship (see Fig. 2 for a description of the assignment). Since we assume ξ is constant, we also assume a constant for the correlation: $\Delta\overline{\mathscr{G}}_{corr} = -0.4\,\mathrm{kcal\,mol^{-1}}$ at 310 K (Section 2.1).

(4) To handle different values of ξ, the local entropy is weighted by the function w_{ξ}. We expect w_{ξ} to be variable and a function of ξ: larger for GC-rich BPs, smallest for GU-rich BPs, and somewhere in the middle for AU-rich BPs. In this work, we must assume that w_{ξ} is the same for all cross-links: the *mean* local CLE.

(5) For a given BP (i, j), the lookup tables of the Turner energy rules for nearest-neighbor (NN) stacking of oligo-nucleotides, dangling bond corrections, etc. are used to calculate $\Delta H_{NN(i,j)} - T\Delta S_{NN(i,j)} = \Delta G_{NN(i,j)}$. First note, we have shifted from "link" notation to the "mer" notation (i, j). Second, note that $T\Delta S_{NN(i,j)}$ is *not* the penalties for $\mathscr{B}$, $\mathscr{I}$, $\mathscr{H}$, and iMBL structures: raw data from NNSS programs will be denoted as $\Delta\mathscr{G}_{ss}$ and the corresponding secondary structure (ss) penalties will be denoted by ΔS_{ss}. Instead, ΔG_{NN} represents the FE found in NNSS algorithms minus the usually applied secondary structure penalties.

The change in FE for BP (i, j) to transition to the folded state $(c \to f)$ becomes

$$\langle \Delta\mathscr{G}^{c\to f}_{i,j}\rangle = \begin{cases} \Delta G_{NN(i,j)} + \langle \Delta\mathscr{G}_{cl(i,j)}\rangle & \\ \quad + w_{\xi}\Delta\overline{\mathscr{G}}^{\mathscr{X}}_{ncl(i,j)} & \text{if } (i,j) \equiv \mathscr{B}, \mathscr{H}, \mathscr{I}, \text{iMBL} \\ & \text{and } \Delta\overline{\mathscr{G}}^{\mathscr{X}}_{ncl} > 0 \\ \quad + w_{\xi}\Delta\overline{\mathscr{G}}_{frz(i,j)} & \Delta\overline{\mathscr{G}}_{frz} > 0 \\ \quad + w_{\xi}\langle\Delta\overline{\mathscr{G}}_{corr(i,j)}\rangle & \langle\Delta\overline{\mathscr{G}}_{corr}\rangle \leqslant 0, \end{cases} \tag{17}$$

where $\langle \Delta \mathscr{G}_{cl(i,j)} \rangle$ ($\equiv \langle \Delta \mathscr{G}^{c \to f}_{cl(i,j)} \rangle$) is the global CLE contribution for cross-link (i, j) and the $\langle \cdots \rangle$ notation is used to indicate that these FE values are averaged over the range of ξ (Part I).

These results are summed to obtain the total FE for the given secondary structure

$$\langle \Delta \mathscr{G}^{c \to f} \rangle = \sum_{all \ \ (i,j) \in \mathbf{S}} \langle \Delta \mathscr{G}^{c \to f}_{i,j} \rangle. \tag{18}$$

2.3. KINETICS FOR A GPC–CLE SYSTEM

As an illustrative example, we consider the progress of a two-state system in which the cross-links have a unique pairing (Fig. 3) where the dark and light circles correspond to sites 1 and 2, respectively. Multiple pairing possibilities would force us to sketch out all of the pathways, but otherwise, the result is not changed in this simple model and these alternative pathways would also be subject to the same treatment described here.

Two different paths are considered in Fig. 3: the upper half ($coil \leftrightarrow (a, 1) \leftrightarrow (b, 2)$) and the lower half ($coil \leftrightarrow (a, 2) \leftrightarrow (b, 1)$), where for the notation (X, m), X denotes the transition step and m denotes the particular bond formed. The "link" concept is assumed in this section but can easily be adjusted to fit a "mer" perspective.

The rate of formation of a given state (X, m) is

$$k^f_{(X,m)} = \chi_m \exp(-\Delta \mathscr{G}/k_B T), \tag{19}$$

where $\Delta \mathscr{G}$ is the difference between the initial state (m unpaired) and the final state (m paired), and χ_m is the mean formation rate for bond m (Mironov *et al.*, 1985). On applying this to the transition ($coil \to a$), the change in the FE for a transition into state a becomes

$$\begin{aligned} \Delta \mathscr{G}[(a, 1) \leftarrow coil] &= \Delta \mathscr{G}^{(1)}_{NN} + \langle \Delta \mathscr{G}^{local}_{cl} \rangle + \langle \Delta \mathscr{G}^{(1)}_{cl} \rangle, \\ \Delta \mathscr{G}[(a, 2) \leftarrow coil] &= \Delta \mathscr{G}^{(2)}_{NN} + \langle \Delta \mathscr{G}^{local}_{cl} \rangle + \langle \Delta \mathscr{G}^{(2)}_{cl} \rangle, \end{aligned} \tag{20}$$

where $\Delta \mathscr{G}^{(m)}_{NN}$ ($= \Delta H^{(m)}_{NN} - T \Delta S^{(m)}_{NN}$) is the FE of bond formation for cross-link m (where all the NNSS $\mathscr{B}$, $\mathscr{H}$, $\mathscr{I}$ and iMBL penalties are completely removed), $\langle \Delta \mathscr{G}^{local}_{cl} \rangle$ is the *local* CLE-FE contribution (assumed the same for both sites), and $\langle \Delta \mathscr{G}^{(m)}_{cl} \rangle$ is the corresponding global CLE-FE for the formation of a cross-link at site m (see Section 2.2).

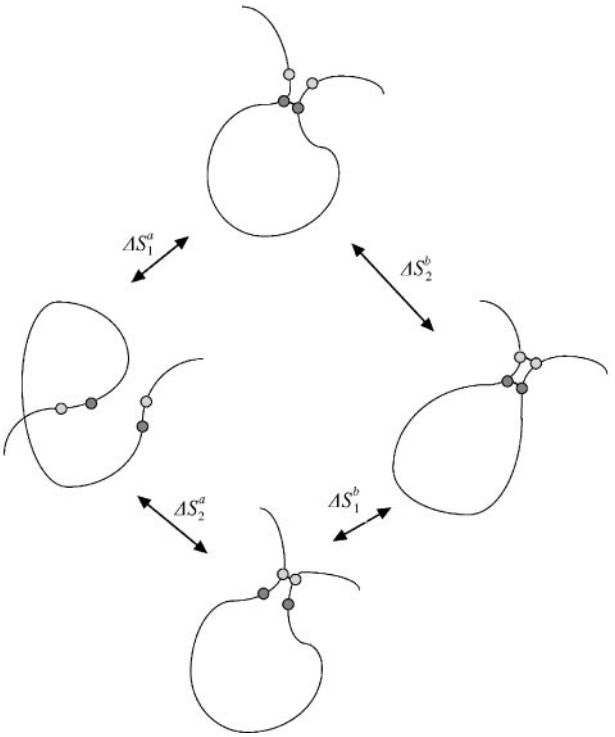

FIG. 3. A model of the folding pathway kinetics for a polymer which forms only two uniquely different bonds: where black can only pair with black (bond 1) and white with white (bond 2). The distance between site 1 (black circle) and site 2 (white circle) depends on the number of *links* that separate them along the chain. If the distance is small, then $\langle \Delta \mathscr{G}^{(1)}_{cl} \rangle \sim \langle \Delta \mathscr{G}^{(2)}_{cl} \rangle$. If the distance is large, then $\langle \Delta \mathscr{G}^{(1)}_{cl} \rangle \ll \langle \Delta \mathscr{G}^{(2)}_{cl} \rangle$. Reversibility is assumed.

If we allow that $\Delta \mathscr{G}^{(1)}_{NN} = \Delta \mathscr{G}^{(2)}_{NN}$ and $\chi_1 = \chi_2$, then the local contributions become a constant $\Delta \mathscr{G}_{NN}$. The rate constants become

$$\begin{aligned} k^f_{(a,1)} &= \chi \ K_{NN} \exp(-(\langle \Delta \mathscr{G}^{(1)}_{cl} \rangle)/k_B T), \\ k^f_{(a,2)} &= \chi \ K_{NN} \exp(-(\langle \Delta \mathscr{G}^{(2)}_{cl} \rangle)/k_B T), \end{aligned} \tag{21}$$

where $K_{NN} = \exp\{-(\Delta \mathscr{G}_{NN} + \langle \Delta \mathscr{G}^{local}_{cl} \rangle)/k_B T\}$.

The CLE-FE is a positive quantity for bond formation and $\Delta \mathscr{G}^{(1)}_{cl} < \Delta \mathscr{G}^{(2)}_{cl}$. Hence, $k^f_{(a,1)} > k^f_{(a,2)}$ (given that all other parameters are held the same). Thus, the rate of formation for the upper half of Fig. 3 will be favored over the lower half even though the two sites are otherwise identical in their thermodynamic properties.

Adjustments to $\Delta H_{NN}^{(1)}$ and $\Delta H_{NN}^{(1)}$ can change the behavior favoring $(a, 2)$; however, that only further supports our point that these features can be understood with some basic tools. It is also true that when $\Delta \mathscr{G}_{cl}^{(1)} \sim \Delta \mathscr{G}_{cl}^{(2)}$ (as when both sites are in close proximity), the formation rates will also be similar (given that all other parameters are the same). Nonetheless, this only shows that long folding regions (where $\tilde{j} - \tilde{i} \gg \tilde{1}$) require very large enthalpies to successfully dominate the folding kinetics over neighboring regions (where $\tilde{j} - \tilde{i} \to \tilde{1}$). Finally, even if we permit mixed pairing between sites 1 and 2, this still yields $\Delta \mathscr{G}_{cl}^{(1 \cdot 1)} < \Delta \mathscr{G}_{cl}^{(1 \cdot 2)}$, which shows that bond 1 $(1 \cdot 1)$ is still favored over the $1 \cdot 2$ bond.

Therefore, given that all other parameters are held the same, the folding of a hairpin loop will tend to proceed from the head of the loop in the direction of the tail (as in Fig. 1). This has been known for some time in protein folding problems (Mironov *et al*., 1985) and RNA folding (Scheffler *et al*., 1970), but this is the first time that we know of that a theory has shown that this is what will happen when all other thermodynamic properties about the structure are the same.

This gives us some important insight into why polymers fold as they do and how we might be able to design new ones. Moreover, we actually do not need to grovel before a machine to arrive at this theoretical biology. This latter point is what we find most encouraging about this work.

2.4. CALCULATION METHODS

All calculations were carried out at the Human Genome Center of the University of Tokyo Institute of Medical Science and the National Institute of Infectious Diseases in Japan.

Although we focus on the outputs from the MFOLD algorithm in this work, the issues raised apply to all NNSS algorithms and the results are not expected to differ drastically as a result of merely exchanging algorithms. Nevertheless, this matter was not verified rigorously.

The FE parameters and penalties used in this work are all from the stock GCG (version 10.1) distribution which utilizes the Turner energy rules (Turner *et al*., 1988). The *general* observations reported here are not changed by simply upgrading to MFOLD (v. 3.1) although *local* predictions are typically improved.** In the MFOLD calculations, the entire sequence was input. There was no cutting of the original RNA sequence and no freezing of specified bonds. The optimal and suboptimal secondary structure results were obtained from connect files produced by the PLOTFOLD utility using a *p*-value of 5% (GCG:- INC = "(default settings) ") and a structural distance value of 1 (GCG:-WIN = 1) (Jaeger *et al*., 1990) with "-LIS = 100" (the number of listed structures). To make a more complete search, a *p*-value of 10% with "-LIS = 150" (the maximum allowed) was also used. All calculations were carried out at 37°C to insure the best fit of the Turner energy rules from the lookup tables (Huynen *et al*., 1997).

Structural analysis was carried out on data from the connect file using a recursive analysis program written by the authors. The order of the structures in the connect file is prioritized in terms of the FE: a higher number means a less negative FE than the previous structure. Hence, structure # 1 corresponds to the so-called "optimal structure" and all larger numbers correspond to successively less optimal (i.e. suboptimal) structures. Here, we use the order in the connect file to index the MFOLD connect file structures. Predictions using both MFOLD and cross-linking entropy are numbered in terms of their FE where a smaller index corresponds to a larger (more negative) FE and presumably a structure which is thermodynamically more probable. The secondary structure of each sequence was analysed in terms of its hierarchy and the details about each secondary structure feature were recorded in terms of the type of structure (i.e. $\mathscr{B}$, $\mathscr{I}$, $\mathscr{H}$, $\mathscr{V}$, iMBL or domain boundary of a pMBL). Based on this information, the cross-linking entropic contribution for each secondary structure was computed [eqns (18), see also eqns (I-13) and (I-17)].

The major departure of this work from the standard NNSS strategy is the use of the CLE to evaluate all the entropic contributions of $\mathscr{B}$, $\mathscr{I}$, $\mathscr{H}$, and iMBL structures in place of the standard lookup tables for the entropy. We already showed in Part I (Section I-3.2) that these

**The structures obtained from MFOLD v 2.3 resemble the "best" structures found in Jaeger *et al*. (1990).

penalties appear to be the result of a generic sequence of RNA ($\xi \sim 3$ nt, $T \sim 310$ K and $N \sim 100$ nt). This will be further substantiated in the results reported in this work. There is no formal theory showing where the penalties in eqn (16) originate. Finally, the issue about using logarithmic penalties for the iMBLs remains to this day (Lyngsø, 1999). Therefore, we have removed the original NNSS penalties for the entropy†† and calculated entropic contributions strictly in terms of cross-linking entropy rules. Our analysis program is able to recalculate the asymmetric penalties for the internal loops (Mathews *et al*., 1999); however, these values are usually small and were left as it is.

In this work, we have assumed that ξ is a constant. With the exception of a general scan of the distribution as a function of ξ, values for the persistence ratio were limited to an experimentally justifiable range ($2.5 < \xi < 9.0$) in this work. We have also set the stacking gap distance between the AU, GC, and GU BP to $\lambda = 2$ (Section I-3.1).

2.4.1. *Secondary Structure Comparisons*

To verify that the cross-linking entropy is finding domain sizes and structures that represent some tangible approximation of an experimentally observed domain, three well-known structures were tested: the 16S ribosomal RNA (rRNA) *Escherichia coli* (GenBank accession number ECORRD), the group I intron *T. thermophila* (Cech, 1988; Pan & Woodson, 1999) and tRNA^{Phe} (Hagerman, 1997). All these sequences have well-established secondary structures (Wimberly *et al*., 2000; Cech, 1988; Hagerman, 1997) published in the literature. The tRNA^{Phe} sequence can be written without a large number of methylated bases or pseudouridine (Hagerman, 1997) which further complicates or obfuscates the calculation due to the non-standard (ACGU) bases that are present. The process for tRNA^{Phe} appears to be reversible (Hagerman, 1997). Likewise, the *T. thermophila* is also known to fold in a reversible process (Pan & Woodson, 1999), hence we should expect that secondary structures at the top of the list will be populated with the actual domains of the group I intron. The distributions for rRNA are not as certain in this regard; however, we should expect that at least some of the representative structures will appear.

The NNSS program was used to generate a list of secondary structures, and from that list, an evaluation of the CLE was used to rank the list using the CLE contribution to the FE. The list that the CLE can analyse is clearly limited by the list that is generated from the NNSS program in the first place. However, since there is no way of determining the correctness of a domain from a random sequence (short of carrying out an experiment), the distribution obtained from this procedure is qualitatively the best we can expect with the current state of the art methods as far as detection or prediction is concerned. Since the primary objective of this work is simply to show that domain size is governed by the cross-linking entropy, it is sufficient to show that correct structures from the given list *are* being found. In as much as the algorithm is finding the most representative structures from the list, the algorithm is also finding the best domain structures (with the attached proviso).

An important issue is the distribution of structures. Structures at the top of the list of suboptimal structures should represent the thermodynamically most probable structures. Non-physiological conditions will somewhat obscure the correct outcome. Nevertheless, the best criteria we have are the domains that are reported, and these should appear *near* the top of the list. Moreover, important domains should appear in groups or clusters because of the higher probability of those domains.

2.4.2. *Functional Domain Comparisons*

To study the nature of functional domains with persistence length, two strategies were used. The first approach was to examine the behavior of particular secondary structures obtained from the secondary structure study of known sequences with respect to ξ and the various local entropic contributions such as $\Delta\overline{\mathcal{G}}^{X}_{ncl}$, $\Delta\overline{\mathcal{G}}_{frz}$, and correlation effects. The second strategy was to find the average maximum domain size from the top five secondary structures of shuffled

††In calculating the $\mathscr{I}$ penalties for $n = 2$, we have assumed that the entropic penalty is 4.1 kcal mol^{-1}.

sequences. From the set of shuffled sequence results (all the same length and sequence composition), a distribution of domain sizes was obtained. By using shuffled sequences, we avoid biasing the results in terms of well-established biologically active sequences and by examining real structures, we avoid obscuring the power of this algorithm in its ability to find the *correct* functional domains when they are actually present.

If the theory is finding the functional domains of biologically active sequences, we can also trust its ability to find correct domains in the shuffled sequences. Since it is likely that the weight of the domain sizes are characteristically overestimated (Rivas & Eddy, 2000) in the MFOLD distribution, the results reported here on the average maximum functional domain size should be seen as only approximate.

3. Results

3.1. RIBOSOMAL RNA (rRNA)

Using "-LIS = 100" and "-WIN = 1", the best *E. coli* 16S rRNA structures that we could find (in the list of MFOLD suboptimal structures) matched domains 1, 2, and about half of domain 3. The remaining domains were not matched and there appear to be no reasonable candidates within the connect files. One of the better structures near the top of the list has the index # 13 in the connect file. (In terms of general domain definitions (Gutell *et al*., 1993), domain I is quite well recovered, some of domain II is also recovered, but very little of domain III is recovered.) Most of the secondary structure in domains 1 and 2 is correct, but there are a few notable deviations in the iMBLs. The domain lengths of # 13 are as follows: 17, 529, 197, 582,

TABLE 1

Domain lengths of results of E. coli 16S *rRNA* (*GenBank accession ECORRD*) *for the first* 12 *outputs, the optimal structure predicted by MFOLD* (*MF*(1)), *and the experimentally determined structure of E. coli rRNA* (*Gutell et al*., 1993)*

CLE ss index	MFOLD index	FE results (kcal mol^{-1})				# of domains	Domain boundary sizes
		$\Delta\mathscr{G}_{ss}$	$\langle\Delta\mathscr{G}_{cl}\rangle$	$\lvert T\Delta S_{ss}\rvert$	$\Delta\mathscr{G}$		
1	79	− 733.30	548.00	604.90	− 790.20	8	**17**, **529**, 197, 8, 343, 38, 15, 363
2	13	− 738.10	539.27	589.40	− 788.23	6	**17**, **529**, 197, 582, 133, 53
3	69	− 733.80	542.54	595.30	− 786.56	6	**17**, **529**, 197, 582, 133, 53
4	51	− 734.90	511.28	562.70	− 786.32	15	29, 36, 36, 118, 47, 23, 22, 12, 17, 12, 859, 32, 53, 133, 53
5	87	− 732.90	561.80	605.50	− 776.60	8	**17**, **529**, 21, 66, 97, 25, 392, 367
6	57	− 734.70	550.94	592.80	− 776.56	7	**17**, 849, 307, 17, 300, 20, 11
7	68	− 733.80	552.76	594.70	− 775.74	5	**17**, **529**, 21, 587, 367
8	40	− 735.60	550.17	590.20	− 775.63	8	**17**, **529**, 21, 66, 97, 25, 711, 40
9	78	− 733.30	549.18	589.00	− 773.12	6	**17**, **529**, 197, 582, 133, 53
10	58	− 734.60	567.34	600.80	− 768.06	4	29, 36, 65, 1399
11	28	− 736.80	523.21	554.30	− 767.89	15	29, 36, 36, 118, 47, 23, 22, 12, 17, 14, 1039, 32, 39, 22, 11
12	77	− 733.30	567.18	600.90	− 767.02	6	**17**, **529**, 742, 22, 153, 53
			…				…
87	MF(1)	− 740.80	639.85	607.60	− 708.55	4	1140, 8, 15, 363
			…				…
rRNA (observed)						7	**17**, **529**, 323, 25, 476, 94, 24

*The left most column indicates the cross-linking entropy (CLE) secondary structure (ss) index. The MFOLD index is simply the order found in the MFOLD connect file. The number of domains corresponds to the number of domain boundaries found in the ss (Section I-2.1). The variables $\Delta\mathscr{G}_{ss}$ and $T\Delta S_{ss}$ refer to the raw data obtained directly from the NNSS calculation and its respective database [see eqn (17)]; likewise, $\Delta\mathscr{G}$ and $\langle\Delta\mathscr{G}_{cl}\rangle$ are defined by eqn (18). In the above calculation, only the freeze out entropy ($\Delta\overline{\mathscr{G}}_{frz} = 0.25$ kcal mol^{-1}) was utilized with $\xi = 3.5$ nt. Domains 1 and 2 lengths are the same as the experimental value for rRNA. Domain sizes that correspond to the known domain structure of 16S rRNA are highlighted in bold text. [Note: $\Delta\overline{\mathscr{G}}_{frz} = 0.25$ kcal mol^{-1} is the fit value for $\xi = 3.0$ and leads to slightly low values for the total FE ($\Delta\mathscr{G}$) with $\xi > 3.0$.]

133, and 53 nt. The domains have the following positions along the pMBL: (9, 25), (27, 555), (563, 759), (762, 1343), (1350, 1482) and (1485, 1537) (Table 2). The structure marked "observed" lists the correct domain sizes (Table 1) and positions (Table 2) for the experimentally determined structure of rRNA.

Using a larger *p*-value and increasing the requested number of structures (-LIS = 150) failed to achieve any closer matching structure to the observed 16S rRNA of *E. coli*. Without introducing constraints such as the option "-FORCe" in MFOLD, this appears to be the best structure we can generate for the 1541 nt long sequence (at least with the default settings of GCG v 10.1).

Table 1 lists the top 12 structures found using cross-linking entropy ($\xi = 3.5$ and $\Delta\overline{\mathscr{G}}_{frz} = 0.25\ \text{kcal mol}^{-1}\ \text{nt}^{-1}$). The cross-linking entropy (CLE) secondary structure (ss) index is indicated in the left column, and the next column indicates the MFOLD ss-index which PLOTFOLD automatically sorts by the FE. Columns 3–6 indicate the original NNSS-FE ($\Delta\mathscr{G}_{ss}$), the CLE-FE ($\langle\Delta\mathscr{G}_{cl}\rangle$), the NNSS ss penalties ($|T\Delta S_{ss}|$), and the estimated FE including CLE contributions [see eqns (17) and (18)]. Columns 7 and 8 indicate the number of domains found, and their domain lengths in units of nt.

Tables 1 and 2 have been calculated using the freezing out entropy. Adding a correlation interaction, or incorporating nucleation does not appear to change the distribution substantially. Hence, the baseline appears to be similar for all approaches. The estimated FE values listed are a bit low due to the value selected for the local CLE ($\Delta\overline{\mathscr{G}}_{frz} = 0.25\ \text{kcal mol}^{-1}$ for $\xi = 3.0$ nt). Changing to $\Delta\overline{\mathscr{G}}_{frz} = 0.35\ \text{kcal mol}^{-1}$ [adjusting w_ξ to fit tRNA^{phe} for $\xi = 3.5$ nt (Sections 2.3 and 3.6)] raises CLE ss index 1 (MFOLD ss index 79) to $-740\ \text{kcal mol}^{-1}$ with some minor redistribution of the data in Table 1.

Eight out of the 12 top structures predicted by the CLE have exactly the same first two domains in the same position as structure # 13. There can be little doubt that the CLE is finding the best structures from the list of suboptimal structure in

TABLE 2

Domain positions of the 16S *rRNA of E. coli* (*GenBank accession # ECORRD*) *for the first* 5 *outputs listed in Table* 1, *the optimal structure predicted by MFOLD* (*MF*(1)), *and the experimentally determined structure of E. coli rRNA* (*Gutell et al.*, 1993)*

CLE ss index	MFOLD index	Number of domains	Domain boundary positions
1	79	8	**(9,25)**, **(27,555)**, (**563**,759), (762,769), (772,1114), (1117,1154), (1160,1174), (1176,1538)
2	13	6	**(9,25)**, **(27,555)**, (**563**,759), (762,1343), (1350,1482), (1485,1537)
3	69	6	**(9,25)**, **(27,555)**, (**563**,759), (762,1343), (1350,1482), (1485,1537)
	…		…
5	87	8	**(9,25)**, **(27,555)**, (**563**,583), (586,651), (654,750), (754,778), (782,1173), (1174,1540)
	…		…
87	MF(1)	4	(9,1148), (1151,1158), (1160,1174), (1176,1538)
	…		…
rRNA (observed)		7	**(9,25)**, **(27,555)**, (**563**,885), (887,911), (920,1395), (1403,1496), (1505,1528)

*The left most column indicates the cross-linking entropy (CLE) secondary structure (ss) index. The MFOLD ss-index is simply the order found in an MFOLD connect file. The number of domains corresponds to the number of domain boundaries found in the ss. In the above calculation, $\xi = 3.5$ and $\Delta\overline{G}_{frz} = 0.25\ \text{kcal mol}^{-1}$. Note that the domain boundaries 1 and 2 are at exactly the same positions as the experimental locations of rRNA. Domain boundaries that correspond to the known domain structure of 16S rRNA are highlighted in bold text.

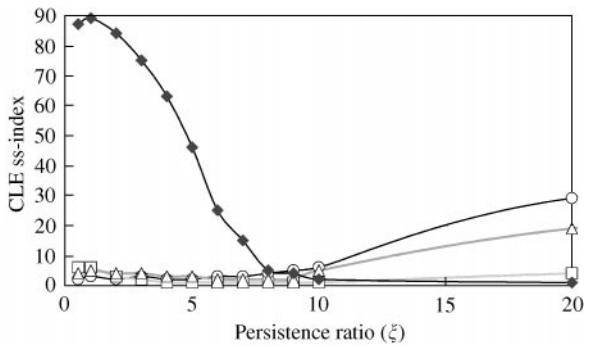

FIG. 4. A plot of the cross-linking entropy (CLE) secondary structure (ss) index as a function of persistence ratio (ξ) for rRNA (1541 nt). Structure # 13 (○), # 79 (□), and # 69 (△) are the top 3 in the CLE ss-index and are some of the best structures predicted by the NNSS algorithm. Structure # 56 (♦) is a single domain ss (1535 nt) that dominates the CLE ss-index at large ξ.

the connect file. A few structures below the CLE-index 12 also contain these same first two domains; however, they also contain other even larger domains in addition to the correct ones.

None of the first 12 MFOLD ss-indices listed in Tables 1 and 2 have the expected domain structure. The optimal structure predicted using the NNSS algorithm showed the following domain lengths: 1140, 8, 15, and 363 nt (Table 1). The domains have the following positions along the pMBL: (9, 1148), (1151, 1158), (1160, 1174), and (1176, 1538) (Table 2).

The cross-linking entropy tended to promote structures possessing domains 1 and 2 to the top of the list in these calculations. This included MFOLD ss-index 13 which became second at the top of the list. The original "optimal" structure was promoted to a CLE ss-index of 87 in the list of the top 100 such suboptimal structures. Structures with long domain lengths (and high BPD) like the optimal structure are typically distributed at the end of the CLE ss-index list. The result is effectively a population inversion.

Figure 4 shows the stability of the CLE ss-index as a function of ξ. MFOLD ss-indices 13 (○), 69 (△) and 79 (□) are stable over the entire range of experimentally reasonable values of ξ ($1 < \xi < 9$). For $\xi > 4.0$, the MFOLD ss-index 13 and 69 gradually climb higher in the suboptimal index list. At the same time, MFOLD ss-index 56 with a single domain comprising 1535 nt rapidly falls to position 1 for $\xi > 4$ (♦). Here, we see striking evidence of functional domain structure size governed by the persistence ratio (ξ). Interestingly, # 79 does not change much over the entire range. In general, as ξ increases, long domains lengths and long stems are greatly favored over short domain lengths and short stems.

3.2. GROUP I INTRONS

A fairly accurate estimation of the group I intron structure for *T. thermophila* can be found in the MFOLD connect files using "-LIS = 150" and "-WIN = 1". The major discrepancies in the structure are at the base of the P4 stem where an additional cross-link is formed. A short stem is found in the region where the P3 stem tertiary structure forms. Likewise, an extra stem and two internal loops are formed at the base of the P7 stem. Extra BPs are found at the base of the P5b and P5c stems connecting the iMBL, at the L6b and L9.1 loops, and at the base of the P9.2 loop. Missing BPs are found at the L2.1 section (Cech, 1988). The MFOLD ss-index is # 7 in the GCG 10.1 connect file outputs. Structures #3 and #4 also show promise but have minor errors at the domain boundary of the P2.1 stem and structure # 4 has some notable differences at P5 where an additional bulge is found. There are also two structures # 33 and # 85 which correspond to a secondary structure attempt at connecting the P3 stem's tertiary structure where the P4–P5abc, and P6 regions are nearly complete but the P7 structure in the iMBL is not found due to conflicts introduced by the secondary structure approximation rules.

It is notable that one of the best domain structures in the MFOLD ss-index (#7) appears in the top ten structures of the MFOLD listing. Two other structures (#3, #4) are also very close in agreement with the experiment. However, intermingled with these excellent fits (#3, #4 and #7), are structures #1, #2, #5, #6, and #8 which contain two huge domains 100 and 205 nt long. Neither of these two domains are strongly suggestive of any naturally occurring structures or even stalled structures of the group I intron. Structure #9 obtains the correct P4–P5abc stem, but misses P6 ~ P9. Structure #10 shows the stable P5abc stem region

(Wu & Tinoco, 1998; Thirumalai, 1998) and splits the two halves of the P4 stem into two additional domains (typical of results for $\xi = 4$).

Using the CLE strategy, MFOLD ss-index #7 fares rather poorly showing a CLE ss-index 35 for $\xi = 9.0$ nt (Table 3), which is a rather large persistence length to currently justify experimentally. Larger values of ξ bring this structure closer to the top, but are probably too large. Therefore, in the current state of development, the CLE strategy clearly has difficulty finding the #7 structure.

However, Table 3 reveals some very important physics of the *T. thermophila* intron (Section 4.3). For $\xi = 4.0$ nt, with the exception of the P4–P5abc region, the majority of the fitted domains are exactly those domains found in MFOLD ss-index #7 (compare the P1, P2, P2.1 columns and the P6 ~ P9 columns in Table 3 with the real structure (labeled "observed"). The stable P5abc structure which is the precursor to the fully formed domain is clearly visible in the $\xi = 4$ list (the 49 and 68 nt segment: Table 3). At the same time, only one fully formed P4–P5abc domain region is listed at CLE ss-index #10 for $\xi = 4.0$ nt. On the other hand, five structures with complete P4–P5abc domain are found for $\xi = 9.0$ nt, yet the P1 ~ P2.1 domains in particular suffer a very poor fit. A quick study of the composition of the P4–P5abc region reveals a considerably different distribution of ACGU compared to the rest of the intron. The P4–P5abc stem region is also the longest domain in the intron (108 nt). Since the CLE currently weights the entropy the same way for *all* domains under the assumption that ξ is constant, it is unlikely that the CLE can find two domains that have vastly different persistence lengths and vastly different domain sizes simultaneously using a single parameter. The results are suggesting that the *T. thermophila* sequence exhibits a particularly variable persistence length which is governed by the GC content. Since GC typically forms much stronger bonds due to the triple H-bond, the P4–P5abc stem is likely to be more stiff (Section 4.3).

It is also noteworthy that *none* of the structures resembling MFOLD ss-index #1 appear anywhere near the top of this list for any choice of ξ. The CLE has again sent the majority of these structures to indices on the order of 100 (out of 150), although some appear in the 50s (at $\xi = 9.0$ nt).

In the composite picture revealed in Table 3, the CLE is finding the correct domains for specified values of ξ. Varying ξ from 4.0 to 20.0 nt, the population of P4–P5abc structures gradually increases while the distinct P1, P2, and P2.1 structures disappear. The remaining domains are well represented in all the structures. Taken in terms of thermodynamic populations, a scan of the structure by varied persistence length passes through all of the vital structures of the *T. thermophila* intron. Again, just like the rRNA evaluation, the results we obtain do not depend on the length of sequence that is input into the NNSS calculation.

Figure 5 compares the stability of the P4–P5abc domain structure listed in Table 3 of MFOLD ss-index 4, 26, and 68 as a function of the persistence length. These are contrasted with structure MFOLD ss-index 53 which shows the opposite trend. Structure 12 appears to be stable throughout the distribution (data not shown). Nevertheless, we clearly see the strong dependence on ξ emerging. In Figure 5, structures #4, #26, and #68 all descend to the range of 1–10 rather rapidly from $\xi = 4.0$ nt. Likewise, structure #53 ascends quite rapidly for $\xi > 4.0$ nt. The results suggest that our *assumption* that persistence length is invariant is not correct (Section 4.3).

These calculations were also verified using the nucleation ($\Delta\overline{\mathcal{G}}^{\mathcal{X}}_{ncl}$) and freeze out ($\Delta\overline{\mathcal{G}}_{frz}$) models for the local CLE. The nucleation results were almost identical to the freeze out results. More subtle differences are observed when applying correlation effects. Again, there was no clear way to ascertain as to which effects dominate the local CLE.

3.3. tRNA

In the $tRNA^{Phe}$ structure, the NNSS algorithm finds the correct D loop, anticodon loop, and acceptor stem; however, the T loop is not correct (Hagerman, 1997). The MFOLD ss-index is # 7 (last in the list). The structure is the well-known "t" shape. MFOLD ss-index 4 also has this "t" shape; however, the fit is less in agreement

TABLE 3

The first 10 *CLE predictions of the domain sizes* [*using* $\xi = 4.0$ *nt and* $\xi = 9.0$ *nt*) *for the self-splicing intron* T. thermophila, *along with the optimal structure predicted by MFOLD* (MF(1)], *and the experimentally determined structure of* T. thermophila (*Cech*, 1988) [(Observed)]*

CLE ss index	MFOLD index	# of domains	Domain boundaries for $\xi = 4.0$ nt					
			P1, P2	P2.1	(P3)	P4–P5abc	P6, P7–P8	P9 ∼ P9.2
1	12	13	32, 26	36	14	21, 49, 29	34, 62	14, 36, 36, 20
2	93	12	32, 26	[59]		11, 49, 29	34, 62	14, 36, 36, 20
3	43	12	32, 26	29		38, 49, 29	34, 62	14, 36, 36, 20
4	132	12	32, 26	36		26, 68, 24	34, 62	14, 36, 36, 20
5	47	12	32, 26	36		26, 68, 24	34, 62	14, 36, 36, 20
6	49	13	32, 26	36	14	17, 49, 29	34, 62	14, 36, 36, 20
7	137	13	32, 26	36		26, 68, 30	21, 20, 42	14, 36, 36, 20
8	115	13	32, 26	36	14	21, 49, 26	34, 62	14, 36, 36, 20
9	53	14	32, 26	36		26, 68, 30	21, 13, 11, 42	14, 36, 36, 20
10	68	11	32, 26	29	12	**110**		
							34, 62	14, 36, 36, 20
	…					…		
112	MF(1)	6	32, 26	36	14	[205]		100
	…					…		
(Observed)		9	32, 26	36	—	**108**	44, 51	16, 36, 34, —
			P3 tertiary structure: 182					
			Domain boundaries for $\xi = 9.0$ nt					
1	26	8	[111]			**110**	34, 62	14, 36, 36, 20
2	93	12	32, 26	[59]		11, 49, 29	34, 62	14, 36, 36, 20
3	12	13	32, 26	36	14	21, 49, 29	34, 62	14, 36, 36, 20
4	68	11	32, 26	29	12	**110**	34, 62	14, 36, 36, 20
5	132	12	32, 26	36		26, 68, 24	34, 62	14, 36, 36, 20
6	4	11	32, 26	29	12	**110**	34, 62	14, 36, 36, 20
7	43	12	32, 26	29		[38], 49, 29	34, 62	14, 36, 36, 20
8	86	8	[111]			**110**	34, 62	14, 36, 36, 20
9	47	12	32, 26	36		26, 68, 24	34, 62	14, 36, 36, 20
10	100	10	67	27	12	**110**	34, 62	14, 36, 36, 20
(Observed)		9	32, 26	36	—	**108**	44, 51	16, 36, 34, —
			P3 tertiary structure: 182					

*The fourth and remaining columns are organized and labeled in terms of the observed domains of *T. thermophila* where "P" (meaning "paired") is the standard notation for stems in the group I intron literature. A comma separating the labels (e.g., P1, P2) indicates stems in separate domains, whereas a dash (–) (e.g., P7–P8) indicates stems that encompass a single domain. The tilde (∼) indicates a group of domains (e.g., P9 ∼ P9.2 implies the P9, P9.1 and P9.2 domains). Label (P3) indicates a stem in the region where the P3 tertiary structure is usually formed. The brackets "[ ]" are used to indicate a region where the domains in two or more columns have overlaps. In these calculations, $\Delta\overline{\mathscr{G}}_{frz} = 0.25$ kcal mol^{-1}. At $\xi = 4.0$ nt, the stable P5abc region is fully formed, but only one structure with a completed P4–P5abc domain is found. On the other hand, at $\xi = 9.0$ nt, five structures with a completed P4–P5abc domain are found. Domain sizes that correspond to the known P4-P5abc domain structure of *T. thermophila* are highlighted in bold text.

than # 7. Two structures on the list form more than one domain (Table 4), the rest are single domain structures. MFOLD ss-index 1 is a straight stem.

For $\xi = 4$, MFOLD ss-index 4 appears at the top and # 7 is fourth on the list (Table 4). The

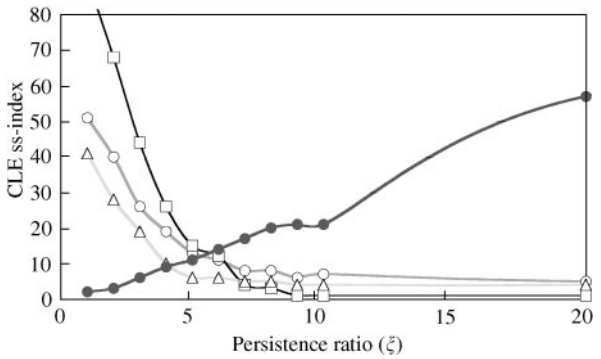

FIG. 5. A plot of the CLE ss-index as a function of persistence ratio (ξ) for the group I intron (433 nt). MFOLD ss-indices #4 (○), 26 (□), and 68 (△) are listed in the top 10 structures of the CLE ss-index at $\xi = 9.0$ nt. These structures contain the complete P4–P5abc domain. MFOLD ss-indices #12 and #53 both appear in the top ten of the CLE ss-index list for $\xi = 4.0$ nt. MFOLD ss-index #53 (●) is typical of structures which have the correct P5abc subdomain and the complete set of P1, P2, and P2.1 domains, but lack the fully formed P4–P5abc structure. The P1, P2, and P2.1 stems are strongly favored at $\xi = 4.0$ and the fully formed P4–P5abc domain is favored at larger ξ. MFOLD ss-index #12 appears to be stable across this spectrum with only a gradual increase (data not shown). The trend of these example structures shows how the value of ξ can greatly influence the distribution of suboptimal structures. Short domains are clearly favored for $\xi < 4$ and longer domains such as the P4–P5abc stem are favored for $\xi > 4$ nt.

calculation used $\Delta\overline{\mathcal{G}}_{frz} = 0.42$ kcal mol^{-1} which leaves the FE in roughly the range of the original sequence when compared with MFOLDs distribution. The distribution of secondary structure is rather stable over the range $4 \leqslant \xi \leqslant 6$, and gradually changes for $\xi > 6$. The domain sizes and distributions for both MFOLD and CLE predictions are essentially the same because only two of the seven predicted structures consist of more than one domain.

The tendency for tRNA shows that the CLE can get roughly similar results as the traditional NNSS approach, and possibly better, although that would require more study. The CLE method functions equally well as traditional NNSS approaches for short sequences such as tRNA (76 nt), nearly as well for intermediate size sequences such as *T. thermophila* (433 nt), and much better for very long sequences such as rRNA (1541 nt). Hence, the CLE performs reasonably well over all decades of sequence lengths in spite of the limitations caused by assuming that persistence length is constant over the entire sequence.

3.4. PERSISTENCE LENGTH AND FUNCTIONAL DOMAIN SIZE

In all the tests that we carried out, the functional domain size appears to be strongly dependent on the persistence ratio (ξ) and the BPD.

Figure 6 shows the average maximum domain size as a function of ξ (○) for shuffled sequences of the *T. thermophila* group I intron. There is

TABLE 4

Results for tRNA^{Phe}. In these calculations, $\Delta\overline{\mathcal{G}}_{frz} = 0.42$ kcal mol^{-1} and $\xi = 4.0$ nt. The organization is the same as in Table 1

CLE ss index	MFOLD index	FE results (kcal mol^{-1})				# of domains	Domain boundary sizes
		$\Delta\mathcal{G}_{ss}$	$\langle\Delta\mathcal{G}_{cl}\rangle$	$\lvert T\Delta S_{ss}\rvert$	$\Delta\mathcal{G}$		
1	4	− 19.60	25.84	28.20	− 21.96	1	**72**
2	3	− 19.90	24.56	26.60	− 21.94	2	41, 31
3	5	− 19.50	24.80	25.70	− 20.40	1	**72**
4	7	− 18.50	22.16	23.00	− 19.34	1	**72**
5	2	− 20.40	27.26	25.60	− 18.74	1	**72**
6	6	− 19.30	23.32	22.60	− 18.58	2	41, 28
7	1	− 20.50	25.84	21.70	− 16.36	1	**72**
tRNA^{phe} (observed)						1	**72**

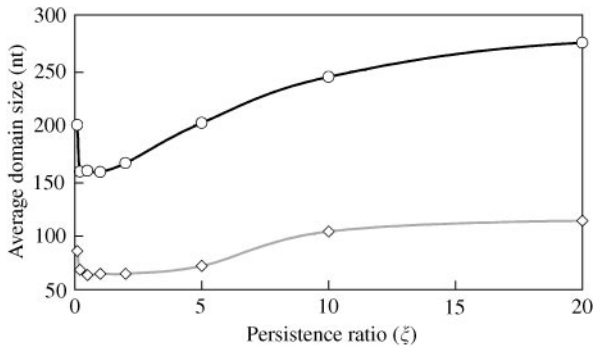

FIG. 6. A plot of the maximum average domain size as a function of the persistence ratio (ξ) for the *T. thermophila* group I intron (433 nt). (○) The average maximum domain size found in the original sequence and 100 shuffled sequences. The average maximum domain size of each sequence was calculated from the top five CLE ss indices listed for each sequence. The distribution of the average maximum domain size was then evaluated for the set of sequences as a function of ξ. (◇) Out of the list of domain size maxima found in these 101 sequences, the smallest maximum domain size of this set: the largest is 429 nt (for all values of ξ).

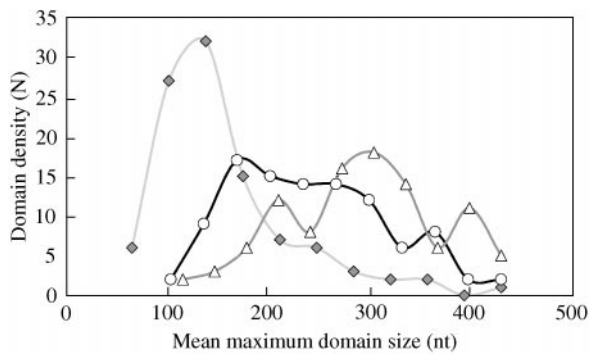

FIG. 7. The distribution of average maximum domain size (from Fig. 6) for various values of ξ based on 100 shuffled sequences of the *T. thermophila* group I intron (433 nt): (◆) $\xi = 1.0$, (○) $\xi = 10.0$ and (△) $\xi = 100.0$. The mean values gradually shift from 157 nt (skew: 1.4) → 243 nt (skew: 0.2) → 295 nt (skew: -0.2) as $\xi = 1 \rightarrow 10 \rightarrow 100$ nt, respectively. The S.D. is about 70 for all these data.

a clear tendency for the domain size to increase with increasing ξ in the shuffled sequences. The *T. thermophila* sequence is rather short (433 nt) which places limits on the growth and extent of the functional domain size. The ◇ indicate the increase in the lower bound of this *maximum* average domain size. The distribution for $\xi = 1$, 10, and 100 are shown in Fig. 7 and a visible lower bound is suggested in the distribution. Whereas the meaning of $\xi = 100$ is somewhat questionable in this context (permitting only four "links" per sequence), our point is to illustrate that the domain size is strongly influenced by ξ. In Fig. 7, a highly skewed normal distribution can be seen for $\xi = 1$ (◆: skew ~ 1.4), a distribution roughly spread over the entire range for $\xi = 10$ (○: skew ~ 0.4), and again returning to a distorted normal distribution for $\xi = 100$ (△: skew ~ -0.2). For large ξ, the constraints on the sequence length are the source of the distortions.

The size of the functional domains can also be seen in Figs 4 and 5 where long domain lengths are favored for large ξ and short domain lengths are favored for small ξ. This shows that ξ tends to weight the domain distribution and therefore the size of the corresponding domains.

3.5. GROUP I INTRON FOLDING KINETICS

The concepts developed from the two state model (Section 2.3) are now applied on the known folding behavior of the group I intron (*T. thermophila*). The outer lying regions of the P4 domain clearly have $\Delta\mathcal{G}_{cl}^{[P4]} \gg \Delta\mathcal{G}_{cl}^{[P5abc]}$. The rates tend to govern which path is favored thermodynamically. Hence, the outer lying regions (P4) will form slower than the inner regions (P5abc and P5) on average. This is even more the case for the P3 region which encompass 182 nt ($\Delta\mathcal{G}_{cl}^{[P3]} \gg \Delta\mathcal{G}_{cl}^{[P4]}$). Likewise, the P1 ~ P2.1 domains and the P6 ~ P9 domains are shorter. Since the domains will tend to fold independently, these other domains are also likely to form early.

From the experimental folding data, the catalytic core region that encompasses the P4 ~ P6 domains is known to form before the appearance of the P3 domain (Pan & Woodson, 1999). This is consistent with the expectations of cross-linking entropy in which the most thermodynamically probable pathway for RNA folding begins from the base of a given loop and works its way toward the 5′-3′ end of a given domain. (The separate loops will tend to form independently.) The P3 domain encompasses 182 nt and is unlikely to nucleate at the 5′-3′ end first; hence, most of the secondary structure will form first. However, the P4–P5abc domain is quite long (108 nt) and the other domains are typically less than 50 nt long (Table 3). This would permit the tertiary structure involving the peripheral domains to form *before* the catalytic core is completed

the appearance of the P13 and P14 tertiary structures preceed formation of the P3 stem: see Pan & Woodson (1999) and references therein).

Hence, even with an extremely rudimentary two state model for the *T. thermophila* in our hands, we can "visualize" the experimental folding data of the group I intron without addressing a computer. The NNSS algorithm cannot make these predictions because they assume that both distant and proximal bonds have the same probability of forming given that all other terms are essentially the same.

3.6. FITTING OF LOCAL CLE

In the CLE evaluations, all the original NNSS penalties were removed. To estimate $\Delta\overline{\mathscr{G}}_{frz}$ and $\Delta\overline{\mathscr{G}}^{\mathscr{X}}_{ncl}$, tRNA^{phe} was used to "tune" the parameters used here. We assume that the NNSS results are in sufficient agreement with specific heat values for tRNA to be used as a standard. To start with, eqn (17) was evaluated with $\Delta\overline{\mathscr{G}}_{frz}$, $\Delta\overline{\mathscr{G}}^{\mathscr{X}}_{ncl}$ and $\langle\Delta\overline{\mathscr{G}}_{corr}\rangle$ set to 0. The unchanged NNSS prediction ($\Delta\mathscr{G}_{ss}$) was also obtained and the difference between the two values was determined.

The freezing out penalty ($\Delta\overline{\mathscr{G}}_{frz}$) was obtained by taking the difference between the NNSS prediction and the CLE prediction and dividing the result by the number of BPs. For $\xi = 3$ nt, $\Delta\overline{\mathscr{G}}_{frz} \sim 0.25$ kcal mol^{-1} at 310 K.

Nucleation estimates $\Delta\overline{\mathscr{G}}^{\mathscr{X}}_{ncl}$ (where $\mathscr{X} \equiv \mathscr{B}$, $\mathscr{H}$, $\mathscr{I}$ and iMBL) were handled in the same way except that the divisor was the sum of the number of $\mathscr{B}$s, $\mathscr{H}$s, $\mathscr{I}$s and iMBLs in the final secondary structure. For $\xi = 3$ nt, $\Delta\overline{\mathscr{G}}^{\mathscr{X}}_{ncl} \sim 1$ kcal mol^{-1} at 310 K.

These same local CLE-FE corrections for $\Delta\overline{\mathscr{G}}^{\mathscr{X}}_{ncl}$ and $\Delta\overline{\mathscr{G}}_{frz}$ were then used on the *T. thermophila* and rRNA secondary structures. After adding the local CLE-FE corrections, the total FE values were nearly identical in their FE values as the original NNSS predictions ($\Delta\mathscr{G}_{ss}$). Indeed, we can cover the entire spectrum of sequence length for random structure using these default values for $\Delta\overline{\mathscr{G}}^{\mathscr{X}}_{ncl}$ and $\Delta\overline{\mathscr{G}}_{frz}$ (with $\xi = 3$) and obtain a similar range for the FE distributions.

This shows that $\Delta S^{c\to f}_{\mathscr{X}_o}$ [eqn (16)] results from a local CLE effect because $\Delta\overline{\mathscr{G}}_{frz}$, $\Delta\overline{\mathscr{G}}^{\mathscr{X}}_{ncl}$ and $\langle\Delta\overline{\mathscr{G}}_{corr}\rangle$ all show stable predictions over the entire spectrum of sequence length based upon a single fit using tRNA^{phe}. The only issue we currently cannot resolve is whether the local CLE is due to nucleation, freezing out, or some combination of the two, and to what extent we need to account for correlation in this problem.

4. Discussion

Several observations have emerged from this study. The CLE reveals size limits on functional domain and provides us with a helpful tool for finding those domains. The major parameters governing domain size are ξ and the BPD. Most of the biologically active RNA is near its thermodynamic equilibrium state; however, there are some indications of $5' \to 3'$ synthesis effects that reveal branching in the secondary structure predictions. The CLE provides a useful tool for understanding the folding dynamics of RNA.

At present, only a few example structures have been studied. Nevertheless, the results show that the theory can aid us in finding these domains. All of the structures were examined together, yet the dominant structures that appeared at the top of the CLE list were exactly those structures that agree with the functional domains of the biologically active structures. In the study of unknown RNA structures of very long sequences (Dawson & Yamamoto, 1999), this latter observation is certainly encouraging. At the same time, a more complete study of the CLE is certainly necessary.

4.1. OBSERVED LIMITS ON FUNCTIONAL DOMAINS SIZES IN NUCLEIC ACIDS

A scan of the known RNA structures suggest that most RNA-based functional domains do not exceed 500 nt. The group I intron ribozymes have a sequence length of about 400 nt (Damberger & Gutell, 1994; Pan & Woodson, 1999). In effect, group I intron never exceeds this domain size limit. For example, the *T. thermophila* has 414 nt in the sequence (Cech, 1988) and the structure is sometimes described as having two major helical domains: P4–P6 and P3–P9. The P3 stem actually closes tertiary structural features (Cech, 1988). Without the tertiary structure, there are 9 domains (Table 3): the largest being the P4–P5abc

stem (108 nt). The P3 domain encloses a domain of 182 nt: positions (105, 283) or (100, 278) using the notation by Cech (1988). Group II introns can extend up to 3000 nt in length, and they form 6 domains (Michel & Ferat, 1995; Michel *et al*., 1989). Moreover, the largest domains of group II introns are rarely longer than 500 nt. The 16S rRNA family appears to build roughly 3 general domains (I, II and III) out of sequences on the order of 1500 nt (Woese *et al*., 1980; Glotz *et al*., 1980; Mueller *et al*., 2000). Likewise, the 23S rRNA family generally forms six large functional subdomains out of sequences of the order of 3000 nt (Gutell *et al*., 1993).‡‡ Each of the domains has a length of the order of 500 nt. Hence, whereas there are no hard and fast rules, there appear to be size limits on function domains that often emerge *around* 500 nt.

There are also some exceptions to this rule. In the Simian virus 40 late pre-mRNA, several very long sequences were reported (Nussinov *et al*., 1982). In the Qβ virus, there is evidence that there are very long sequences that form (approximately 1600 nt) (Jacobson & Zuker, 1993; Skripkin & Jacobson, 1993). Likewise, some group II introns have coding regions that greatly exceed 500 nt in domain 4. The 23S rRNA subunit (*E. coli*) also closes at its 5′-3′ ends. Likewise, twintrons may close large domains. We propose five alternative explanations with respect to this theoretical model: (1) the persistence length in the respective regions is *enormous* compared to that of the free segment (i.e. the RNA structure is effectively "crystalline"); (2) the BPD is relatively low in the domain (as is the case for 23S rRNA of *E. coli*), (3) the final structure links together a variety of tertiary structure, (4) there are a variety of protein binding interactions that greatly extend the length of these domains, and (5) ribozymes such as the group II intron and twintrons may require additional unresolved steps in the splicing process (including (4)). Cases (1) and (2) can be tested with the current theory; case (3) should still conform to BPD and ξ requirements; and cases (4) and (5) extend beyond the current treatment. Often, a virus must pack into a small capsid, which would lend support for (1). Likewise, proteins are often associated with group II introns. Group II introns may also require a flexible coding regions suggesting a low BPD. At any rate, it is important to remember that domain size can be large but only if the BPD is small, ξ is large, or there is additional "help" from somewhere.

‡‡This has some important exceptions: the most notable is *E. coli* 23S. However, the maximum MBL hierarchal complexity (HC) of 23S is only of order 5 (Section I-2.1). The first-order iMBL is an *enormous* loop and the majority of the higher-order iMBLs are also quite large. The result is that the BPD (Section I-2.1) is not so far from the average. An enormous loop is *easily* accommodated if only a few BP close the entire region. None of this actually contradicts the CLE predictions, it merely makes generalizations about domain size dependent on several factors.

4.2. EQUILIBRIUM AND NON-EQUILIBRIUM FORMATION OF FUNCTIONAL DOMAINS

There are essentially two reasonable scenarios in which the length of a functional domain of an RNA sequence would exhibit limits. One possibility is that the folding process itself yields the limits on domain size of a biologically active structure and that functional domains are inherently metastable. Another possibility is that the native state is fairly close to thermodynamic equilibrium and some other effect (such as CLE) is causing these limits.

First, even *given* that a non-equilibrium structure is formed, stable equilibrium structures must strongly influence the folding process; otherwise, the possibility of misfolding or rapid decay into a more thermodynamically stable structure (which is non-functional) becomes a serious problem for biologically active RNA. In experiments conducted on complementary sequences of the group I intron (*T. Thermophila*), the introns eventually folded into the native state when sufficient Mg^{2+} was added (if the sequences were not drastically mutated; (Pan & Woodson, 1999). This would suggest that the native state of catalytic group I introns is essentially quite close to the thermodynamic equilibrium structure.

For all the model RNA structures studied in this work, the CLE pulled out structures that are more characteristic of natural sequences (i.e. short domain structures) suggesting that CLE limits the domain size. Qualitatively, the cost of forming a functional domain grows as $N\ln(\xi N)/\xi$. The size limit is weighted most heavily by the length of the domain and the BPD;

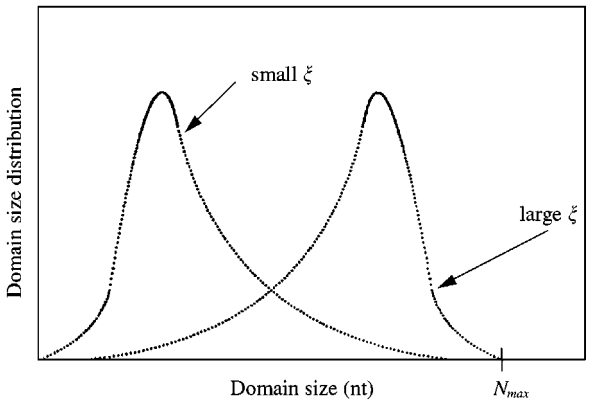

FIG. 8. A schematic of the distribution of functional domain sizes for two extreme values of ξ: small ξ and large ξ (extrapolated from Figs 6 and 7). N_{max} indicates the sequence length of some arbitrarily long sequence. If $\xi \sim N_{max}$, then the large ξ condition is met. Typical distributions for real structures should fall somewhere in between.

however, the extent of crystallization (large ξ: Fig. I-2) tends to reduce the magnitude of this effect at the expense of flexibility (small ξ: Fig. I-3). It can be seen in Fig. 6 that the average maximum domain size increases as a function of ξ. Likewise, the distribution for the lower bound of the average maximum domain size also increases with increasing ξ. As shown in Fig. 7, the skew in the distribution tends to be large for extreme values of ξ ($\xi = 1, 100$) and small for intermediate values ($\xi = 10$). This is qualitatively expressed by the dotted lines in Fig. 8. From this, we can conclude that in the absence of cross-linking entropy ($\xi \to \infty$), there is no selectivity on the domain size and all the sequence space is available for domain formation with a preferential selection of the longest domain lengths and the longest stems. The limit on the domain size becomes the sequence length itself. As seen in Figs 4 and 5, the longer domain lengths are preferred at larger ξ (squares: Fig. 4; open symbols: Fig. 5).

In this view, when a sequence gradually relaxes into its native functional state, it is blocked from folding into very long functional domains as a result of the cross-linking entropy. The size of the functional domain is then governed by the characteristics of the persistence length. If the persistence length is long, then the structure is likely to form larger domains with long stems. Conversely, if the persistence length is short, then only short domains and short stems are likely to form. This is strongly suggested in the behavior of the average maximum domain size of shuffled sequences of *T. thermophila* (Fig. 6). Standard NNSS algorithms only take into account the nearest-neighbor interactions; hence, they do not account for this phenomena.

There may be some important examples of non-equilibrium structures suggested in this study. For the 16S rRNA (*E. Coli*) structure (Section 3.1: Tables 1 and 2), the NNSS program could not even produce the correct structures for domains $3 \sim 7$ without first cutting the sequence. Since all "possible" structures that are thermodynamically stable *should* appear on the list of potential structures, the inability to create reasonable approximations of the correct structure may be a problem related to the non-equilibrium conditions under which the rRNA is synthesized. A way of testing this experimentally would be to denature the rRNA and check that the renatured structure matches one of the candidates listed in the NNSS suboptimal structure solutions. If the renatured product resembles the domains of MFOLD ss 13 for example, then we can be certain that these domains are a result of non-equilibrium processes that occur during the $5' \to 3'$ synthesis of rRNA.

Hence, whereas it is quite likely that there are a variety of metastable structures that exist in the world of biology; in light of the CLE, it is not so clear that such a class of structures is normative, nor is it apparent that when such structures are present, they are extremely far from the equilibrium structures in FE.

4.3. GLOBAL CLE EFFECTS: VARIATIONS IN THE PERSISTENCE LENGTH

The results of rRNA showed a clear improvement over the original FE distribution and were fairly stable for a wide range of values of ξ. From this, we can infer that rRNA (*E. coli*) and $tRNA^{Phe}$ may actually have a fairly invariant persistence length (at least in the properly fitted domains found in this work).

This "invariance" was clearly not the case for the *T. thermophila* results where two different distributions are found for $\xi = 4$ and 9. The

P4–P5abc domain appears to be quite sensitive to the choice of persistence length. Even the MFOLD ss 26 which appears at the top of the list at $\xi = 9.0$ is strongly suppressed at $\xi < 4.0$ (Fig. 5).

In the results of the *T. thermophila* group I intron ribozyme, the favored equilibrium structures (Table 3) all lack the P3 domain. Likewise, for $\xi = 4$, the P5abc subdomain and all of the other domains except the P4–P5 stem region of the P4–P5abc domain are fully formed and at the top of the list. Physically, this suggests that secondary structure resembling Table 3 for $\xi = 4$ forms rather early followed by completion of the P4–P5 stem. It is known that the P5abc part of the domain is quite stable and forms quite early in the folding process (Wu & Tinoco, 1998; Thirumalai, 1998). The structures of size 49 and 68 nt in the P4–P5abc column of Table 3 ($\xi = 4$) are all representative of this structure. Meanwhile, the peripheral tertiary structure has time to fold. Finally, the P3 domain completes the tertiary structure of the catalytic core of the intron. All the major components of the catalytic core are present in the top 10 CLE ss indices except the P3 segment. It was mentioned in (Section I-2.2.1) that higher-order folding is possible in flexible structures (Fig. I-3). This is all consistent with a kinetic model incorporating CLE (Sections 2.3 and 3.5).

The P4–P5abc domain is likely to be quite stiff compared to the rest of the domains in *T. thermophila*. The majority of bonds in the P4–P5abc domain consist of GC and GU stacking and nearly equal percentages of purine (A: 28.7%, G: 30.6%), and pyrimidine (C: 21.3%, U: 19.4) bases. On the other hand, the segment comprising the P1, P2, and P2.1 domains contains mostly AU stacking and a predominance of AU in the sequence (A: 32.7%, C: 18.3%, G: 19.4%, and U: 29.6%). The overall distribution is (A: 29.9%, C: 18.7%, G: 25.4%, and U: 26.0%); hence, the distribution in the P4–P5abc domain is quite rich in GC and GU pairing. Due to the loss in rotational degrees of freedom, the triple hydrogen bond in the GC is more "stiff" compared to the double H-bond of AU or the single H-bond of GU (Searle & Williams, 1993). This would suggest that the P4–P5abc domain should also be less flexible compared to the surrounding AU-rich domains. When CLE observations are combined with considerations about coaxial stacking (Mathews *et al*., 1999; Holbrook & Kim, 1997), there is reason to presume that these separate domains will form since polyelectrolytic effects related to the GC/AU rich regions (Grosberg & Khokhlov, 1994) would tend to separate the P4–P5abc domain from the other domains in the structure.

Metal ions often occupy the internal loops and bulges (Hermann & Patel, 1999) and would tend to lead to a hardened structure with a long persistence length in a structure like group I intron. The metallic ions tend to stabilize the secondary structure and at the same time, the ions tend to make the structure less flexible due to the ionic bonding that forms. The P4–P6 domain regions have been reported to have 24 metallic ions within the structure (Holbrook & Kim, 1997; Hermann & Patel, 1999). At least some of these are occupying the interstitial regions of the stems (Tinoco & Bustamante, 1999) and tend to "crystallize" the P5abc region (Wu & Tinoco, 1998). The incorporation of water ions also tends to "harden" the structure (Hermann & Patel, 1999; Holbrook & Kim, 1997). Again, the reductions in degrees of freedom will make such "free" segments more stiff. Hence, the regions occupied by metallic ions or coordinated water molecules will tend to behave more like "stems" than free segments. The P4–P5abc region is sometimes thought to form the "scaffolding" for group I intron.

We recognize that the value we use ($\xi = 9.0$) may be too large. This seems to mostly reflect our naive assumption that ξ is invariant in ssRNA. Indeed, we should have expected a variable persistence length in *all* these structures. Nevertheless, this naive assumption in itself has also revealed some significant physics in biologically active structures which is rarely discerned except by way of experiment and certainly not understandable from an NNSS stand point. In Section 2.1 we have shown that a variable ξ can be modeled into the CLE at the resolution scale of a link. Hence, at the scale of domains, we can surely vary our monolithic ξ and this is not speculation.

Finally, it was mentioned in Part I (Sections I-2.2 and I-3.1) that the loop region may have

a smaller persistence length than the stem. Section 2.1 shows that this can be accommodated in the theory. However, at least for $\xi \sim 3$ nt, there currently does not appear to be any necessity to "adjust" these values. This issue may be more relevant when ξ is very large in the stem regions. A variable ξ will be considered in all future work.

Ultimately, the context dependence of ξ needs to be measured. One way is via differential melting curves (Liang & Draper, 1994; Brion Westhof, 1997). Another more recent and promising route is atomic force microscopy (AFM) (Rief *et al*., 1999; Essevaz-Roulet *et al*., 1997). Considerable work has been done on proteins using this technique (Mueller *et al*., 1999). Currently, only very general measurements of ssDNA (and proteins) have been studied. There are also experimental problems related to the limits of resolution for current AFM spectrometers.

4.4. AN ORIGIN FOR FUNCTIONAL DOMAINS IN THE RNA WORLD

It was pointed out in Section 1 that any model of the RNA world (Noller, 1999; Tomizawa, 1993; Turner & Bevilacqua, 1993; Wyatt *et al*., 1993; Volkenstein, 1994) should show a strong dependence on the equilibrium thermodynamics of functional domains. Unlike modern organisms which might conceivably utilize proteins or ribonucleoproteins to help stabilize their functional domains in metastable structures, it is important to recognize that this status of affairs is not so reasonable for the first RNA structures of the RNA world.

These ancestral structures would not have a full range of proteins to service them. Moreover, catalytic processes would have been more likely to occur on time-scales that would have allowed equilibrium conditions to dominate the processes. In such an environment, the evolution of the RNA from shorter sequences to longer sequences (Noller, 1999) would have required long-range coding strategies to successfully lock in a functional domain. In the formation of poly(tRNA) domains (Noller, 1999), the poly(tRNA) would only yield an enormous bramble of secondary structure if no limits on domain size exist. On the other hand, if the cross-linking effect is present, the window size of long-range pairing required for the maintenance of function domain structures is greatly reduced even in near equilibrium conditions (Figs 6–8). In appealing to CLE effects, it is more conceivable that extrapolations from these much shorter structures (Fontana & Schuster, 1998; Noller, 1999; Huynen *et al*., 1993) to much longer functional sequences with multiple domain structures *could* have formed, particularly if the ancestral RNA structures were derived from smaller units such as tRNA (Noller, 1999). The cross-linking entropy permits more reliable domain segmentation and corruption of existing structure is less probable from such combined structures (Section 4.2).

As shown in Part I, the theory also predicts locomotive properties in iMBLs which suggests in part how RNA can do work. However, the source of the precursor engines (analogous to ATP motors) and the subsequent transition to the current ATP engines must be explained. Without ATP engines or some catalysts, such RNA engines would only be capable of a one time operation.

The results of this work suggest that current native functional domain structures still depend primarily on equilibrium thermodynamic conditions. However, non-equilibrium factors such as the $5' \rightarrow 3'$ synthesis of RNA is likely to influence the type of allowed branching of rRNA structures particularly if the energy differences are small between an "optimal" structure and the actual biologically active RNA. In addition, it is likely that a variety of "chaperones" have developed which "aid" in the folding of functional domains. This is well known from protein chaperones (Hlodan & Hartl, 1994). An analogy to this phenomena may appear in the case of alternative splicing where a variety of splicing factors are needed to select a particular splicing pattern (Manley & Roland, 1996; Smith *et al*., 1989; Puig *et al*., 1999; Zhang *et al*., 1999).

5. Conclusions

It is now clear that double-stranded RNA (or DNA), and folded single-stranded RNA (or DNA) should not be treated too liberally as the same thing in drawing conclusions and extrapolations from thermodynamic data. Whereas

the major features of these two systems have some common and important similarities, there are also some striking differences.

The limits on the size of functional domains is at least in part a consequence of the cross-linking entropy. These entropic effects are associated with the freezing out of the total number of degrees of freedom that are available to an N particle system. The CLE weight for a given domain size is a function of the persistence length (the degree of "stiffness") and the relative density of cross-links in the folded structure.

Limits on the domain size would suggest ways in which the current nearest-neighbor secondary structure calculations could be speeded up. Since the domain size is not infinite, it suggests that a cut off is acceptable. This cut off can be easily estimated from the CLE strategies. With improved prediction strategies for secondary structures, molecular dynamics simulations on large molecules become more tractable.

The traditional loop, bulge, internal loop, and multibranch loop penalties used in nearest-neighbor secondary structure algorithms are actually an averaged set of parameters for typical sequences of ACGU whose length is somewhere around 100 nt. With further development, the cross-linking entropy model can be used to estimate these same penalties for sequences of any length, base composition, and base distribution. The primary effect governing RNA structure is the persistence length.

The mathematical formalism developed in this work provides a way of modelling the folding dynamics of RNA and to do so using more realistic models for the persistence length. The model is sufficiently robust to follow group I intron folding under reversible conditions. Results of the cross-linking entropy on known structures of RNA also suggests that at least *some* of the biologically active RNA appears to be quite close to its thermodynamic equilibrium structure. This is particularly important to our understanding of how functional domain structures can be designed. It may also help in explaining how functional domains evolved from an RNA world.

Perhaps most remarkable is that even with the crude and coarse-grained strategy we were forced to employ in this current work, the physics of RNA was revealed through a simple model and RNA secondary structure prediction for long sequences showed some encouraging improvements.

As in the first part of this series, we do not expect everyone to agree with us on all the issues presented in this work. Nevertheless, we have benefited from the advice and comments of the following people. We graciously thank Prof. M. Doi (Nagoya University) and Prof. Schuster (Vienna group) for their insights which have certainly helped us to pull this work together. We also kindly thank Prof. Zuker for exposing some weaknesses in our arguments and encouraging us to look in more detail at the correlation effects. Dr Roger Ruber (Uppsala University) kindly granted his much needed technical assistance on OS and LATEX problems: he and Dr Yasuhiro Futamura (Tokyo University) both provided helpful comments on the manuscript. We also thank Yucong Zhu for her encouragement. Research was supported in part by a fellowship JISTEC, the SMF, and MTB. We also extend our gratitude to the staff at the National Institute of Infectious Diseases.

REFERENCES

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

BOYLE, J., ROBILLARD, G. T. & KIM, S.-H. (1980). Sequential folding of transfer RNA: a nuclear magnetic resonance study of successively longer tRNA fragments with a common 5′ end. *J. Mol. Biol.* **139,** 601–625.

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*, (Gestland, R. E., Cech, T. R. & Atkins, J. F., eds), 2nd Edn., Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

CECH, T. R. (1988). Conserved sequences and structures of group I introns: building an active site for RNA catalysis —a review. *Gene* **73,** 259–271.

CHEN, J-H., LE, SH-Y, & MAIZEL, J. V. (2000). Prediction of common secondary structures of RNAs: a genetic algorithm approach. *Nucl. Acids Res.* **28,** 991–999.

DAMBERGER, S. H. & GUTELL, R. R. (1994). A comparative database of group I intron structures. *Nucl. Acids Res.* **22,** 3508–3510.

DAWSON, W. K., SUZUKI, K., YAMAMOTO, K. (2001). A physical origin for functional domain structure in nucleic acids as evidenced by cross-linking entropy: I. *J. theor. Biol.* **213,** 359–386.

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

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

ESSEVAZ-ROULET, B., BOCKELMANN, U. & HESLOT, F. (1997). Mechanical separation of the complimentary strands of DNA *Proc. Natl Acad. Sci.* **94,** 11935–11940.

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.

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

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

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

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.

FRESCO, J. R., ADAINS, A., ASCIONE, R., HENLEY, D. & LINDAHL, T. (1966). Tertiary structure in transfer ribonucleic acids. *Cold Springs Harbor Symp. Quant. Biol.* **31,** 527–537.

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.

GRALLA, J. & CROTHERS, D. M. (1973a). Free energy of imperfect nucleic acid helices. II. Small hairpin loops. *J. Mol. Biol.* **73,** 497–511.

GRALLA, J. & CROTHERS, D. M. (1973b). Free energy of imperfect nucleic acid helices III. Small internal loops resulting from mismatches. *J. Mol. Biol.* **78,** 301–319.

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

GULTYAEV, A. P., VAN BATENBURG, F. H. D. & PLEIJ, C. W. A. (1995). The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.* **250,** 37–51.

GUTELL, R. R., GRAY, M. W. & SCHNARE, M. N. (1993). A compilation of large subunit (23S and 23S-like) ribosomal RNA structures: 1993. *Nucl. Acids. Res.* **21,** 3055–3074.

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

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

HLODAN, R. & HARTL, F. U. (1994). How the protein folds in the cell. In: *Mechanisms of Protein Folding* (Pain, R. H. ed), pp. 194–228. Oxford: Oxford University Press.

HOFACKER, I. L. (1998). RNA secondary structures: a tractable model of biopolymer folding. In: *Workshop on Monte Carlo Approach to Biopolymers and Protein Folding*, pp. 171–182. Singapore: World Scientific Publishing.

HOFACKER, I. L., FONTANA, W., STADLER, P. F., BONHOEFFER, S., TACKER, M. & SCHUSTER, P. (1994a). Fast folding and comparison of RNA secondary structures. *Monats. Chem.* **125,** 167–188.

HOFACKER, I. L., FONTANA, W., STADLER, P. F. & SCHUSTER, P. (1994b). The Vienna Package. "http://www.tbi.univie.ac.at/ivo/RNA/". (Free Software). (cited in Hofacker, 1998.)

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

HUYNEN, M., GUTELL, R. & KONINGS, D. (1997). Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.* **267,** 1104–1112.

HUYNEN, M. A., KONINGS, D. A. M. & HOGEWEG, P. (1993). Multiple coding and the evolutionary properties of RNA secondary structure. *J. theor. Biol.* **165,** 251–267.

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

JACOBSON, A. B. & ZUKER, M. (1993). Structural analysis by energy dot plot of a large mRNA. *J. Mol. Biol.* **233,** 261–269.

JAEGER, J. A., TURNER, D. H. & ZUKER, M. (1990). Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymol.* **183,** 281–306.

JAEGER, J. A., TURNER, D. H. & ZUKER, M. (1989). Improved predictions of secondary structures for RNA. *Proc. Natl Acad. Sci. U.S.A.* **86,** 7706–7710.

LAING, L. G. & DRAPER, D. E. (1994). Thermodynamics of RNA folding in a conserved ribosomal RNA domain. *J. Mol. Biol.* **237,** 560–576.

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

MANLEY, J. L. & ROLAND, T. (1996). SR proteins and splicing control. *Genes Dev.* **10,** 1569–1579.

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.

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

MICHEL, F. & FERAT, J.-L. (1995). Structure and activities of group II introns. *Annu. Rev. Biochem.* **64,** 435–461.

MICHEL, F., UMESONO, K. & OZEKI, H. (1989). Comparative and functional anatomy of group II catalytic introns —a review. *Gene* **82,** 5–30.

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.

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* 50S ribosomal subunit based on a cryo-electron microscopic reconstruction at 7.5 Å resolution. *J. Mol. Biol.* **298,** 35–59.

NAKAYA, A., YONEZAWA, A. & YAMAMOTO, K. (1996). Classification of RNA secondary structures using the techniques of cluster analysis. *J. theor. Biol.* **183,** 105–117.

NOLLER, H. F. (1999). On the origin of ribosome coevolution of subdomains of tRNA and rRNA. In: *The RNA World*, (Gesteland, R. E., Cech, T. R. & Atkins, J. F., eds),

2nd Edn. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

NOTREDAME, C., O'BRIEN, E. A. & HIGGINS, D. G. (1997). RAGA: RNA sequence alignment by genetic algorithm. *Nucl. Acids Res.* **25,** 4570–4580.

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.

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.

PIPAS, J. & MCMAHON, J. (1975). Method for predicting RNA secondary structure. *Proc. Natl Acad. Sci. U.S.A.* **72,** 2017–2021.

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

PUIG, O., GOTTSCHALK, A., FABRIZIO, P. & SERAPHIN, B. (1999). Interaction of the U1 snRNP with nonconserved intronic sequences affects 5′ splice site selection. *Genes Dev.* **13,** 569–580.

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

RIVAS, E. & EDDY, S. R. (2000). Secondary structure alone is generally not statistcally significant for the detection of noncoding RNAs. *Bioinformatics* **16,** 583–605.

SALSER, W. (1977). Globin mRNA sequences: analysis of base pairing and evolutionary implications. *Symp. Quantitative Biol.* **42,** 987–1004.

SANTALUCIA JR., J. & TURNER, D. H., (1998). Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers* **44,** 309–319.

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.

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.

SKRIPKIN, E. A. & JACOBSON, A. B. (1993). A two-dimensional model at the nucleotide level for the central hairpin of coliphage Qβ RNA. *J. Mol. Biol.* **233,** 245–260.

SMITH, C. W. J., PATTON, J. G. & NADAL-GINARD, B. (1989). Alternative splicing in the control of gene expression. *Annu. Rev. Genet.* **23,** 527–577.

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.

TINOCO, I., UHLENBECK, O. & LEVINE, M. (1971). Estimation of secondary structure in ribonucleic acids. *Nature* **230,** 362–367.

TOMIZAWA, J.-I. (1993). Evolution of functional structures of RNA. In: *The RNA World* (Gesteland, R. E. and Atkins, J. F., eds), pp. 419–445. Cold Springs Harbor: Cold Springs Harbor Laboratory Press.

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

TURNER, D. H. & BEVILACQUA, P. C. (1993). Thermodynamic considerations for evolution by RNA. In: *The RNA World* (Gesteland, R. E. and Atkins, J. F., eds), pp. 465–596. Cold Springs Harbor: Cold Springs Harbor Laboratory Press.

VAN BATENBURG, F. H. D., GULTYAEV, A. P. & PLEIJ, C. W. A. (1995). An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. theor. Biol.* **174,** 269–280.

VOLKENSTEIN, M. V. (1994). *Physical Approaches to Biological Evolution.* New York: Springer Verlag.

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.

WOESE, C. R., MAGRUM, L. J., GUPTA, R., SIEGEL, R. B., STAHL, D. A., KOP, J., CRAWFORD, N., BROSIUS, J., GUTELL, R., HOGAN, J. J. & NOLLER, H. F. (1980). Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucl. Acids Res.* **8,** 2275–2293.

WU, M. & TINOCO JR., I. (1998). RNA folding causes secondary structure rearrangement. *Proc. Natl Acad. Sci. U.S.A.* **95,** 11 555–11 560.

WUCHTY, S., FONTANA, W., HOFACKER, I. L. & SCHUSTER, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers.* **49,** 145–165.

WYATT, J. R. & TINOCO, JR., I. (1993). RNA structure elements and RNA function. In: *The RNA World* (Gesteland, R. E. 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.

YAMAMOTO, K. & YOSHIKURA, H. (1986). Relation between genomic and capsid structures in RNA viruses. *Nucl. Acids Res.* **14,** 389–396.

ZHANG, D. & ROSBASH, M. (1999). Identification of eight proteins that cross-link to pre-mRNA in the yeast commitment complex. *Genes Deve.* **13,** 581–592.

ZUKER, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science* **244,** 48–52.

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/".)