SIAM REVIEW
Vol. 49, No. 1, pp. 65–88

Thermodynamic Analysis of Interacting Nucleic Acid Strands*

Robert M. Dirks†
Justin S. Bois‡
Joseph M. Schaeffer§
Erik Winfree¶
Niles A. Pierce‖

Abstract. Motivated by the analysis of natural and engineered DNA and RNA systems, we present the first algorithm for calculating the partition function of an unpseudoknotted complex of multiple interacting nucleic acid strands. This dynamic program is based on a rigorous extension of secondary structure models to the multistranded case, addressing representation and distinguishability issues that do not arise for single-stranded structures. We then derive the form of the partition function for a fixed volume containing a dilute solution of nucleic acid complexes. This expression can be evaluated explicitly for small numbers of strands, allowing the calculation of the equilibrium population distribution for each species of complex. Alternatively, for large systems (e.g., a test tube), we show that the unique complex concentrations corresponding to thermodynamic equilibrium can be obtained by solving a convex programming problem. Partition function and concentration information can then be used to calculate equilibrium base-pairing observables. The underlying physics and mathematical formulation of these problems lead to an interesting blend of approaches, including ideas from graph theory, group theory, dynamic programming, combinatorics, convex optimization, and Lagrange duality.

Key words. DNA, RNA, equilibrium, base pair, secondary structure, partition function, minimum free energy, multiple strands, dynamic programming, redundancy, distinguishability, symmetry, overcounting, dilute solution, convexity, duality

AMS subject classifications. 82B05, 80A50, 92E10, 05C62, 20B99, 90C39, 05A15, 90C25, 90C46

DOI. 10.1137/060651100

1. Introduction.

DNA is the primary genetic storage medium for life. RNA plays a more varied role in biology, participating in storage, regulation, catalysis, and synthesis [45]. The unique structural properties of nucleic acids (DNA and RNA)

*Received by the editors January 27, 2006; accepted for publication (in revised form) March 30, 2006; published electronically January 30, 2007. The first and second authors contributed equally to this work. This work was supported by grants NSF-CNS-PECASE-0093486, NSF-EIA-0113443, NSF-DMS-0506468 (IMAG), NSF-ACI-0204932, and NSF-CCF-CAREER-0448835, the Charles Lee Powell Foundation, and the Ralph M. Parsons Foundation.

http://www.siam.org/journals/sirev/49-1/65110.html

†Department of Bioengineering, California Institute of Technology, Pasadena, CA 91125 (dirks@caltech.edu).

‡Department of Chemical Engineering, California Institute of Technology, Pasadena, CA 91125 (bois@caltech.edu).

§Department of Computer Science, California Institute of Technology, Pasadena, CA 91125 (schaeffer@dna.caltech.edu).

¶Departments of Computer Science and Computation & Neural Systems, California Institute of Technology, Pasadena, CA 91125 (winfree@caltech.edu).

‖Departments of Applied & Computational Mathematics and Bioengineering, California Institute of Technology, Pasadena, CA 91125 (niles@caltech.edu).

make them attractive materials for engineering nanoscale structures and devices. By appropriately designing the sequence of bases in each strand, synthetic nucleic acid systems can be programmed to self-assemble into complex structures implementing dynamic mechanical tasks [33, 32, 38]. The field of nucleic acid nanotechnology is devoted to exploring and developing these capabilities for applications in molecular robotics, fabrication, computation, biosensing, electronics, and medicine. To support both biological inquiry and technological innovation, we present a model and algorithms for analyzing the thermodynamics of interacting nucleic acid strands.

1.1. Background. We begin by reviewing some basic terminology and prior work to facilitate precise descriptions of the problem formulations and solution strategies. The *primary structure* of a nucleic acid strand is specified as a sequence of *bases* (of types A, C, G, or T for DNA, with U replacing T for RNA). Nucleic acid energetics are dominated by the formation of *base pairs* between complementary bases (Watson–Crick pairs C-G and A-T (DNA) or A-U (RNA) and less frequent pairs G-T (DNA) or G-U (RNA)), each base participating in at most one base pair. By convention, $i \cdot j$ denotes that base i is paired to base j. Strands are directional (with the beginning denoted $5'$ and the end denoted $3'$) and base pairing occurs in an antiparallel fashion (e.g., $5'$-GCTCA-$3'$ is the reverse-complement of $5'$-TGAGC-$3'$, allowing complete base pairing to yield a familiar DNA double helix). Here, we are interested in the general scenario of base pairing within a single DNA or RNA strand, or between multiple strands that have arbitrary degrees of complementarity.

The *secondary structure* of a nucleic acid strand in a particular physical conformation is simply the set of base pairs present in the molecule. In general, each sequence is compatible with multiple secondary structures. Figure 1.1a depicts a secondary structure in which some bases are paired and others are unpaired, and illustrates the decomposition of this secondary structure into different *loop* types. Each secondary structure is compatible with an ensemble of *tertiary structures* corresponding to the three-dimensional atomic coordinates of the strand. Remarkably, empirical potential functions based on secondary structure alone [41, 31, 23] have great utility for studying the properties of natural and engineered RNA and DNA structures [23, 40, 36, 35, 12]. For a given sequence, the free energy of secondary structure s is estimated as the sum of the empirically determined free energies[1] of the constituent loops [41, 31, 23],

$$\Delta G(s) = \sum_{\text{loop} \in s} \Delta G(\text{loop}), \tag{1.1}$$

each defined with respect to the free energy of the unpaired reference state.

Secondary structure models have enabled the development of efficient dynamic programming algorithms for characterizing the equilibrium properties of a DNA or RNA molecule. For algorithmic purposes, it is convenient to represent a secondary structure as a *polymer graph*, with the strand drawn along the circumference of a circle and base pairs depicted as straight lines joining complementary bases (Figure 1.1b). The class of secondary structures that are considered in dynamic programs is usually defined to exclude *pseudoknots* (Figure 1.1c), which correspond to polymer graphs with crossing lines (Figure 1.1d). The first dynamic programming algorithms for predicting the minimum free energy (MFE) secondary structure were proposed by Waterman and Smith [44] and Nussinov et al. [26]. In a seminal 1981 paper, Zuker and Stiegler [50] described a dynamic program for MFE determination for a nucleic acid strand over the ensemble of unpseudoknotted secondary structures Ω. In 1990, McCaskill [24] described a different dynamic program for calculating the partition

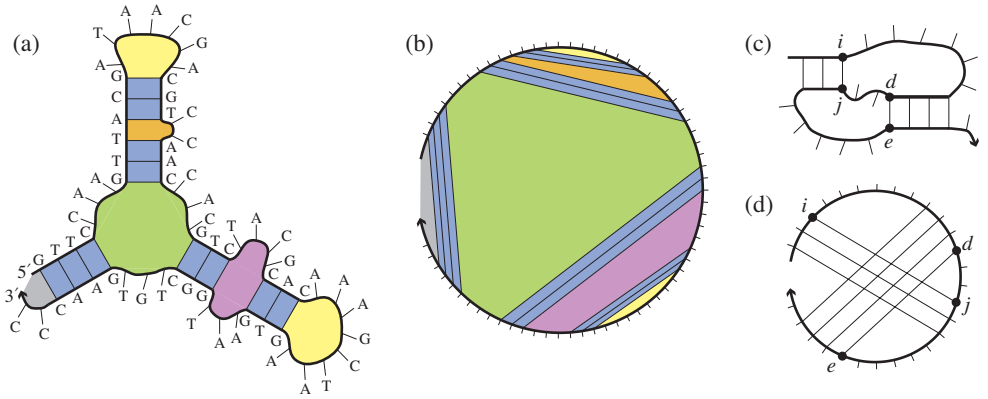


Fig. 1.1 *Secondary structure model for a single nucleic acid strand.* (a) *A sample secondary structure with the strand depicted as a directed thick line (an arrow marks the 3′ end), base pairs depicted as thin lines joining complementary bases, and unpaired bases depicted as thin protruding lines. This structure can be decomposed into canonical loop types* [41, 50]: *hairpin loops (a stretch of unpaired bases closed by one base pair; yellow), stacked base pairs (two consecutive base pairs with no unpaired bases between them; blue), an interior loop (two base pairs separated by unpaired bases on both sides of the loop; purple), a bulge loop (two base pairs separated by unpaired bases on only one side of the loop; orange), a multiloop (three or more base pairs; green), and an exterior loop (the loop containing the two ends of the strand; gray).* (b) *An equivalent polymer graph representation, with the strand depicted as a directed thick circular arc, bases depicted as protruding tick marks, base pairs depicted as straight lines joining complementary bases, and loops colored as in* (a). (c) *A sample pseudoknot with base pairs $i{\cdot}j$ and $d{\cdot}e$ (with $i < d$) that fail to satisfy the nesting property $i < d < e < j$, yielding crossing lines in the corresponding polymer graph* (d).

function over Ω. The *partition function* [18],

$$Q = \sum_{s \in \Omega} e^{-\Delta G(s)/kT},$$

where k is the Boltzmann constant and T is temperature, can be used to calculate the equilibrium probability of any secondary structure $s \in \Omega$,

$$p(s) = \frac{1}{Q}\, e^{-\Delta G(s)/kT}, \tag{1.2}$$

and therefore has profound implications for the development of rigorous sequence design methods [9]. Adaptations of the partition function algorithm allow the calculation of other important equilibrium properties including the probability of any base pair [24], thermodynamically representative samplings of secondary structures in the ensemble Ω [8], and the average number of incorrectly paired bases relative to a design target [9]. These tools are useful in practice for the analysis and design of functional nucleic acid systems [8, 22, 12, 29, 28].

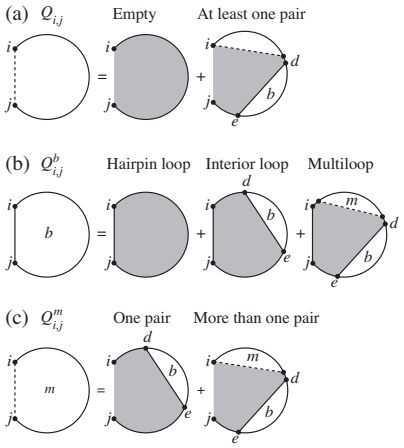
The exclusion of pseudoknots from Ω is founded on both modeling and algorithmic considerations. Energy models for pseudoknots are difficult to formulate due to the increased significance of geometric issues and tertiary interactions. Furthermore, if the ensemble is augmented to include all possible pseudoknots, MFE determination can be *NP*-hard [20, 1]. Consideration of restricted classes of pseudoknots enables the

specification of polynomial-time MFE determination [30, 1] and partition function [10, 11] algorithms. Although pseudoknots exist in nature [43] and have been incorporated in synthetic DNA systems [46, 42], many natural and synthetic structures of interest do not include pseudoknots [23, 32, 38], and we will not consider them here.

By employing a dynamic program, the unpseudoknotted MFE determination and partition function algorithms require $O(N^4)$ time and $O(N^2)$ storage to consider an ensemble of secondary structures Ω that grows exponentially with strand length N [49]. To provide a basis for describing the multistranded partition function algorithm of the present work, it is helpful to revisit McCaskill's dynamic program for the single-stranded case [24].[2] The dynamic program calculates the subsequence partition function $Q_{i,j}$ for each subsequence $[i, j]$, starting from short subsequences and iteratively considering longer subsequences until the full partition function $Q_{1,N}$ is obtained. The calculation of $Q_{i,j}$ relies on additional restricted partition functions Q^b and Q^m as detailed by the recursion diagrams and equations of Figure 1.2. The key conceptual challenge in evaluating the partition function is the avoidance of algorithmic redundancy. Zuker and Stiegler's MFE determination recursions [50] are *redundant* in the sense that a single secondary structure can be encountered by multiple recursion trajectories through the dynamic program. Although this may affect efficiency, it does not affect accuracy, since the outcome of selecting the MFE structure from an ensemble of competing structures is unaffected by the number of times a given structure is considered. However, the partition function algorithm evaluates a weighted sum over all secondary structures in Ω, so repetition implies overcounting the contributions of some structures. McCaskill's partition function algorithm [24] relies on strictly nonredundant recursions that incorporate the contribution of each secondary structure with exactly one trajectory in the recursive process.

While MFE determination and partition function algorithms for a single unpseudoknotted strand have existed for many years, progress on the multistranded case is quite recent. In 2004, Dimitrov and Zuker [7] described a partition function algorithm for unpseudoknotted interactions between two strands with the restriction that intrastrand base pairs are prohibited. Given initial concentrations (e.g., x_A^0 and x_B^0) of two strand species (A and B) and calculated partition functions for each of the possible monomer and dimer complexes ($Q_A, Q_B, Q_{AA}, Q_{BB}, Q_{AB}$), they described an approach for calculating the equilibrium concentration of each species ($x_A, x_B, x_{AA}, x_{BB}, x_{AB}$) in a dilute solution. The approach is suitable for the study of short nearly complementary strands of the type used for free energy parameterization studies [7] but is not applicable to the diverse multistranded structures that are the hallmark of DNA nanotechnology applications [38, 32, 33] or to RNA regulatory processes in biology involving intrastrand and interstrand base pairing [15, 3]. Alternatively, Andronescu, Zhang, and Condon [2] developed an MFE determination algorithm that considers all unpseudoknotted secondary structures for complexes with an arbitrary number of strands. The issue of algorithmic redundancy precludes the extension of this approach to partition function calculations. Here, we address the thermodynamic analysis of an arbitrary number of interacting nucleic acid strands, providing a rigorous treatment of the physical and algorithmic subtleties that arise, including new challenges associated with secondary structure modeling, molecular symmetry, algorithmic redundancy, and convexity.

1.2. Outline. We start by defining a multistranded secondary structure model, proving a representation result that is the first step to ensuring that the partition function algorithm is free of redundancies. For complexes of interacting strands in



$$Q_{i,j} = 1 + \sum_{i \le d < e \le j} Q_{i,d-1}\, Q^b_{d,e}$$

$$\begin{aligned} Q^b_{i,j} = {} & \exp\{-\Delta G^{\text{hairpin}}_{i,j}/kT\} \\ & + \sum_{i<d<e<j} \Big[Q^b_{d,e} \exp\{-\Delta G^{\text{interior}}_{i,d,e,j}/kT\} \\ & + Q^m_{i+1,d-1} Q^b_{d,e} \exp\big\{-[\Delta G^{\text{multi}}_{\text{init}} + 2\Delta G^{\text{multi}}_{\text{bp}} + \Delta G^{\text{multi}}_{e+1,j-1}]/kT\big\} \Big] \end{aligned}$$

$$\begin{aligned} Q^m_{i,j} = \sum_{i \le d < e \le j} \Big[& Q^b_{d,e} \exp\big\{-[\Delta G^{\text{multi}}_{\text{bp}} + \Delta G^{\text{multi}}_{i,d-1} + \Delta G^{\text{multi}}_{e+1,j}]/kT\big\} \\ & + Q^m_{i,d-1}\, Q^b_{d,e} \exp\big\{-[\Delta G^{\text{multi}}_{\text{bp}} + \Delta G^{\text{multi}}_{e+1,j}]/kT\big\} \Big] \end{aligned}$$

Fig. 1.2 *Single-stranded partition function algorithm described by recursion diagrams (left) and recursion equations (right). By convention, a solid straight line indicates a base pair and a dashed line demarcates a region without implying that the connected bases are paired. Shaded regions correspond to loop free energies*[3] *that are explicitly incorporated at the current level of recursion.* (a) $Q_{i,j}$ *represents the partition function for subsequence* $[i,j]$. *There are two cases: either there are no base pairs (corresponding to the reference state* $\Delta G^{\text{empty}} \equiv 0$ *and a partition function contribution of unity) or there is a* $3'$*-most base pair* $d \cdot e$. *In the latter case, determination of the partition function contribution makes use of previously computed subsequence partition functions* $Q^b_{d,e}$ *and* $Q_{i,d-1}$. *By the distributive law, multiplication of these subsequence partition functions (each representing a sum over substructures) implicitly sums over all pairwise combinations of substructures. The independence of the loop contributions in the energy model* (1.1) *implies that these products appropriately add the free energies in the exponents.* (b) $Q^b_{i,j}$ *is the partition function for subsequence* $[i,j]$ *with the restriction that bases* i *and* j *are paired. There are three cases: either there are no additional base pairs (corresponding to a hairpin loop), there is exactly one additional base pair* $d \cdot e$ *(corresponding to an interior loop), or there is more than one additional base pair (corresponding to a multiloop) with* $3'$*-most pair* $d \cdot e$ *and at least one additional pair specified in a previously computed subsequence partition function* $Q^m_{i+1,d-1}$. (c) $Q^m_{i,j}$ *is the partition function for subsequence* $[i,j]$ *with the restrictions that the subsequence is inside a multiloop and contains at least one base pair. There are two cases: either there is exactly one additional base pair* $d \cdot e$ *defining the multiloop or there is more than one additional base pair defining the multiloop (with* $3'$*-most pair* $d \cdot e$*). Initialization requires* $Q_{i,i-1} = 1$ *and* $Q^m_{i,i-1} = 0$ *for* $i = 1, \ldots, N$.

which some strands are identical (e.g., AA or $ABAB$), issues of molecular and algorithmic distinguishability necessitate symmetry corrections to the physical model and overcounting corrections to the partition function recursions. Although it is not clear how to address these two effects in isolation, we prove that they can be simultaneously and exactly corrected. We then describe dynamic programming recursions for evaluating the partition function of a complex containing an arbitrary number of strands. We further address modeling and algorithmic issues surrounding the conversion of the partition function algorithm into an MFE determination algorithm over the same ensemble of multistranded secondary structures. The partition function of a fixed volume (or "box") containing a dilute solution of complexes can be expressed in terms of the number of solvent molecules and the partition function and population of each type of complex in the box. For small numbers of complexes, direct evaluation of the partition function of the box is feasible, enabling calculation of the equilibrium probability distribution for the population of each complex species. Alternatively,

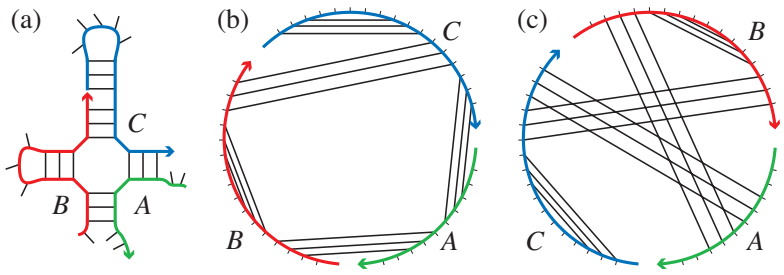


Fig. 2.1 *Multistranded secondary structure model.* (a) *A connected unpseudoknotted secondary structure for a complex of three distinct strands with sequences A, B, and C. The set of distinct circular permutations is* $\overline{\Pi} = \{123, 132\}$. (b) *Polymer graph representation of the secondary structure with no crossing lines corresponding to* $\pi = 123$. *Loop classifications are the same as for the single-stranded case (Figure* 1.1a*).* (c) *Alternative polymer graph with crossing lines corresponding to* $\pi = 132$.

in the thermodynamic limit of large populations that is most relevant to typical experimental conditions in a test tube, it is still possible to calculate the equilibrium concentration of each complex species. Concentration determination leads to a convex programming problem that can be solved efficiently in the dual form to yield the unique concentrations corresponding to thermodynamic equilibrium. Partition function information can then be used to calculate the expected number of each type of base pair in the solution. Throughout the paper, we refer to the notes of Appendix B for technical details of interest to specialists.

2. Partition Function of a Complex of Interacting Strands. We now consider the modeling and algorithmic issues surrounding the calculation of the partition function and MFE secondary structure for a complex with an arbitrary number of interacting strands.

2.1. Secondary Structure Model. For L interacting strands, we assign to each a unique identifier in $\{1, \ldots, L\}$. As for the single-stranded case, the secondary structure of multiple interacting strands is defined by a list of base pairs, where here each base is specified by a strand identifier and a position on that strand. For example, $i_n \cdot j_m$ denotes base i of strand n pairing with base j of strand m.[4]

A polymer graph for a secondary structure can be constructed by ordering the strands and drawing them in succession from $5'$ to $3'$ around the circumference of a circle with a *nick* between each strand and straight lines connecting paired bases (Figure 2.1). The distinct ways to order the L strands on a circle correspond to the set $\overline{\Pi}$ of *circular permutations* containing $(L-1)!$ *orderings* (e.g., $\overline{\Pi} = \{123, 132\}$ for a complex of three strands). Each circular permutation defines a distinct polymer graph. *Cyclic permutations* change the location of the strands on the circle without affecting the relative orderings and hence contribute no additional orderings to $\overline{\Pi}$ (e.g., the orderings 123, 231, and 312 are indistinguishable on a circle, as are the orderings 132, 321, and 213).

For a given secondary structure, if every circular permutation $\pi \in \overline{\Pi}$ corresponds to a polymer graph with crossing lines, then the secondary structure is *pseudoknotted*. A polymer graph with no crossing lines can be decomposed into *loops* as for the single-stranded case, and all loops containing one nick are *exterior loops*. A secondary structure is *connected* if no loop contains more than one nick (i.e., no subset of the

strands is free of the others), in which case the L strands constitute a *complex*. Let $\overline{\Omega}$ be the set of all unpseudoknotted secondary structures for a given complex of L strands, and let $\overline{\Omega}(\pi)$ be the subset of $\overline{\Omega}$ that can be represented as polymer graphs with no crossing lines using strand ordering π. The following representation theorem ensures that these subsets comprise a partitioning of $\overline{\Omega}$, i.e., each secondary structure is in exactly one subset.

THEOREM 2.1 (Representation). *For every unpseudoknotted connected secondary structure $s \in \overline{\Omega}$, there is exactly one circular permutation $\pi \in \overline{\Pi}$ that yields a polymer graph with no crossing lines.*

Proof. The proof is by induction on the number of strands in the complex, L. The theorem holds for $L = 1$ since there is only one circular permutation. We now assume that the theorem holds for $L-1$ strands and attempt to show that it holds for L strands.

Let s be an unpseudoknotted connected secondary structure of L strands so that there must exist a circular permutation π for which the polymer graph has no crossing lines. We create a connectivity graph for s in which each strand is represented by exactly one node and there is an edge between nodes if there exists at least one base pair between the corresponding strands. Since s is connected, the connectivity graph must have either a leaf node or a node that is part of a cycle whose removal will not break the connectedness of the resulting graph. Let l be some such node and let s' be the secondary structure which has had the corresponding strand removed. Then s' is a connected secondary structure of $L-1$ strands. By supposition, the circular permutation π' of the strands in s' that corresponds to π (omitting strand l) is the only one that yields a polymer graph with no crossing lines. Hence, the only polymer graphs for structure s that have the possibility of no crossing lines are those that are obtained by inserting strand l between two strands of s' in circular permutation π'.

Now we show that the only position where strand l can be added back into the polymer graph with circular permutation π' without introducing crossing lines is the original position n corresponding to circular permutation π. Consider inserting l into π' at positions $m \neq n$. A line drawn from m to n must cross some base pair $i \cdot j$ in the polymer graph or s' would not be connected. In the original strand ordering π for structure s, there must exist a base pair $d \cdot e$ connecting strand l to another strand in the complex. This base pair $d \cdot e$ cannot cross $i \cdot j$, so both d and e are on the n side of $i \cdot j$. If we now insert strand l at a position m, crossing lines are produced because one end of $d \cdot e$ is on the m side of $i \cdot j$ and the other is on the n side. This implies that n is the only position where l can be added back to s' without introducing crossing lines. Hence, the original polymer graph corresponding to circular permutation π is the only one without crossing lines. □

Remark 1. A simple counterexample illustrates that this result does not hold if $\overline{\Omega}$ is permitted to contain secondary structures that are not connected. Consider three strands with circular permutations $\overline{\Pi} = \{123, 132\}$. Any unpseudoknotted secondary structure in which strands 2 and 3 form a complex that is disconnected from strand 1 will yield a polymer graph with no crossing lines for both orderings $\pi \in \overline{\Pi}$.

Remark 2. Secondary structure kinetics are often modeled as a sequence of elementary moves, each corresponding to the formation, breakage, or shifting of a single base pair [13]. The representation theorem has an interesting physical implication: a complex of L strands cannot transition via elementary moves between two secondary structures corresponding to different circular permutations $\pi \in \overline{\Pi}$ without temporarily leaving the ensemble $\overline{\Omega}$ of unpseudoknotted connected secondary structures. Either the strands must dissociate and reconnect or they must transition through a pseudo-

knotted state. As a result, for some systems it may be desirable to consider *ordered complexes*, each defined by a single subensemble $\overline{\Omega}(\pi)$.

For each secondary structure $s \in \overline{\Omega}$, the free energy, $\overline{\Delta G}(s)$, is the sum of the free energies of the constituent loops plus a strand association penalty ΔG^{assoc} [4] applied $L-1$ times for a complex of L strands:

$$\overline{\Delta G}(s) = (L-1)\,\Delta G^{\text{assoc}} + \sum_{\text{loop}\in s} \Delta G(\text{loop}).$$

All cyclic strand permutations of a polymer graph with no crossing lines are equivalent, having identical loop decompositions and identical free energies.

To calculate the partition function over the ensemble $\overline{\Omega}$, our strategy is to consider each subensemble $\overline{\Omega}(\pi)$ separately, calculating $\overline{Q}(\pi) \equiv \sum_{s\in\overline{\Omega}(\pi)} e^{-\overline{\Delta G}(s)/kT}$ by applying a generalization of the single-stranded partition function algorithm (section 2.3) to each strand ordering $\pi \in \overline{\Pi}$. The representation theorem then guarantees that the partition function

$$\overline{Q} = \sum_{s\in\overline{\Omega}} e^{-\overline{\Delta G}(s)/kT} = \sum_{\pi\in\overline{\Pi}} \overline{Q}(\pi)$$

considers every secondary structure $s \in \overline{\Omega}$ exactly once.

2.2. Distinguishability Issues. In most experimental settings, strands with the same sequences behave identically with respect to relevant measurable quantities. In this context, we consider strands with identical sequences to be *indistinguishable*, while strands that differ in sequence are *distinct*. Similarly, two secondary structures are indistinguishable if there exists a permutation of the strand identifiers that maps indistinguishable strands onto each other while preserving all base pairs; otherwise, the structures are distinct. To describe experimental observables, we must correct the calculated partition function $\overline{Q}$ to obtain the partition function Q that accounts for the indistinguishability of secondary structures in $\overline{\Omega}$. This necessitates corrections for physical symmetries and algorithmic overcounting. These two challenges can be simultaneously (but not separately) addressed in a mathematically rigorous way.

We consider a maximal subset of distinct secondary structures $\Omega \subseteq \overline{\Omega}$. Thus, we define $\Omega(\pi) \subseteq \overline{\Omega}(\pi)$ to be a maximal subset of distinct secondary structures for strand ordering π. Then Ω is the union of $\Omega(\pi)\ \forall\ \pi \in \Pi$, where $\Pi \subseteq \overline{\Pi}$ is a maximal subset of distinct orderings with respect to sequences. Consequently, Π may contain fewer than $(L-1)!$ members (e.g., $\Pi = \{AAB\}$ for a complex of strands with sequences labeled A, A, and B). It follows from Theorem 2.1 that for every $s \in \Omega$, there is exactly one $\pi \in \Pi$ yielding a polymer graph with no crossing lines.

The first challenge pertains to the free energy model for L interacting strands. The free energy of a rotationally asymmetric secondary structure is calculated by adding the free energies of the constituent loops and strand association penalties. However, a secondary structure with an R-fold rotational symmetry corresponds to an R-fold reduction in the distinguishable conformational space of the molecule. A free energy model based on summing local contributions cannot account for the entropy reduction implied by this global R-fold symmetry, so the free energy must be adjusted by a symmetry correction[5] of $kT\log R$:

$$\Delta G(s) = kT\log R + \overline{\Delta G}(s) = kT\log R + (L-1)\,\Delta G^{\text{assoc}} + \sum_{\text{loop}\in s} \Delta G(\text{loop}). \tag{2.1}$$

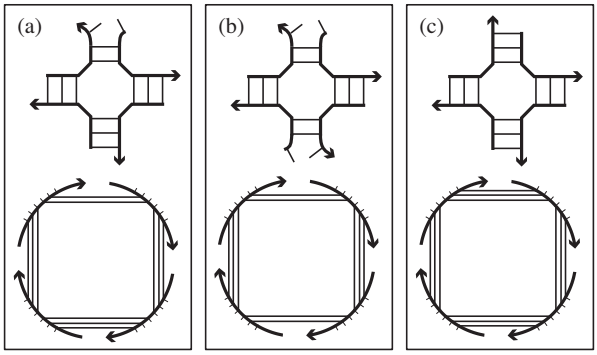


Fig. 2.2 *Sample secondary structures and polymer graphs for a complex of four indistinguishable strands.* (a) 1*-fold (i.e., no) rotational symmetry.* (b) 2*-fold rotational symmetry.* (c) 4*-fold rotational symmetry.*

While the $5'$ to $3'$ directionality of nucleic acid strands prevents a single-stranded secondary structure from being rotationally symmetric, a complex of multiple strands may have a rotational symmetry whenever a cyclic permutation maps a strand ordering onto itself. For example, a complex of four indistinguishable A strands with $\Pi = \{AAAA\}$ (Figure 2.2) can have secondary structures with either 1-fold (i.e., no), 2-fold, or 4-fold rotational symmetries, depending on whether the secondary structures are identical within each $AAAA$, AA, or A subunit, respectively. Algorithmically, the difficulty is that the free energies for different secondary structures in the same distinct circular permutation π require different symmetry corrections. Since the contributions of each structure are calculated in a recursive fashion rather than by explicit enumeration, it is not obvious how to introduce the appropriate corrections in a dynamic program.

The second challenge arises because the dynamic programming algorithm computes the partition function over the set of secondary structures in $\overline{\Omega}(\pi)$ rather than $\Omega(\pi)$. If the strand ordering can be mapped onto itself by a cyclic permutation, some indistinguishable secondary structures may be treated as distinct by the algorithm, resulting in an overcounting of the corresponding partition function contributions. For example, the 1-fold symmetric secondary structure of Figure 2.2a will be encountered four times during the recursive process, as each of four stems can play the role of introducing the asymmetry. Hence, the partition function contribution of this secondary structure will be overcounted by a factor of four. By comparison, the 2-fold symmetric structure of Figure 2.2b will be encountered twice and its partition function contribution overcounted by a factor of two. Meanwhile, the 4-fold symmetric structure of Figure 2.2c can only be represented by a single polymer graph so its contribution will be counted only once. In general, the partition function contribution of each secondary structure must be divided by the number of indistinguishable representations of that secondary structure among the polymer graphs for a given strand ordering. Again, it is not obvious how to correct the contributions of individual secondary structures in the absence of explicit enumeration. Fortunately, the symmetry and overcounting corrections are intimately linked, and they can be simultaneously treated as described by the following theorem.

Consider an ordering $\pi \in \Pi$ of L strands, where some of the strands may be indistinguishable. Let $\mathcal{G}$ be the group of $v(\pi)$ cyclic permutations mapping each

strand to a strand of the same species. For example, $v(\pi) = 4$ for $\pi = AAAA$, $v(\pi) = 3$ for $\pi = ABABAB$, and $v(\pi) = 2$ for $\pi = ABAABA$, where the elements of $\mathcal{G}$ correspond to all rotations of a polymer graph that map strands of type $A \to A$ and strands of type $B \to B$.

THEOREM 2.2 (Distinguishability Correction). *For an ordering $\pi \in \Pi$ of L strands, if the multistranded partition function algorithm yields $\overline{Q}(\pi)$, then the corrected partition function, $Q(\pi)$, that properly accounts for both symmetry and overcounting corrections is $Q(\pi) = \overline{Q}(\pi)/v(\pi)$.*

Proof. Consider an arbitrary secondary structure $s \in \overline{\Omega}(\pi)$. A permutation $g \in \mathcal{G}$ acts on a secondary structure s by relabeling strand identifiers: $g(s) = \{i_{g(n)} \cdot j_{g(m)} : i_n \cdot j_m \in s\}$. The stabilizer of s, $\mathcal{G}_s = \{g \in \mathcal{G} : g(s) = s\}$, is the set of cyclic permutations of strand identifiers (rotations of the polymer graph) that map s onto itself. The order of the rotational symmetry of the physical complex with secondary structure s is given by $|\mathcal{G}_s|$, requiring a correction of $+kT \log |\mathcal{G}_s|$ to the standard loop-based free energy.

The orbit of s in $\mathcal{G}$, $\mathcal{G}(s) = \{g(s) \in \overline{\Omega}(\pi) : g \in \mathcal{G}\}$, is the subset of $\overline{\Omega}(\pi)$ corresponding to the images of s under the permutations of the group $\mathcal{G}$. Note that the members of $\mathcal{G}(s)$ represent indistinguishable secondary structures within $\overline{\Omega}(\pi)$. Consequently, the partition function contribution of secondary structure $s \in \Omega(\pi)$ will be overcounted by a factor of $|\mathcal{G}(s)|$ because the recursion algorithm treats elements of the orbit as algorithmically distinct even though they are physically indistinguishable.

The orbit-stabilizer theorem of group theory [14] provides the useful relationship

$$|\mathcal{G}_s||\mathcal{G}(s)| = |\mathcal{G}| = v(\pi) \; \forall \, s \in \overline{\Omega}(\pi),$$

linking the symmetry and overcounting effects. We will make use of the fact that the product $|\mathcal{G}_s||\mathcal{G}(s)|$ depends only on the strand ordering π and is independent of the specific secondary structure $s \in \overline{\Omega}(\pi)$.

The partition function algorithm applied to strand ordering π yields $\overline{Q}(\pi) = \sum_{s \in \overline{\Omega}(\pi)} \exp\{-\overline{\Delta G}(s)/kT\}$. The corrected partition function then takes the form

$$\begin{aligned} Q(\pi) &= \sum_{s \, \in \, \Omega(\pi)} \exp\{-\Delta G(s)/kT\} \\ &= \sum_{s \, \in \, \overline{\Omega}(\pi)} \frac{1}{|\mathcal{G}(s)|} \exp\{-(\overline{\Delta G}(s) + kT \, \log |\mathcal{G}_s|)/kT\} = \frac{\overline{Q}(\pi)}{v(\pi)}. \end{aligned} \tag{2.2}$$

Thus, the symmetry and overcounting corrections conveniently combine to give a uniform correction factor of $v(\pi)^{-1}$ to the partition function $\overline{Q}(\pi)$ for a given ordering $\pi \in \Pi$. The corrected partition function $Q(\pi)$ can then be associated with the subensemble $\Omega(\pi)$. ∎

The partition function for a complex of L strands can then be obtained by summing the corrected partition functions $Q(\pi) = \overline{Q}(\pi)/v(\pi)$ for each distinct circular permutation

$$Q = \sum_{\pi \in \Pi} Q(\pi). \tag{2.3}$$

The representation theorem ensures that the contributions for different strand orderings $\pi \in \Pi$ are nonredundant, and the distinguishability correction theorem ensures that the contributions within each ordering $\pi \in \Pi$ are nonredundant and use the correct physical model.

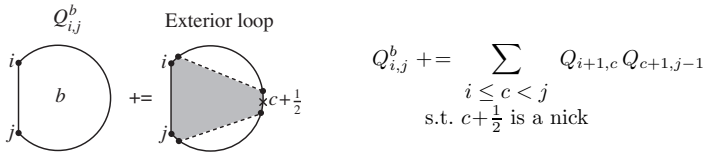


Fig. 2.3 *Updates to the single-stranded recursion diagrams (left) and recursion equations (right) of Figure* 1.2 *to enable calculation of the partition function* $\overline{Q}(\pi)$ *for circular permutation* π *for a complex of* L *strands. The recursions for* $Q_{i,j}$ *and* $Q^m_{i,j}$ *remain unchanged and the recursion for* $Q^b_{i,j}$ *is augmented with a case treating exterior loop structures based on the possible locations for a single top-level nick (i.e., a nick that has not been accounted for within a previously calculated* $Q^b_{d,e}$ *recursion). Retaining the single-stranded diagram conventions, we additionally denote nicks between strands by half indices (e.g.,* $c+\frac{1}{2}$*) and adopt the convention that there is a top-level nick between neighboring indices if and only if there is an* $\times$ *between them on the circle. These conventions must be applied when reinterpreting the single-stranded recursions for use in multistranded calculations. Implementation details are provided in the pseudocode of Appendix* A.

The equilibrium probability $p(s)$ of a secondary structure $s \in \Omega$ for a complex of L strands is then calculated using (1.2) with the symmetry-corrected free energy (2.1). To illustrate the influence of distinguishability effects, consider calculating $p(s)$ for the complex $AAAA$ with $v(\pi) = 4$ and $\Pi = \{AAAA\}$. If we omit the distinguishability correction (2.2), the partition function is calculated to be four times the correct value. If we also ignore the symmetry correction (2.1) to the standard free energy model, the Boltzmann factor $e^{-\Delta G(s)/kT}$ for a secondary structure with a 4-fold rotational symmetry will also be four times the correct value so (1.2) produces the correct probability $p(s)$. However, for a secondary structure with a 2-fold rotational symmetry, $p(s)$ will be calculated to have half the correct value, and for a structure with no rotational symmetry, $p(s)$ will be calculated to have one-quarter the correct value. On the other hand, when calculating the free energy of the complex[6] as $\Delta G = -kT \log Q$, the logarithm reduces the influence of errors in the partition function, yielding a free energy that is too low by $kT \log 4$ for this example. The effect of neglecting the distinguishability correction becomes more complicated for larger complexes in which Q contains contributions $Q(\pi)$ from multiple orderings $\pi \in \Pi$, each with a different correction $v(\pi)$. Fortunately, the distinguishability correction theorem ensures that these effects can be treated in a straightforward manner for a complex with an arbitrary number of strands.

2.3. Partition Function Recursions. It now remains to define recursions for evaluating the partition function $\overline{Q}(\pi)$ over subensemble $\overline{\Omega}(\pi)$ corresponding to a particular circular permutation π of L strands. If the ordering requires symmetry and overcounting corrections (i.e., $v(\pi) > 1$), these distinguishability issues are treated a posteriori as described in the previous section. In the context of the partition function recursions, each strand in an ordering π is distinguishable based on its position in the ordering (regardless of possible indistinguishable strands at other positions in the ordering). If strand l has length N_l, then the recursions operate on a single concatenated strand of length $N \equiv \sum_{l=1}^{L} N_l$. The location of the nicks between strands are known based on the specifications of the strand lengths in the particular ordering.

The spirit of the multistranded algorithm is identical to that of the single-stranded algorithm of Figure 1.2. At the level of recursion diagrams and equations, Figure 2.3 illustrates that the only change is the addition of an extra case to incorporate exterior

loop contributions to the Q^b recursion. The details of the implementation are complicated by the requirement that every polymer graph encompassed by the recursions must be connected. This amounts to ensuring that no loop contains more than one nick, which is handled implicitly by the recursion diagram conventions of Figure 2.3 and explicitly by the conditionals in the pseudocode of Appendix A (see Figure A.1). Like the single-stranded algorithm, the recursions require $\mathcal{O}(N^4)$ time and $\mathcal{O}(N^2)$ space in their most transparent form, but the time complexity can be reduced to $\mathcal{O}(N^3)$ using standard methods [21, 10].

2.4. MFE Determination Recursions. MFE determination is often used for structure prediction since the MFE secondary structure has the maximum equilibrium probability in the ensemble Ω. However, it is possible for a subensemble of competing structures to dominate the MFE secondary structure, so care should be used in applying this approach [9]. In the single-stranded case, partition function recursions can be converted into MFE determination recursions in a straightforward way.[7] To perform the conversion, every product of exponentiated free energies is replaced by a sum of free energies, and every sum over alternative partition function contributions is replaced by a minimization over alternative free energy contributions. At the end of the process, the output is the value of the free energy $\Delta G(s)$ for the MFE structure in Ω. A backtracking algorithm employing the intermediate subsequence results can then be used to identify the corresponding secondary structure s [50].

If the ensemble Ω contains no rotationally symmetric secondary structures, the same conversion approach can be used to obtain MFE determination recursions in the multistranded case. However, if secondary structure $s \in \Omega$ has an R-fold rotational symmetry, the free energy $\overline{\Delta G}(s)$ employed by the MFE determination recursions will be missing the appropriate symmetry correction $kT \log R$. Unlike in the partition function setting, there is no difficulty with algorithmic overcounting for MFE determination since the repeated taking of minimums over free energies of indistinguishable polymer graphs does not change the outcome of the comparisons. Ironically, the absence of the overcounting problem makes it harder to correct the remaining symmetry problem; it is not clear that these corrections can be directly incorporated into the dynamic programming recursions since the individual structures are never explicitly enumerated.

One approach is to calculate the MFE secondary structure ignoring the symmetry corrections to the free energy model. If the predicted MFE structure is asymmetrical, then it is the true minimum using the corrected free energies because the symmetry correction is always positive. However, if the predicted MFE structure contains an R-fold symmetry, then its free energy must be increased by $kT \log R$. The true minimum can then be identified by exhaustively enumerating [47] all secondary structures with free energies within $kT \log R$ of the original predicted minimum, applying any needed symmetry correction to the free energy for each structure. The disadvantage of this approach is that it scales exponentially with N. The difficulty in modeling rotational symmetries for multistranded MFE calculations highlights the significance of the distinguishability correction theorem in facilitating partition function calculations for complexes of interacting strands.

3. Partition Function Analysis of Dilute Solutions of Interacting Strands. Using the partition function algorithm for a single complex as a starting point, we now turn to the problem of analyzing the equilibrium properties of a fixed volume (or "box") containing an arbitrary number of strand species that interact to form complexes in a dilute solution.

3.1. Contents of the Box. Consider a box containing M_s solvent molecules and the set of strand species Ψ^0 with total strand populations $m^0 \in \mathbb{Z}_{>0}^{|\Psi^0|}$. Suppose the strands can interact to form the set of complexes Ψ (so $\Psi^0 \subseteq \Psi$) with populations $m \in \mathbb{Z}_{\geq 0}^{|\Psi|}$. The total number of strands is $M^0 \equiv \sum_{i \in \Psi^0} m_i^0$ and the total number of complexes is $M \equiv \sum_{j \in \Psi} m_j$.

We seek to determine the partition function of the box corresponding to the Boltzmann weighted sum over all possible states of the system. It is first necessary to calculate the partition function Q_j for each complex $j \in \Psi$. In principle, a complex can contain arbitrarily many strands, so in practice we limit the size of Ψ. This can be achieved, for example, by specifying Ψ to contain only those complex species that are expected to be physically significant. Here, we consider all possible complexes with L strands for $1 \leq L \leq L_{\max}$.[8] For the set of strand species Ψ^0 with sequences such that all complexes in this size range can form, the total number of complex species is then given by[9]

$$|\Psi| = \binom{L_{\max} + |\Psi^0|}{|\Psi^0|} - 1. \tag{3.1}$$

Hence, we typically have $|\Psi^0| \ll |\Psi|$. By (2.3), calculation of the partition function Q_j requires $|\Pi_j|$ applications of the multistranded partition function algorithm of Appendix A, where Π_j is the set of distinct circular permutations for complex j. By the Pólya enumeration theorem, calculation of Q_j for all complexes $j \in \Psi$ then requires a total of

$$\sum_{L=1}^{L_{\max}} \sum_{l=1}^{L} \frac{|\Psi^0|^{\gcd(l,L)}}{L}$$

applications of the multistranded partition function algorithm.[10] Recalling the $\mathcal{O}(N^3)$ time complexity of the multistranded partition function algorithm (for a complex containing N bases), the time complexity for calculating the partition functions for all complexes in the box is $\mathcal{O}(|\Psi^0|^{L_{\max}} N_{\max}^3 / L_{\max})$, where $N_{\max}$ is the largest number of bases in a complex.

3.2. Partition Function of the Box. We now formulate the partition function of the box in terms of the number of solvent molecules M_s and the complex populations m_j and partition functions Q_j $\forall j \in \Psi$. The complexes are much bigger than the solvent molecules, but experimental studies demonstrate that their contribution to the free energy of the system is independent of their size [48]. We therefore introduce a small (but negligible) error by assuming that the solvent molecules and complexes are comparable in size. Dividing the box into small cells, we place one solvent molecule or complex in each. The solution is assumed to be *dilute* ($M_s \gg M^0$) so that interactions between complexes are negligible at equilibrium. We also assume that the free energy of each complex is independent of its position within the box. Thus, the free energy of the box is additive in the free energies of the complexes and the partition function for the box is multiplicative in the complex partition functions Q_j:

$$Q_{\text{box}} = Q_{\text{ref}} \sum_{m \in \Lambda} \left[\frac{(M_s + M)!}{M_s! \prod_{j \in \Psi} m_j!} \prod_{j \in \Psi} Q_j^{m_j} \right]. \tag{3.2}$$

Here, Q_{ref} may be chosen to set the reference state of the free energy and Λ is the set of population vectors m satisfying the conservation of mass constraint $A\,m = m^0$, where

$A \in \mathbb{Z}_{\geq 0}^{|\Psi^0|\times|\Psi|}$ has entries A_{ij} denoting the number of strands of species i in complex j. The quotient in (3.2) is the standard multinomial expression for the number of possible ways the complexes and solvent can be arranged in the cells (effectively an entropy of mixing), given that members of each species are indistinguishable.

Since $M_s \gg M^0 \geq M$, we have $(M_s+M)!/M_s! = M_s^M \left(1 + \mathcal{O}(M_s^{-1})\right)$ and hence[11]

$$Q_{\text{box}} \approx Q_{\text{ref}} \sum_{m\in\Lambda} q(m), \tag{3.3}$$

where

$$q(m) \equiv \prod_{j\in\Psi} \frac{M_s^{m_j}\, Q_j^{m_j}}{m_j!}$$

is interpreted as the partition function corresponding to a particular vector of populations m. The free energy of the box is given by $\Delta G_{\text{box}} = -kT \log Q_{\text{box}}$. It is convenient to define ΔG_{box} to be zero when all strands are contained in the box and there are no base pairs, so we specify $Q_{\text{ref}} \equiv \prod_{i\in\Psi^0}(m_i^0!/M_s^{m_i^0})$.

The probability of population vector m at equilibrium is

$$p(m) = Q_{\text{box}}^{-1}\, Q_{\text{ref}}\, q(m) \tag{3.4}$$

and the expected value of each population m_j is[12]

$$\langle m_j \rangle = \sum_{m\in\Lambda} m_j\, p(m). \tag{3.5}$$

The probability distribution for each m_j is then found by calculating

$$p_j(n) = \sum_{\substack{m\,\in\,\Lambda \\ \text{s.t. } m_j = n}} p(m)$$

for each value n taken by m_j in the set Λ. For a box containing a small number of strands, Q_{box} and the equilibrium population distributions can be evaluated explicitly. For a large box containing a large number of strands, explicit enumeration of all population vectors m in Λ is no longer feasible.

3.3. Concentration Determination in the Thermodynamic Limit. We now describe an approach for determining the equilibrium concentration for each species of complex in the thermodynamic limit of large populations. This problem corresponds to typical experimental conditions in a test tube and is relevant to designing and analyzing experiments for both technological and biological studies.

For large systems, the distributions of extensive thermodynamic variables (the value of an extensive variable is proportional to the size of the system it describes) are Gaussian with variance scaling as the mean [18]. Hence, for large numbers of interacting strands, the distribution of populations, $p(m)$, is a Ψ-dimensional Gaussian with variance proportional to $\langle m_j \rangle$ in coordinate j. By (3.4), the distribution of $q(m)$ is a rescaling of $p(m)$ and hence Gaussian, so the sum of (3.3) may be approximated by a product of the height $Q_{\text{ref}}\, q(\langle m \rangle)$ and the width $\langle m_j \rangle^{1/2}$ in each coordinate $j \in \Psi$:

$$Q_{\text{box}} \approx Q_{\text{ref}}\, q(\langle m \rangle) \prod_{j\in\Psi} \langle m_j \rangle^{\frac{1}{2}}\,.$$

Substituting for $q(\langle m\rangle)$ and applying Stirling's approximation ($\log n! = n\log n - n + \mathcal{O}(\log n)$) we obtain the free energy

$$\Delta G_{\text{box}} = -kT\log Q_{\text{ref}} + kT\sum_{j\in\Psi}\left\{\langle m_j\rangle\left[\log\left(\frac{\langle m_j\rangle}{M_s}\right) - \log Q_j - 1\right] + \mathcal{O}\left(\log\langle m_j\rangle\right)\right\}.$$

The contribution to ΔG_{box} by each complex $j\in\Psi$ scales as $\langle m_j\rangle\log\langle m_j\rangle$, while the error in this contribution resulting from neglecting the width of the distribution and from using Stirling's approximation is only $\mathcal{O}(\log\langle m_j\rangle)$. Hence, for large systems, ΔG_{box} can be accurately calculated by replacing (3.3) with $Q_{\text{box}} \approx Q_{\text{ref}}\, q(\langle m\rangle)$. On a per-solvent basis, the dimensionless free energy is then

$$g(\langle x\rangle) \equiv \frac{\Delta G_{\text{box}}}{M_s kT} \approx g_{\text{ref}} + \sum_{j\in\Psi}\left[\langle x_j\rangle\left(\log\langle x_j\rangle - \log Q_j - 1\right)\right],$$

where $\langle x_j\rangle \equiv \langle m_j\rangle / (M_s + \sum_{k\in\Psi}\langle m_k\rangle) \approx \langle m_j\rangle / M_s$ is the equilibrium concentration[13] of complex species $j\in\Psi$ and $g_{\text{ref}} \equiv -M_s^{-1}\log Q_{\text{ref}} = \sum_{i\in\Psi^0}[x_i^0(1-\log x_i^0)]$. The sharply peaked Gaussian population distributions allow us to equate $\langle m\rangle$ with the population vector m that maximizes $q(m)$ subject to conservation of mass. Alternatively, we may equate $\langle x\rangle$ with the concentrations $x \approx m/M_s$ that minimize $g(x)$ while conserving total strand concentrations $x^0 \approx m^0/M_s$.

The equilibrium concentrations $\langle x\rangle$ for the complexes[14] can therefore be determined by solving the optimization problem

$$\begin{aligned} &\min_x \quad g(x) \\ &\text{subject to} \quad Ax = x^0 \end{aligned} \tag{3.6}$$

for $g(x) : \mathbb{R}^{|\Psi|}_{>0} \to \mathbb{R}$, where the constraint enforces conservation of mass. Expressions (3.4) and (3.5) indicate that the equilibrium concentrations are strictly positive.

3.4. Convexity and Duality. We now seek an efficient, globally convergent algorithm for solving (3.6) to determine the equilibrium concentration of each species of complex. The constraint is linear so the feasible set is convex and the free energy is a strictly convex function of the concentrations [39], as can be observed by noting that the Hessian of $g(x)$ is a diagonal positive definite matrix with entries $[\nabla^2 g(x)]_{jj} = x_j^{-1}$. Hence, (3.6) has at most one solution x^* [5].

Defining the Lagrange multipliers $\lambda \in \mathbb{R}^{|\Psi^0|}$ to enforce mass conservation, the Lagrangian is

$$\mathcal{L}(x,\lambda) = g_{\text{ref}} + x^T(\log x - \log Q - \mathbf{1}) + \lambda^T(x^0 - Ax).$$

Here, and in subsequent expressions, we adopt the convention that $\log x$ and e^x denote the termwise logarithm and exponential of a vector x; $\mathbf{1}$ denotes a vector of ones of the appropriate length. The corresponding dual function has the form

$$h(\lambda) = \inf_x \mathcal{L}(x,\lambda) = g_{\text{ref}} + \lambda^T x^0 - Q^T e^{A^T\lambda}, \tag{3.7}$$

and the dual problem corresponding to (3.6) is the unconstrained optimization problem

$$\max_\lambda \quad h(\lambda) \tag{3.8}$$

with $h(\lambda) : \mathbb{R}^{|\Psi^0|} \to \mathbb{R}$.

Suppose the primal problem (3.6) has optimal value p^* and the dual problem (3.8) has optimal value d^*. For a convex primal problem, if the constraints satisfy the strong Slater conditions (full row rank for A in addition to feasibility), then strong duality holds ($p^* = d^*$) and the Karush–Kuhn–Tucker (KKT) optimality conditions

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = \log x^* - \log Q - A^T \lambda^* = 0, \tag{3.9}$$

$$A\, x^* = x^0 \tag{3.10}$$

are necessary and sufficient for x^* and λ^* to be primal and dual optimal, respectively [16, 5].[15] The constraint matrix A has full row rank because $A_{ij} = \delta_{ij}$ for $j \in \Psi^0$. Primal feasibility is verified by letting $x_j = \epsilon$ for $j \in \Psi \backslash \Psi^0$ and $x_i = x_i^0 - \epsilon \sum_{j \in \Psi \backslash \Psi^0} A_{ij}$ for $i \in \Psi^0$ with $\epsilon > 0$ sufficiently small.

By the following lemma, the Hessian of $h(\lambda)$ is negative definite, so the dual problem (3.8) is strictly concave with at most one solution λ^*. The strong Slater conditions further ensure that d^* is finite and that there exists a corresponding finite λ^* [16, 5].

LEMMA 3.1. *The Hessian $\nabla^2 h(\lambda)$ is real symmetric negative definite.*

Proof. The Hessian entries are given by

$$[\nabla^2 h(\lambda)]_{mn} = -\sum_{j \in \Psi} A_{mj}\, A_{nj}\, Q_j\, \exp\left\{ \sum_{i \in \Psi^0} \lambda_i A_{ij} \right\} \qquad \forall m, n \in \Psi^0,$$

so the Hessian is real and symmetric by inspection. The Hessian is negative definite if $y^T \nabla^2 h\, y < 0$ for $y \neq 0$. We note that $\nabla^2 h = -R^T R$, where $R \in \mathbb{R}_{\geq 0}^{|\Psi| \times |\Psi^0|}$ has entries $R_{ji} = A_{ij}[Q_j \exp\{\sum_{i \in \Psi^0} \lambda_i A_{ij}\}]^{1/2}$. Hence, $y^T \nabla^2 h\, y = -y^T R^T R\, y = -\|Ry\|^2$, which is negative provided R has linearly independent columns. A has full row rank so R has full column rank and hence $\nabla^2 h(\lambda)$ is negative definite. ∎

We now show that λ^* fully determines x^* so we are free to solve the dual problem (3.8) instead of the primal one (3.6). This is advantageous because the number of complex species $|\Psi|$ can be large even when the number of strand species $|\Psi^0|$ is small (see, e.g., (3.1)). The dual solution λ^* satisfies $\nabla h(\lambda^*) = 0$ or

$$A\, e^{A^T \lambda^* + \log Q} = x^0. \tag{3.11}$$

The first KKT condition (3.9) gives an explicit representation for $x^* \in \mathbb{R}_{>0}^{|\Psi|}$ in terms of λ^*,

$$x^* = e^{A^T \lambda^* + \log Q}, \tag{3.12}$$

and referring to (3.11) we see that the second KKT condition (3.10) is also satisfied. Equating $\langle x \rangle \approx x^*$, the (positive) concentrations corresponding to thermodynamic equilibrium represent the unique solution to (3.6).

Any globally convergent unconstrained optimization algorithm applied to the dual problem (3.8) will suffice to find λ^*. We consider the equivalent dual problem $\min_\lambda f(\lambda)$ with $f(\lambda) \equiv -h(\lambda)$ and apply a trust-region method with a Newton dog-leg step [25] that exploits the symmetric positive definiteness of $\nabla^2 f(\lambda)$ by using Cholesky decomposition for the Newton matrix inversions.[16] For this problem,

the trust-region method converges globally (in arbitrary-precision arithmetic) with quadratic local convergence [37, 25].[17]

3.5. Base-Pairing Observables. We have already shown how to calculate important experimental observables in the form of equilibrium population distributions for small systems and equilibrium concentrations for large systems. Here, we describe the calculation of more detailed base-pairing information about the ensemble of states in a given system.

For a complex of L distinct strands, the equilibrium probability of each intrastrand and interstrand base pair for a given strand ordering π can be calculated by backtracking through the partition function algorithm (Appendix A) applying a particular algorithmic transformation at each step (see [24, 11] for details). The probability of base pair $i_n \cdot j_m$ at equilibrium is simply

$$p(i_n \cdot j_m) = \frac{1}{\overline{Q}} \sum_{\pi \in \overline{\Pi}} \overline{Q}(\pi)\, p(i_n \cdot j_m; \pi), \tag{3.13}$$

where the pair probabilities $p(i_n \cdot j_m; \pi)$ for each $\pi \in \overline{\Pi}$ represent the output of the backtracking recursions.

If a complex contains some indistinguishable strands, we have already seen that distinguishability effects arise at the secondary structure level in the form of rotational symmetry corrections and algorithmic overcounting corrections. When we examine the probabilities of individual base pairs in the ensemble Ω, new distinguishability issues arise. For example, consider a complex involving two indistinguishable copies of strand A (with identifiers 1 and 2) and one copy of strand B (with identifier 3). There is only one distinct circular permutation $\pi = AAB$ and $v(\pi) = 1$ so no symmetry and overcounting corrections are required for any structure $s \in \Omega$. However, base pairs $i_1 \cdot j_3$ and $i_2 \cdot j_3$ are indistinguishable since strands 1 and 2 are both of type A. Likewise, without the global structural context, we cannot distinguish between the inter- and intrastrand base pairs $i_1 \cdot j_2$ and $i_1 \cdot j_1$.

We now develop a quantity analogous to base pair probabilities that appropriately treats the indistinguishability of strands in a complex. Let Θ be the set of strand species in the complex and $\{\theta\}$ be the set of all strand identifiers corresponding to strands of type $\theta \in \Theta$ (hence $L = \sum_{\theta \in \Theta} |\{\theta\}|$).

We define the expected number of base pairs between base i on strands of type $A \in \Theta$ and base j on strands of type $B \in \Theta$ to be $E(i_{\{A\}} \cdot j_{\{B\}}) \in [0, \min(|\{A\}|, |\{B\}|)]$. For a given strand ordering π,

$$E(i_{\{A\}} \cdot j_{\{B\}}; \pi) = \sum_{l_A \in \{A\}} \sum_{l_B \in \{B\}} p(i_{l_A} \cdot j_{l_B}; \pi)$$

represents a sum over the contributions of each type of distinct base pair, where each term $p(i_{l_A} \cdot j_{l_B}; \pi)$ is an output of a backtracking recursion.[18] The expected value for each type of distinct base pair in the complex is then given by

$$E(i_{\{A\}} \cdot j_{\{B\}}) = \frac{1}{Q} \sum_{\pi \in \Pi} Q(\pi)\, E(i_{\{A\}} \cdot j_{\{B\}}; \pi). \tag{3.14}$$

This result can be used to calculate experimental observables for a dilute solution of complexes at equilibrium. For each species of complex $k \in \Psi$, (3.14) is used to calculate the corresponding $E_k(i_{\{A\}} \cdot j_{\{B\}})$, representing the expectation value that base i of strand species $A \in \Theta_k$ pairs to base j of strand species $B \in \Theta_k$, where $\Theta_k \subseteq \Psi^0$ denotes the set of strand species that appear in complex k. For a mixture of strands at equilibrium, the expected concentration of base pairs between base i of strands of type A and base j of strands of type B is

$$\langle x(i_A \cdot j_B) \rangle = \sum_{k \in \Psi} E_k(i_{\{A\}} \cdot j_{\{B\}}) \, \langle x_k \rangle \, .$$

For experimental studies, it is usually more convenient to measure the expected fraction of A strands or B strands that form this base pair: $f_A(i_A \cdot j_B) = \langle x(i_A \cdot j_B) \rangle / x^0_A$ and $f_B(i_A \cdot j_B) = \langle x(i_A \cdot j_B) \rangle / x^0_B$, respectively. Similarly, the expected concentration $\langle x(i_A) \rangle$ of strand species $A \in \Psi^0$ with base i paired to any other base is

$$\langle x(i_A) \rangle = \sum_{B \in \Psi^0} \sum_{j=1}^{N_B} \langle x(i_A \cdot j_B) \rangle \, ,$$

and the expected fraction of A strands that have base i paired is $f_A(i_A) = \langle x(i_A) \rangle / x^0_A$.

4. Conclusions. These algorithms provide new tools for analyzing the equilibrium properties of interacting nucleic acid strands in synthetic and biological systems. For a complex of an arbitrary number of strands, it is now possible to calculate the partition function over all unpseudoknotted connected secondary structures. The approach rigorously treats representation and distinguishability effects that arise in the multistranded setting and provides a basis for the analysis of dilute solutions containing multiple complexes at equilibrium.

If the number of strands and complexes is small, the partition function can be used to calculate the equilibrium population distribution of each complex species. Alternatively, if the number of strands is large, the equilibrium concentration of each complex species can be determined by minimizing the free energy of the system subject to conservation of mass. This is a (high-dimensional) strictly convex programming problem; strong duality and the special form of the KKT conditions imply that we may instead solve the (low-dimensional) unconstrained strictly concave dual problem, leading to an efficient solution framework with uniqueness and global convergence guarantees. Partition function information can then be used to calculate base-pairing expectations for individual complexes or for the system as a whole. Software for performing all of these calculations is available for research purposes at nupack.org.

Figure 4.1 characterizes the size of system for which it is practical to use these methods to analyze the thermodynamics of interacting nucleic acid strands. These results are encouraging, indicating that it is reasonable to analyze dilute solutions containing, for example, strands of length 40 bases that interact to form complexes of up to 10 strands for 2 strand species, complexes of up to 5 strands for 5 strand species, or complexes of up to 4 strands for 10 strand species. After calculating the partition functions for all complexes in a system, the low-dimensional dual formulation makes it very inexpensive to evaluate equilibrium concentrations, $\langle x \rangle$, for many different total strand concentrations, x^0.

The preceding observations on convexity and duality are more general than the present context of aggregating nucleic acid strands—they apply equally well to the

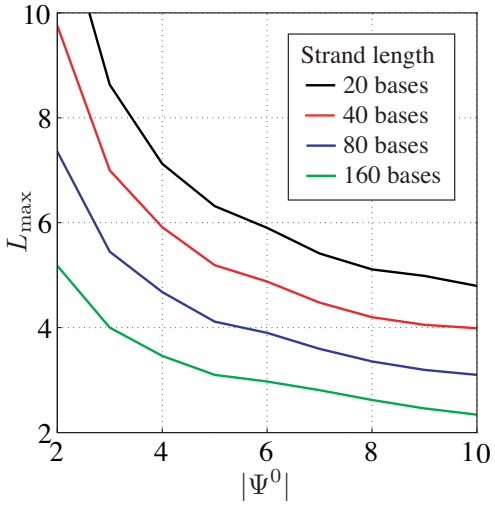


Fig. 4.1 *Computational results demonstrating the size of problem that can be solved in one hour on a single 3 GHz Intel Xeon processor. For a solution containing $|\Psi^0|$ strand species, each with a different random sequence of uniform length (20, 40, 80, or 160 bases), the partition function is calculated at 37° C for all possible complexes of up to $L_{\max}$ strands. Each curve indicates $(|\Psi^0|, L_{\max})$ values corresponding to one hour of wall clock time for strands of a particular length (based on the mean timings for three different sets of sequences). For each of the solutions considered in generating these curves, the trust-region method returns the equilibrium concentrations for all complexes in no more than one second (using $x_i^0 = 10^{-9}\ \forall\ i \in \Psi^0$).*

calculation of equilibrium concentrations for chemical species interacting in a dilute solution (with elements replacing strands and molecules replacing complexes). The convex structure of this classical optimization problem has been noted [39], but the framework of Lagrange duality appears to have been neglected until recently [17].

While we have chiefly exploited the mathematical properties of convexity, it is also interesting to consider its physical significance. The equilibrium and kinetic properties of a nucleic acid system are determined by the features of the underlying free energy landscape [27, 6]. A free energy landscape based on nucleic acid secondary structure may be represented as a graph with each vertex corresponding to a different state of the system (i.e., the pairing status of every base in the system) and each edge corresponding to an elementary step between states (e.g., formation, breakage, or shifting of a single base pair [13]). States that are likely at equilibrium are represented by deep valleys in the landscape, and the rate of conversion between two different states is dependent on the nature of the valleys and ridges that separate them in the landscape. No underlying convexity is evident in this discrete free energy landscape. Now suppose we coarse-grain the state space so that each state corresponds to a different set of complexes (with the fine-grained base-pairing information captured by the ensemble and partition function of each complex). In the thermodynamic limit of large species populations, the free energy landscape may be treated as continuous in the complex concentrations, in which case it becomes convex.

Further work is required to develop algorithms for simulating the kinetics of interacting nucleic acid strands. Additional work is also required to develop thermodynamic and kinetic algorithms for strands that interact to form pseudoknots.

Appendix A. Pseudocode for the Multistranded Partition Function Algorithm.

Initialize (Q, Q^b, Q^m) // $\mathcal{O}(N^2)$ space
Set all values to 0 except $Q_{i,i-1} = 1$ for $i = 1, \ldots, N$
for $l = 1, N$
  for $i = 1, N-l+1$
    $j = i+l-1$
    // Q^b **recursion equations**
    if $\eta[i+\frac{1}{2}, j-\frac{1}{2}] == 0$
      $Q^b_{i,j} = \exp\{-\Delta G^{\text{hairpin}}_{i,j}/kT\}$
    for $d = i+1, j-2$ // loop over all possible $3'$-most pairs $d{\cdot}e$
      for $e = d+1, j-1$
        if $\eta[i+\frac{1}{2}, d-\frac{1}{2}] == 0$ **and** $\eta[e+\frac{1}{2}, j-\frac{1}{2}] == 0$
          $Q^b_{i,j}$ += $\exp\{-\Delta G^{\text{interior}}_{i,d,e,j}/kT\}\, Q^b_{d,e}$
        if $\eta[e+\frac{1}{2}, j-\frac{1}{2}] == 0$ **and** $\eta[i+\frac{1}{2}] == 0$ **and** $\eta[d-\frac{1}{2}] == 0$ // multiloop: no top-level nicks
          $Q^b_{i,j}$ += $Q^m_{i+1,d-1}\, Q^b_{d,e}\, \exp\{-[\Delta G^{\text{multi}}_{\text{init}} + 2\Delta G^{\text{multi}}_{\text{bp}} + (j-e-1)\Delta G^{\text{multi}}_{\text{base}}]/kT\}$
    for $c \in \{i, \ldots, j-1\}$ s.t. $\eta[c+\frac{1}{2}] == 1$ // loop over all top-level nicks $\in [i+\frac{1}{2}, j-\frac{1}{2}]$
      if $(\eta[i+\frac{1}{2}] == 0$ **and** $\eta[j-\frac{1}{2}] == 0)$ **or** $(i == j-1)$ **or**
        $(c == i$ **and** $\eta[j-\frac{1}{2}] == 0)$ **or** $(c == j-1$ **and** $\eta[i+\frac{1}{2}] == 0)$ **then**
        $Q^b_{i,j}$ += $Q_{i+1,c}\, Q_{c+1,j-1}$ // exterior loops
    // Q, Q^m **recursion equations**
    if $\eta[i+\frac{1}{2}, j-\frac{1}{2}] == 0$ **then** $Q_{i,j} = 1$ // empty substructure
      else $Q_{i,j} = 0$ // unconnected substructure
    for $d = i, j-1$ // loop over all possible $3'$-most pairs $d{\cdot}e$
      for $e = d+1, j$
        if $\eta[e+\frac{1}{2}, j-\frac{1}{2}] == 0$
          if $\eta[d-\frac{1}{2}] == 0$ **or** $d == i$
            $Q_{i,j}$ += $Q_{i,d-1}\, Q^b_{d,e}$
          if $\eta[i+\frac{1}{2}, d-\frac{1}{2}] == 0$
            $Q^m_{i,j}$ += $\exp\{-[\Delta G^{\text{multi}}_{\text{bp}} + (d-i)\Delta G^{\text{multi}}_{\text{base}} + (j-e)\Delta G^{\text{multi}}_{\text{base}}]/kT\}\, Q^b_{d,e}$ // single pair in Q^m
          if $\eta[d-\frac{1}{2}] == 0$
            $Q^m_{i,j}$ += $Q^m_{i,d-1}\, Q^b_{d,e}\, \exp\{-[\Delta G^{\text{multi}}_{\text{bp}} + (j-e)\Delta G^{\text{multi}}_{\text{base}}]/kT\}$ // more than one pair in Q^m
return $\left[Q_{1,N} \exp\{-(L-1)\Delta G^{\text{assoc}}/kT\}\right]$ // partition function $\overline{Q}(\pi)$ for ordering π

Fig. A.1 *Pseudocode for an $\mathcal{O}(N^4)$ algorithm for calculating the partition function $\overline{Q}(\pi)$ for circular permutation π for a complex of L strands with total length N. The recursion diagrams and equations corresponding to this dynamic programming implementation were introduced in Figures* 1.2 *and* 2.3. *Nicks between strands are denoted by half indices (e.g., $c+\frac{1}{2}$). The function $\eta[i+\frac{1}{2}, j+\frac{1}{2}]$ returns the number of nicks in the interval $[i+\frac{1}{2}, j+\frac{1}{2}]$. The shorthand $\eta[i+\frac{1}{2}]$ is equivalent to $\eta[i+\frac{1}{2}, i+\frac{1}{2}]$, and by convention, $\eta[i+\frac{1}{2}, i-\frac{1}{2}] = 0$. For clarity, the implementation details for incorporating dangle free energies* [34] *and penalties for helices not ending in C-G pairs are omitted; these terms are treated as in previous work* [10]. *This implementation has been validated by direct comparison with partition functions calculated by explicit enumeration for small multistranded cases and by a previously validated partition function algorithm for large single-stranded cases.*

Appendix B. Notes.

[1]These energies are reported as standard state free energies ΔG° corresponding to 1 mol/liter NaCl and 37°C.

[2]See [10] for an extended description of the algorithm, including details on how to reduce the time complexity to $O(N^3)$ [21] by exploiting certain intricacies of the loop-based free energy model.

[3]Using the standard energy model for dynamic programs [31, 23], hairpin and interior loop free energies are treated as the black-box functions $\Delta G^{\text{hairpin}}_{i,j}$ and $\Delta G^{\text{interior}}_{i,d,e,j}$, but multiloop free energies are assumed to have the linear form $\Delta G^{\text{multi}} = \Delta G^{\text{multi}}_{\text{init}} + n_{\text{bp}}\Delta G^{\text{multi}}_{\text{bp}} + \Delta G^{\text{multi}}_{i,j}$, where $\Delta G^{\text{multi}}_{\text{init}}$ is the penalty for formation of a multiloop, $\Delta G^{\text{multi}}_{\text{bp}}$ is the penalty for each base pair that borders the interior of the multiloop, and $\Delta G^{\text{multi}}_{i,j} = (j-i+1)\Delta G^{\text{multi}}_{\text{base}}$ is a penalty for each unpaired base inside the multiloop. This model allows the partition function contribution of a multiloop to

be calculated incrementally without knowing the full structure of the multiloop at any point in the recursive process. More accurate logarithmic multiloop models can be used to evaluate the energy of specific structures [23]. Note that stacked bases and bulge loops are special cases of interior loops and so do not appear explicitly in the recursions. Exterior loops do not contribute to the free energy because by definition $\Delta G^{\text{empty}} = 0$. We omit details of dangle free energies [34] (corresponding to the energetic contributions of any unpaired bases adjacent to a base pair) and penalties for helices not ending in G-C pairs.

[4]The expressions $i_n \cdot j_m$ and $j_m \cdot i_n$ denote the same base pair. For convenience, we adopt the convention that the bases in a pair are ordered first by strand identifier and then by position on the strand.

[5]The free energy $\Delta G = \Delta H - T\Delta S$ can be decomposed into enthalpic (ΔH) and entropic (ΔS) contributions. The entropy of a system with Γ states at the same energy (in this case, distinct orientations of a complex with a given secondary structure) is given by $k \log \Gamma$ [18], so a reduction of the number of states by a factor of R alters the entropy by $-k \log R$ and ΔG by $+kT \log R$.

[6]The free energy of the complex, ΔG, based on the partition function over the ensemble Ω should not be confused with $\Delta G(s)$, the free energy of a particular secondary structure $s \in \Omega$, which is computed from empirically measured loop free energies but is based conceptually on the partition function over the ensemble of tertiary structures consistent with s.

[7]The reverse is not necessarily true because MFE determination recursions may contain redundancies and partition function recursions must not.

[8]For many physical systems, significant concentrations will be observed only for small complexes due to the entropic cost of strand association. For such systems, an effective strategy is to start with a small value of $L_{\max}$, calculate the probability distributions, increment $L_{\max}$, and then recalculate the distributions to check that there are no significant changes, repeating this process if necessary. This strategy will not work for crystals and polymerization reactions for which there is a substantial nucleation barrier (requiring a critical complex size to be achieved before further aggregation becomes energetically favorable).

[9]This is equivalent to the number of ways to distribute $L_{\max}$ indistinguishable balls amongst $|\Psi^0| + 1$ distinct urns [19].

[10]This is equivalent to the number of distinct necklaces with size $1 \le L \le L_{\max}$ that can be made from $|\Psi^0|$ types of beads [19].

[11]Expression (3.3) for Q_{box} remains valid if all strands are distinct (and solvent molecules are still indistinguishable). This is just the special case where $m_i^0 = 1 \; \forall \; i \in \Psi^0$ and $m_j \in \{0, 1\} \; \forall \; j \in \Psi$. In this case, no distinguishability correction is required when calculating the complex partition functions because $v(\pi) = 1$ for all strand orderings $\pi \in \Pi$ of any complex.

[12]Our convention that the strands in the box can interact to form the set of strand complexes in Ψ implies $m_j > 0 \; \forall \; j \in \Psi$ for at least one $m \in \Lambda$. Hence, $\langle m_j \rangle > 0 \; \forall \; j \in \Psi$.

[13]Based on dimensional analysis, we define our concentrations as mole fractions rather than molarities. Therefore, the free energy of strand association for a complex of L strands is $(L-1)\Delta G^{\text{assoc}} = (L-1)\{\Delta G^{\text{assoc}}_{\text{pub}} - kT \log[\rho_{\text{H}_2\text{O}}/(1 \text{ mol/liter})]\}$, where $\Delta G^{\text{assoc}}_{\text{pub}}$ is the published value for two strands associating [4] and $\rho_{\text{H}_2\text{O}}$ is the molarity of water (e.g., $\rho_{\text{H}_2\text{O}} = 55.14$ mol/liter at 37°C). For the concentration determination problem, in which Q_{box} is approximated by the largest term in the sum of (3.3), this change is merely cosmetic because all factors of $\rho_{\text{H}_2\text{O}}$ ultimately cancel. However, the same is not true when enumerating Q_{box} using the full sum because different terms yield different factors of $\rho_{\text{H}_2\text{O}}$. For dimensional consistency, we use the adjusted association penalty for all formulations. At the end of a calculation, the molarity of species j can be found using $[j] = x_j \rho_{\text{H}_2\text{O}}$.

[14]The equilibrium concentration of the ordered complex corresponding to distinct circular permutation $\pi \in \Pi$ of complex j is simply $\langle x_j(\pi) \rangle = \langle x_j \rangle \, Q_j(\pi)/Q_j$, where the ordered complex is identified with the subensemble $\Omega_j(\pi)$.

[15]To reveal the familiarity of (3.9), it is helpful to note that for $j \in \Psi^0$, $A_{ij} = \delta_{ij}$, so $\lambda_j = \log(x_j/Q_j)$. Substitution of these expressions into the remaining x_j equations yields the standard product/reactant equilibrium equations $x_j = K_j \prod_{i \in \Psi^0} x_i^{A_{ij}} \; \forall \; j \in \Psi \backslash \Psi^0$ with equilibrium constant $K_j = Q_j / \prod_{i \in \Psi^0} Q_i^{A_{ij}}$ (e.g., for complex AAB, the equilibrium expression is $x_{AAB} = K_{AAB} x_A^2 x_B$ with $K_{AAB} = Q_{AAB}/(Q_A^2 Q_B)$). In combination with the mass-conservation constraints (3.10), this $|\Psi|$-dimensional root-finding problem represents the classical formulation for determining the equilibrium concentrations of chemical species reacting in a dilute solution [18].

[16]Implementation details are provided as comments in the source code, which is freely available for research purposes at nupack.org.

[17]This result follows for a function $f(\lambda)$ that is twice continuously differentiable and bounded below with $\nabla^2 f(\lambda)$ Lipschitz continuous and $\|\nabla^2 f(\lambda)\| \le \beta$ on the level set $\mathcal{S} \equiv \{\lambda \,|\, f(\lambda) \le f(\lambda_0)\}$, where λ_0 is the initial guess [37, 25]. In our case, $f(\lambda)$ is infinitely differentiable. The strong

Slater conditions ensure that $f(\lambda^*)$ is finite and furthermore that λ^* is finite [16, 5]. By Lemma 3.1, the Hessian $\nabla^2 f(\lambda)$ is positive definite so the outward normal derivative df/dn on the ball $B(\lambda^*, \epsilon) = \{\lambda^* + \epsilon u \,|\, \|u\| = 1\}$ satisfies a bound $df/dn \geq \delta(n) \geq \delta_0 > 0$, where the uniform bound follows from the continuity of $f'(\lambda)$ on the compact set $\overline{B}(\lambda^*, \epsilon)$. The normal derivative continues to increase as we proceed outward from the ball along any normal $n \in \mathbb{R}^{|\Psi^0|}$, so the distance $s(n)$ from λ^* to the boundary of $\mathcal{S}$ satisfies $s(n) \leq \epsilon + [f(\lambda_0) - f(\lambda^*)]/\delta_0$. Hence, the level set $\mathcal{S}$ is bounded. The continuity of the Hessian entries $[\nabla^2 f(\lambda)]_{ij}$ ensures that they are bounded on the closure of the bounded set $\mathcal{S}$ (say, $\max_{\lambda \in \mathcal{S}} [\nabla^2 f(\lambda)]_{ij} \leq \alpha \ \forall \ i, j \in \Psi^0$). Hence, $\mathrm{tr}(\nabla^2 f(\lambda)) = \sigma^T \mathbf{1} \leq \alpha |\Psi^0| \equiv \beta$, where $\sigma(\lambda) : \mathbb{R}^{|\Psi^0|} \to \mathbb{R}^{|\Psi^0|}_{>0}$ denotes the eigenvalues of $\nabla^2 f(\lambda)$. For a symmetric positive definite Hessian we have $\|\nabla^2 f(\lambda)\| = \sigma_{\max}$ and hence $\|\nabla^2 f(\lambda)\| \leq \beta$ on $\mathcal{S}$.

[18]Calculation of $p(i_{l_A} \cdot j_{l_B}; \pi)$ for a strand ordering with $v(\pi) > 1$ does not require an explicit correction (2.2) for symmetry or overcounting effects. To see this, consider again the complex of four indistinguishable strands with $\pi = AAAA$ and $v(\pi) = 4$; the algorithm computes $\overline{Q}(\pi)$, which is four times the desired $Q(\pi)$. During the reverse pass, this partition function is in the denominator as the backtracking algorithm calculates the probability of each base pair. For a structure with 4-fold rotational symmetry, the absence of the symmetry correction in the physical model causes the numerator to be four times too large so that the correct base-pairing probability is recovered. For a structure with 2-fold rotational symmetry, there are two indistinguishable versions of each base pair, but the absence of the symmetry correction causes each of them to be calculated at half the desired value; the distinct base-pairing expectations then recover the desired value by summing these contributions. Finally, for a structure with no rotational symmetry, four indistinguishable versions of a base pair are each calculated to be one-quarter of the desired value and then summed to give the appropriate base-pairing expectation. Hence, symmetry effects are accounted for by the use of both a forward and a reverse pass of the algorithm and overcounting effects are accounted for by summation to obtain distinct base-pairing expectations.

Acknowledgments. We wish to thank Z.-G. Wang, M. Cook, and L. B. Pierce for helpful discussions during the course of the work.

REFERENCES

[1] T. Akutsu, *Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots*, Discrete Appl. Math., 104 (2000), pp. 45–62.
[2] M. Andronescu, Z.C. Zhang, and A. Condon, *Secondary structure prediction of interacting RNA molecules*, J. Mol. Biol., 345 (2005), pp. 987–1001.
[3] T.S. Bayer and C.D. Smolke, *Programmable ligand-controlled riboregulators of eukaryotic gene expression*, Nat. Biotechnol., 23 (2005), pp. 337–343.
[4] V.A. Bloomfield, D.M. Crothers, and I. Tinoco, Jr., *Nucleic Acids: Structures, Properties, and Functions*, University Science Books, Sausalito, CA, 2000.
[5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
[6] S.-J. Chen and K.A. Dill, *RNA folding energy landscapes*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 646–651.
[7] R.A. Dimitrov and M. Zuker, *Prediction of hybridization and melting for double-stranded nucleic acids*, Biophys. J., 87 (2004), pp. 215–226.
[8] Y. Ding and C.E. Lawrence, *A statistical sampling algorithm for RNA secondary structure prediction*, Nucleic Acids Res., 31 (2003), pp. 7280–7301.
[9] R.M. Dirks, M. Lin, E. Winfree, and N.A. Pierce, *Paradigms for computational nucleic acid design*, Nucleic Acids Res., 32 (2004), pp. 1392–1403.
[10] R.M. Dirks and N.A. Pierce, *A partition function algorithm for nucleic acid secondary structure including pseudoknots*, J. Comput. Chem., 24 (2003), pp. 1664–1677.
[11] R.M. Dirks and N.A. Pierce, *An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots*, J. Comput. Chem., 25 (2004), pp. 1295–1304.
[12] R.M. Dirks and N.A. Pierce, *Triggered amplification by hybridization chain reaction*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 15275–15278.
[13] C. Flamm, W. Fontana, I.L. Hofacker, and P. Schuster, *RNA folding at elementary step resolution*, RNA, 6 (2000), pp. 325–338.
[14] J.A. Gallian, *Contemporary Abstract Algebra*, Houghton Mifflin, New York, 2002.
[15] L. Good, *Diverse antisense mechanisms and applications*, Cell. Mol. Life Sci., 60 (2003), pp. 823–824.

[16] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Springer, New York, 1993.
[17] I.K. Karpov, K.V. Chudnenko, and D.A. Kulik, *Modeling chemical mass transfer in geochemical processes: Thermodynamic relations, conditions of equilibria, and numerical algorithms*, Amer. J. Sci., 297 (1997), pp. 767–806.
[18] L.D. Landau and E.M. Lifshitz, *Statistical Physics Part* 1, 3rd ed., Butterworth-Heinemann, New York, 1980.
[19] L. Lovász, *Combinatorial Problems and Exercises*, Elsevier, Amsterdam, The Netherlands, 1993.
[20] R.B. Lyngsø and C.N.S. Pedersen, *RNA pseudoknot prediction in energy-based models*, J. Comput. Biol., 7 (2000), pp. 409–427.
[21] R.B. Lyngsø, M. Zuker, and C.N.S. Pedersen, *Fast evaluation of internal loops in RNA secondary structure prediction*, Bioinformatics, 15 (1999), pp. 440–445.
[22] D.H. Mathews, *Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization*, RNA, 10 (2004), pp. 1178–1190.
[23] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner, *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure*, J. Mol. Biol., 288 (1999), pp. 911–940.
[24] J.S. McCaskill, *The equilibrium partition function and base pair binding probabilities for RNA secondary structure*, Biopolymers, 29 (1990), pp. 1105–1119.
[25] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer, New York, 1999.
[26] R. Nussinov, G. Pieczenik, J.R. Griggs, and D.J. Kleitman, *Algorithms for loop matchings*, SIAM J. Appl. Math., 35 (1978), pp. 68–82.
[27] J.N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes, *Theory of protein folding: The energy landscape perspective*, Annu. Rev. Phys. Chem., 48 (1997), pp. 545–600.
[28] V. Patzel, S. Rutz, I. Dietrich, C. Köberle, A. Sheffold, and S.H.E. Kaufmann, *Design of siRNAs producing unstructured guide-RNAs results in improved RNA interference efficiency*, Nat. Biotechnol., 23 (2005), pp. 1440–1444.
[29] R. Penchovsky and R.R. Breaker, *Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes*, Nat. Biotechnol., 23 (2005), pp. 1424–1433.
[30] E. Rivas and S.R. Eddy, *A dynamic programming algorithm for RNA structure prediction including pseudoknots*, J. Mol. Biol., 285 (1999), pp. 2053–2068.
[31] J. SantaLucia, Jr., *A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 1460–1465.
[32] N.C. Seeman, *From genes to machines: DNA nanomechanical devices*, Trends Biochem. Sci., 30 (2005), pp. 119–125.
[33] N.C. Seeman and P.S. Lukeman, *Nucleic acid nanostructures: Bottom-up control of geometry on the nanoscale*, Reports Prog. Phys., 68 (2005), pp. 237–270.
[34] M.J. Serra and D.H. Turner, *Predicting thermodynamic properties of RNA*, Methods Enzymol., 259 (1995), pp. 242–261.
[35] W.M. Shih, J.D. Quispe, and G.F. Joyce, *A* 1.7*-kilobase single-stranded DNA that folds into a nanoscale octahedron*, Nature, 427 (2004), pp. 618–621.
[36] J.-S. Shin and N.A. Pierce, *A synthetic DNA walker for molecular transport*, J. Amer. Chem. Soc., 126 (2004), pp. 10834–10835.
[37] G.A. Shultz, R.B. Schnabel, and R.H. Byrd, *A family of trust-region-based algorithms for unconstrained minimization with strong global convergence properties*, SIAM J. Numer. Anal., 22 (1985), pp. 47–67.
[38] F.C. Simmel and W.U. Dittmer, *DNA nanodevices*, Small, 1 (2005), pp. 284–299.
[39] W.R. Smith and R.W. Missen, *Chemical Reaction Equilibrium Analysis: Theory and Algorithms*, John Wiley & Sons, New York, 1982.
[40] M.N. Stojanovic and D. Stefanovic, *A deoxyribozyme-based molecular automaton*, Nat. Biotechnol., 21 (2003), pp. 1069–1074.
[41] I. Tinoco, Jr., O.C. Uhlenbeck, and M.D. Levine, *Estimation of secondary structure in ribonucleic acids*, Nature, 230 (1971), pp. 362–367.
[42] A.J. Turberfield, J.C. Mitchell, B. Yurke, A.P. Mills, Jr., M.I. Blakey, and F.C. Simmel, *DNA fuel for free-running nanomachines*, Phys. Rev. Lett., 90 (2003), article 118102.
[43] F.H.D. van Batenburg, A.P. Gultyaev, and C.W.A. Pleij, *Pseudobase: Structural information on RNA pseudoknots*, Nucleic Acids Res., 29 (2001), pp. 194–195.
[44] M.S. Waterman and T.F. Smith, *RNA secondary structure: A complete mathematical analysis*, Math. Biosci., 42 (1978), pp. 257–266.
[45] J.D. Watson, T.A. Baker, S.P. Bell, A. Gann, M. Levine, and R. Losick, *Molecular Biology of the Gene*, 5th ed., Benjamin Cummings, San Francisco, 2004.

[46] E. Winfree, F. Liu, L.A. Wenzler, and N.C. Seeman, *Design and self-assembly of two-dimensional DNA crystals*, Nature, 394 (1998), pp. 539–544.

[47] S. Wuchty, W. Fontana, I.L. Hofacker, and P. Schuster, *Complete suboptimal folding of RNA and the stability of secondary structures*, Biopolymers, 49 (1999), pp. 145–165.

[48] T.B. Xia, J. SantaLucia, M.E. Burkard, R. Kierzek, S.J. Schroeder, X.Q. Jiao, C. Cox, and D.H. Turner, *Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs*, Biochemistry, 37 (1998), pp. 14719–14735.

[49] M. Zuker and D. Sankoff, *RNA secondary structures and their prediction*, Bull. Math. Biol., 46 (1984), pp. 591–621.

[50] M. Zuker and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information*, Nucleic Acids Res., 9 (1981), pp. 133–147.