JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 9, 2006

Pp. 1640–1657

Combinatorics of Saturated Secondary Structures of RNA

P. CLOTE

ABSTRACT

Following Zuker (1986), a saturated secondary structure for a given RNA sequence is a secondary structure such that no base pair can be added without violating the definition of secondary structure, e.g., without introducing a pseudoknot. In the Nussinov-Jacobson energy model (Nussinov and Jacobson, 1980), where the energy of a secondary structure is -1 times the number of base pairs, saturated secondary structures are local minima in the energy landscape, hence form kinetic traps during the folding process. Here we present recurrence relations and closed form asymptotic limits for combinatorial problems related to the number of saturated secondary structures. In addition, Python source code to compute the number of saturated secondary structures having *k* base pairs can be found at the *web servers* link of *bioinformatics.bc.edu/clotelab/*.

Key words: generating functions, recurrence relations, RNA, saturated secondary structure.

1. INTRODUCTION

In recent years, biologists have discovered that ribonucleic acid (RNA) molecules play a surprising and important regulatory role in the cell. Apart from well-known messenger RNA and transfer RNA, it is now understood that RNA performs diverse biological functions, including *retranslation* of the genetic code [selenocysteine insertion (Böck et al., 1991; Heider et al., 1992), ribosomal frameshift (Moon et al., 2004)], post-transcriptional regulation via small interfering RNA and microRNA (Tuschl, 2003; Lim et al., 2003). In addition, there are conformational switches (Voss et al., 2004), metabolite-sensing *riboswitches* which interact with small ligands and up- or down-regulate certain genes (Barrick et al., 2004) and small nucleolar RNAs which guide the methylation of specific ribosomal nucleotides (Schattner et al., 2004).

The function of the RNA examples just cited is known to depend on RNA 3-dimensional structure, which itself is largely determined by the *secondary structure* (Banerjee et al., 1993). Secondary structure, defined in Section 2, is essentially a well-balanced parenthesis expression in the alphabet {'(', ')', '.'} consisting of left and right parenthesis together with a dot. Just as the Catalan numbers correspond to the number of balanced parenthesis expressions (without dot), the *Motzkin numbers* correspond to the number of balanced parenthesis expressions with dot. The number $S(n)$ of secondary structures on sequence $1, \ldots, n$, was asymptotically computed by Stein and Waterman (1979), by using generating function theory; see also Chapter 13 of Waterman (1995).

The recurrence relations developed by Stein and Waterman (1979) have been extended by (Nussinov and Jacobson, 1980) and especially by Zuker and Stiegler (1981) to dynamic programming algorithms

Department of Biology, Boston College, Chestnut Hill, Massachusetts.

which compute the minimum free energy (mfe) secondary structure of a given RNA sequence. Most current algorithmic work on RNA secondary structure concerns the thermodynamic equilibrium mfe structure [`mfold` (Zuker and Stiegler, 1981; Zuker, 2003), `RNAfold` (Hofacker et al., 1994), `RNAstructure` (Mathews et al., 2000)], the low energy ensemble of structures [`Sfold` (Ding and Lawrence, 2003), `RNAsubopt` (Wuchty et al., 1999)], multiple sequence/structure alignment [`Foldalign` (Havgaard et al., 2005), `Dynalign` (Mathews and Turner, 2002)].

Before continuing, we must insert a remark concerning nomenclature. In 1986, Zuker (1986) introduced the concept of *saturated* secondary structure, defined to be maximal with respect to inclusion of base pairs; i.e., secondary structure S for a given RNA sequence is *saturated* if no base pairs can be added without violating the definition of secondary structure—see Definition 1. In 2001, Evers and Giegerich (2001) used the term *saturated secondary structure* to describe a related, but distinct notion; they defined secondary structure as *saturated* if "stacking regions extend maximally in both directions" and there are no isolated base pairs (i.e., not adjacent to a stacked base pair). With this definition, given 17 nt. RNA sequence `GGGGGUUUUUGGGCCCC`, the secondary structure `((.....)).......` consisting of only base pairs $(1, 10), (2, 9)$ is saturated in the sense of Evers and Giegerich, though not in the sense of Zuker; indeed, this secondary structure is not maximal with respect inclusion, since the structure `((.....))((...))` consisting of base pairs $(1, 10), (2, 9), (11, 17), (12, 16)$ properly includes it.

While the intent of Evers and Giegerich (2001) was clearly to extend Zuker's notion of saturated secondary structure from the Nussinov and Jacobson (1980) energy model to the Turner energy model (Mathews et al., 1999; Xia et al., 1999), this is technically not quite the case. Consider the 11 nt. RNA sequence `AGGGGUUUUUU`, having secondary structure `.((.....)).` consisting of base pairs $(2, 10), (3, 9)$. Using version 1.4 of `RNAeval` from Vienna RNA Package (Hofacker et al., 1994), the free energy for this secondary structure is $+5.2$ kcal/mol. While this secondary structure is not saturated in the sense of Evers and Giegerich, the structure `.(((...))).`, obtained by adding the base pair $(4, 8)$, is indeed saturated in the sense of Evers and Giegerich. However, using `RNAeval`, the free energy for this latter structure is $+5.3$ kcal/mol; i.e., by adding a base pair adjacent to a stack, the free energy, according to the Turner energy model, goes *up* rather than down.

Aware of Evers and Giegerich (2001), though not of Zuker (1986), we defined, in Clote (2005a), the notion of *locally optimal* secondary structure, defined identically as in Zuker's notion of saturated structure. In Clote (2005a, 2005b), we additionally defined k-locally optimal structures to be those which are locally optimal, yet which have k base pairs fewer than that of the Nussinov-Jacobson optimal structure. Given the historical precedence of Zuker's notion of saturation, we now rename k-locally optimal structures to be k-saturated.[1] The main result of Clote (2005a 2005b), was a dynamic programming algorithm running in $O(n^5)$ time and $O(n^3)$ space to compute the number of k-saturated secondary structures for any given RNA sequence of length n. With respect to the Nussinov-Jacobson energy model, saturated secondary structures, defined in the next section, are local minima in the Nussinov-Jacobson energy surface, and hence constitute *kinetic traps* in the folding process, assuming that an RNA molecule finds its mfe structure by a Markovian process of addition and deletion of base pairs.

Having clarified previous confusion in the literature concerning the term of *saturated* secondary structure, we now briefly mention additional results related to the topic of this paper. In Cupal et al. (1996), a $O(n^3m^2)$ time and $O(n^2m)$ space algorithm was described to compute the *density of states* of RNA secondary structures for a given RNA molecule. Here, n is the length of the RNA sequence, and m is the number of energy bins. In Hofacker et al. (1998), asymptotic limits were established for the number of base pairs, number of hairpins, etc. with respect to the collection of *all* secondary structures of random RNA. In Evers and Giegerich (2001), a dynamic programming algorithm was given to compute the number of secondary structures which are *saturated* in the sense of Evers and Giegerich, given any RNA sequence. In Nebel (2002), by using generating function techniques, closed form expressions for the asymptotic number of base pairs, hairpins, etc. were given, where the asymptotic limit was taken over *all* secondary structures of random RNA, In Clote et al. (2005b), an asymptotic limit was proved to exist for the number

[1]The notation $LO(n, k)$ is introduced later in this paper to represent the number of saturated secondary structures on a sequence of length n having k base pairs. Symbols such as LO, an acronym for *locally optimal*, are kept in the formal development, since they were used in a draft of this manuscript prior to knowledge of Zuker (1986).

of base pairs in the Nussinov-Jacobson minimum free energy structure of random RNA. This result was extended in Clote et al. (2005a) to an asymptotic limit of the Turner minimum free energy of random RNA; moreover asymptotic limits were established for all higher order moments for the distribution of free energy in kcal/mol for the Turner minimum free energy secondary structures of random RNA. Note that while Hofacker et al. (1998) and Nebel (2002) concern asymptotic limits taken over all secondary structures of random RNA, the Clote et al. (2005b, 2005a) studies concern asymptotic limits with respect to the minimum free energy structure of random RNA. Although unrelated to asymptotic properties of random RNA, we mention the recent article (Clote et al., 2005) which describes new algorithms to compute the minimum free energy and partition function over all k-point mutants of a given RNA sequence.

In this paper, we employ generating function methods to provide an asymptotic limit for the number of saturated secondary structures, as well as the number of 0-saturated and 1-saturated structures. Section 2 presents the formal definition of secondary structure, saturated secondary structure, and introduces necessary notation to formalize our results. In Section 3, recurrence relations are given to compute the number $LO(n,k)$ [resp. $MO(n,k)$] of saturated structures for a sequence of length n, with k base pairs [resp. additionally with no visible positions]. By dynamic programming, these relations give rise to an algorithm running in time $O(n^4)$ and space $O(n^3)$ to compute $LO(n,k)$ and $MO(n,k)$ (source code in Python is available at the *web servers* link of *bioinformatics.bc.edu/clotelab/*). Section 4 establishes a functional identity for the generating functions associated with the number $LO(n)$ [resp. $MO(n)$] of saturated secondary structures for a sequence of length n. Subsequent application of a general tool from generating function theory, known as Bender's Theorem (Bender, 1974; Meir and Moon, 1989), yields an asymptotic limit for $LO(n)$ and $MO(n)$.[2] While Section 3 yields an algorithm to compute the number $LO_k(n)$ of k-saturated secondary structures for a sequence of length n, it is as yet unclear whether $LO_k(n)$ is *small* (polynomial in n) or *large* (exponential in n) for fixed values of k. In Section 5, we settle this question for the most important values $k = 0, 1$. From a physics standpoint, where saturated secondary structures form *kinetic traps* in the folding process, 0-saturated structures have energy close to that of the minimum free energy structure. Knowledge of the number and distribution of saturated secondary structures is thus important to understand a simplified model for RNA folding.

2. NOTATION

We begin by recalling the formal definition of a secondary structure for a given RNA sequence.

Definition 1. *A secondary structure S on RNA sequence $s_1, \ldots, s_n$ is defined to be a set of ordered pairs (i, j), such that $1 \leq i < j \leq n$ and the following are satisfied.*

1. Watson-Crick or GU wobble pairs: *If (i, j) belongs to S, then pair (s_i, s_j) must be one of the following canonical basepairs: (A, U), (U, A), (G, C), (C, G), (G, U), (U, G).*
2. Threshold requirement: *If (i, j) belongs to S, then $j - i > \theta$, where θ, generally taken to be equal to* 3*, is the minimum number of unpaired bases in a hairpin loop; i.e., there must be at least θ unpaired bases in a hairpin loop.*
3. Nonexistence of pseudoknots: *If (i, j) and (k, ℓ) belong to S, then it is not the case that $i < k < j < \ell$.*
4. No base triples: *If (i, j) and (i, k) belong to S, then $j = k$; if (i, j) and (k, j) belong to S, then $i = k$.*

In this paper, we are interested in the asymptotic number of saturated secondary structures of a sequence of length n, in the same sense that Stein and Waterman (1979) [see also Chapter 13 of Waterman (1995)] provided an asymptotic limit for the number of all secondary structures of a sequence of length n. For such purposes, we assume that any position i can base-pair with any any position j, provide only that

[2] We are indebted to an anonymous referee for pointing out that a counterexample to Bender's Theorem was given in Canfield (1984), and that Meir and Moon (1989) subsequently proved a result, based on Bender's underlying idea, which suffices for our asymptotic limit result in Section 4. Note that the result of Stein and Waterman (1979) on asymptotic number of secondary structures was similarly obtained using Bender's theorem. It follows that similar care, using Meir and Moon's correction of Bender's theorem, is necessary for the Stein-Waterman result.

$|j-i| > \theta$; i.e., condition (1) of Definition 1 is dropped. From this point on, we will speak of a secondary structure S on the sequence $1, \ldots, n$, rather than on the nucleotide sequence $s_1, \ldots, s_n$. For brevity, we may say that S is a secondary structure on n. Since the nature of the nucleotide or base s_i located at position i is not pertinent to the combinatorial study in this paper, by abuse of notation, we may say that i is a *base*.

A position $i \in \{1, \ldots, n\}$ is *visible* in a secondary structure S if for all base pairs $(x, y) \in S$, it is not the case that $x \leq i \leq y$. A visible position is said to be *external* to all base pairs of S. The base pair $(i, j) \in S$ is an *external* base pair in S if there is no base pair $(x, y) \in S$, such that $x < i < j < y$. If S is a secondary structure on sequence $1, \ldots, n$, and $1 \leq x \leq y \leq n$, then the *restriction* of S to $\{x, \ldots, y\}$, denoted by $S \restriction \{x, \ldots, y\}$, is defined by $\{(i, j) \in S : x \leq i < j \leq y\}$.

We now come to the main notion studied in this paper. A secondary structure S for sequence $1, \ldots, n$ is *saturated* if no base pairs can be added without violating the definition of secondary structure, by adding a pseudoknot for instance, i.e., for any $1 \leq i < j \leq n$, if $(i, j) \notin S$ then $S \cup \{(i, j)\}$ is not a valid secondary structure. A secondary structure S is defined to be k-saturated, if S is saturated and additionally S contains k fewer base pairs than the maximum possible number of base pairs.

Define $mbp_\theta(n) = \lfloor \frac{n-\theta}{2} \rfloor$. It is easy to see that the structure $S_0 = \{(k, n+1-k) : 1 \leq k \leq \lfloor \frac{n-\theta}{2} \rfloor\}$ has the maximum number $mbp_\theta(n)$ of base pairs, hence is 0- saturated. Thus it follows that $mbp_\theta(n)$ is the maximum number of base pairs on sequence $1, \ldots, n$. More generally, a saturated secondary structure S on n is k-saturated, if $|S| = mbp_\theta(n) - k$.

Let $LO(n, k)$ [resp., $MO(n, k)$] denote the number of saturated secondary structures on sequence $1, \ldots, n$, which have k base pairs [resp. and there are no *visible* bases]. At times, we may ambiguously refer to $LO(n, k)$ and $MO(n, k)$ as *sets* of secondary structures, rather than the *cardinality* of these sets. With this ambiguous use of term, structures in $MO(n, k)$ are simply those in $LO(n, k)$ which additionally have no visible bases.

Let $LO_k(n)$ denote the number (or ambiguously, the set) of k-saturated secondary structures on $1, \ldots, n$, and let $MO_k(n)$ denote the number (or ambiguously, the set) of k-saturated secondary structures on $1, \ldots, n$, such that there are no visible bases. It follows from definitions that $LO_k(n) = LO(n, mbp_\theta(n) - k)$ and $MO_k(n) = MO(n, mbp_\theta(n) - k)$.

Recall again that throughout this paper, we assume that any position i can base-pair with any other position j, provided only that $|j - i| > \theta$. The value θ is a fixed constant, often explicitly omitted when clear from context. Results in Section 3 hold for arbitary $\theta \geq 1$.

3. COMPUTING SATURATED STRUCTURES WITH k BASE PAIRS

In this section, we provide recurrence relations to compute the number of k-saturated secondary structures. In Clote (2005a), a substantially more complicated algorithm is given, which directly computes by dynamic programming the number of k-saturated secondary structures for a given RNA sequence $s_1, \ldots, s_n$ (respecting requirement 1 of Definition 1).

To study asymptotics, in this paper we have dropped requirement 1 of Definition 1, thus obviating the use of two complicated *visibility* predicates necessary for the algorithm of Clote (2005a). Additionally, we follow a suggestion of R. Bundschuh and D. Mathews (personal communication), and first compute the number of saturated secondary structures having k base pairs. It then follows that the number $LO_k(n)$ of k-saturated secondary structures on a sequence of length n is equal to $LO(n, mbp_\theta(n) - k)$.

While it is biologically unrealistic to drop requirement 1 of Definition 1, by doing so we obtain elegant asymptotic results, which provide information on the potential kinetic traps in the RNA folding process.

Fix $\theta \geq 1$. Throughout the remainder of this section, we suppress explicit notational reference to θ. We now define $M(n, k)$ and $L(n, k)$ simultaneously by double induction on n, k; i.e., outermost induction on n, and for n fixed by inner induction on k. Thus after having defined $M(n, k)$ and $L(n, k)$ for all k by induction on k, we define $M(n+1, 0)$, $L(n+1, 0)$, $M(n+1, 1)$, $L(n+1, 1)$, etc. In Theorem 2, we show that $LO(n, k) = L(n, k)$ and $MO(n, k) = M(n, k)$. Using the recurrence relations for $M(n, k)$ and $L(n, k)$, by dynamic programming it is straightforward to compute numerical values for the number $LO(n, k)$ [resp. $MO(n, k)$] of saturated secondary structures on n having k base pairs [resp. saturated secondary structures on n with no visible positions and having k base pairs].

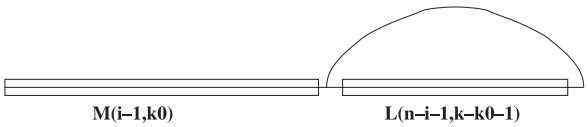

FIG. 1. Contribution to $M(n,k)$ in Equation (1), namely $M(i-1,k_0)\cdot L(n-i-1,k-k_0-1)$, for some $1\le i\le n-\theta-1$. Note that if $i=1$, then $M(i-1,k_0)$ is 0 unless $k_0=0$, in which case $M(i-1,k_0)\cdot L(n-i-1,k-k_0-1)$ equals $L(n-2,k-1)$, i.e., the contribution when there is a base bair $(1,n)$.

For the base case, define

$$M(n,k)=\begin{cases}1 & \text{if } n=0,k=0\\ 0 & \text{if } 1\le n\le\theta+1,k=0\\ 0 & \text{if } k\ge mbp_\theta(n)\end{cases}$$

and for the inductive case, define

$$M(n,k)=\sum_{i=1}^{n-\theta-1}\sum_{k_0=0}^{mbp_\theta(i-1)} M(i-1,k_0)\cdot L(n-i-1,k-k_0-1). \tag{1}$$

Figure 1 depicts a typical term of Equation (1), where the existence of a base pair (i,n) divides the set $MO(n,k)$ of secondary structures into two regions. The contribution from the left region is $MO(i-1,k_0)$ for some $k_0\in\{0,\ldots,mbp_\theta(i-1)\}$, while the contribution from the right region is $LO(n-i-1,k-k_0-1)$. There are k_0 base pairs in the left region, $k-k_0-1$ base pairs in the right region, and one base pair contributed by base pair (i,n). Since the left and right regions are independent, the terms $MO(i-1,k_0)$ and $LO(n-i-1,k-k_0-1)$ are multiplied. For the base case of $L(n,k)$, define

$$L(n,k)=\begin{cases}1 & \text{if } 0\le n\le\theta+1,k=0\\ 0 & \text{if } k\ge mbp_\theta(n)\end{cases}$$

and for the inductive case, define

$$L(n,k)=\sum_{r=n-\theta-1}^{n-1} M(r,k)+\sum_{i=1}^{n-\theta-1}\sum_{k_0=0}^{mbp_\theta(i-1)} L(i-1,k_0)\cdot L(n-i-1,k-k_0-1). \tag{2}$$

The set $LO(n,k)$ of saturated secondary structures on n, having k base pairs, can be divided into two groups: *(i)* those, in which position n is not base-paired, and *(ii)* those, in which n is base-paired. In the former case, if r is the rightmost base-paired position, then $r\ge n-\theta-1$; indeed, otherwise, $r<n-\theta-1$, and the base pair $(r+1,n)$ could be added since $n-1-(r+2)+1=n-r-2\ge\theta$ satisfies condition (2) of Definition 1. The left panel of Figure 2 depicts the case *(i)*.

FIG. 2. ***(i) Left panel:*** Contribution to $L(n,k)$ by the first term of Equation (2), namely $M(r,k)$, for some $n-\theta-1\le r\le n-1$. Note that in this case r must be strictly less than n, for otherwise, there would be a base pair $(1,n)$, which case is considered in the second term of Equation (2). ***(ii) Right panel:*** Contribution to $L(n,k)$ by the second term of Equation (2), namely $L(i-1,k_0)\cdot L(n-i-1,k-k_0-1)$, for some $1\le i\le n-\theta-1$. Note that if $i=1$, then $L(i-1,k_0)$ is 0 unless $k_0=0$, in which case $L(i-1,k_0)\cdot L(n-i-1,k-k_0-1)$ equals $L(n-2,k-1)$, i.e., the contribution when there is a base bair $(1,n)$.

The right panel of Figure 2 depicts the case (*ii*), where the base pair (i, n) divides every remaining structure in $LO(n, k)$ into two regions. The contribution from the left region is $LO(i-1, k_0)$ for some $k_0 \in \{0, \ldots, mbp_\theta(i-1)\}$, while the contribution from the right region is $LO(n-i-1, k-k_0-1)$. There are k_0 base pairs in the left region, $k-k_0-1$ base pairs in the right region, and one base pair contributed by base pair (i, n). Since the left and right regions are independent, the terms $LO(i-1, k_0)$ and $LO(n-i-1, k-k_0-1)$ are multiplied. Finally, an inductive proof, carried out in Theorem 2, then shows that $MO(n, k) = M(n, k)$ and $LO(n, k) = L(n, k)$ for all n, k.

Theorem 2. *For all $0 \le k \le n$, it is the case that $MO(n, k) = M(n, k)$ and $LO(n, k) = L(n, k)$.*

Proof. We establish $LO(n, k) = L(n, k)$ and $MO(n, k) = M(n, k)$ by by simultaneous double induction on n, k.

BASE CASE. When $0 \le n \le \theta + 1$, due to requirement 2 of Definition 1, the only secondary structure is the empty structure. This structure is saturated and has no base pairs, so $LO(n, 0) = 1$. Since there are no structures having any base pairs, $LO(n, k) = 0$ for $k > 0$. Additionally, all bases are visible in the empty structure, if $1 \le n \le \theta + 1$, then $MO(n, k) = 0$ for all k. There are at most $msp_\theta(n) = \lfloor \frac{n-\theta}{2} \rfloor$ base pairs in a secondary structure on sequence $1, \ldots, n$, so $LO(n, k) = 0 = MO(n, k)$ if $k > msp(n)$.

INDUCTIVE CASE. By the induction hypothesis, assume that $MO(n', k') = M(n', k')$ and $LO(n', k') = L(n', k')$, for all $n' < n$ and all k'. The set $MO(n, k)$ of all saturated secondary structures on sequence $1, \ldots, n$, which have k base pairs *and* which have no visible base, can be partitioned into $MO(n, k) = \mathcal{S}_1 \cup \mathcal{S}_2$, where $\mathcal{S}_1 = \{S \in MO(n, k) : (1, n) \in S\}$, $\mathcal{S}_2 = \{S \in MO(n, k) : \exists i[1 < i < n - \theta \wedge (i, n) \in S \wedge S \restriction \{1, \ldots, i-1\} \in MO \wedge S \restriction \{i+1, \ldots, n-1\} \in LO\}$. This is easily seen as follows.

Let S be a saturated secondary structure on sequence $1, \ldots, n$, which has no visible bases. Then one of the following two cases holds.

CASE 1. The base pair $(1, n) \in S$ and $S_0 = S \restriction \{2, \ldots, n-1\}$ is saturated and has $k-1$ bases. Note that S_0 may have visible bases, which however are rendered invisible in S because of the external base pair $(1, n)$.

CASE 2. For some $1 < i < n - \theta$, the base pair $(i, n) \in S$ and $S_0 = S \restriction \{1, \ldots, i-1\}$ has k_0 base pairs, $S_1 = S \restriction \{i+1, \ldots, n-1\}$ has k_1 base pairs, where $k = k_0 + k_1 + 1$, and $S_0 \in MO$ and $S_1 \in LO$. Note that S_1 may have visible bases, since these are made invisible in S because of the external base pair (i, n).

The contribution for Case 1 is given by $L(n-2, k-1)$, while that of Case 2 is given by $\sum_{i=2}^{n-\theta-1} \sum_{k_0=0}^{mbp_\theta(i-1)} M(i-1, k_0) \cdot L(n-i-1, k-k_0-1)$. As discussed in Figure 1, when $i = 1$, $M(i-1, k_0)$ is nonzero only when $k_0 = 0$, in which case $M(i-1, k_0) \cdot L(n-i-1, k-k_0-1) = M(0, 0) \cdot L(n-2, k-1)$. Thus both Case 1 and Case 2 are taken into account in Equation (1). This establishes the inductive case that $MO(n, k) = M(n, k)$. We now establish that $LO(n, k) = L(n, k)$.

The set $LO(n, k)$ of all saturated secondary structures on sequence $1, \ldots, n$, which have k base pairs can be partitioned into $LO(n, k) = \mathcal{S}_1 \cup \mathcal{S}_2$, where $\mathcal{S}_1 = \{S \in LO(n, k) : (1, n) \in S\}$, $\mathcal{S}_2 = \{S \in LO(n, k) : \exists i[1 < i < n - \theta \wedge (i, n) \in S \wedge S \restriction \{1, \ldots, i-1\} \in LO \wedge S \restriction \{i+1, \ldots, n-1\} \in LO\}$. This is easily seen as follows.

Let S be a saturated secondary structure on sequence $1, \ldots, n$. Then one of the following three cases holds.

CASE 1. The base pair $(1, n) \in S$ and $S \restriction \{2, \ldots, n-1\}$ is saturated and has $k-1$ bases. Note that S_0 may have visible bases, which however are rendered invisible in S because of the external base pair $(1, n)$.

CASE 2. The bases $r+1, \ldots, n$ are all visible in S, for some $n - \theta - 1 \le r < n$, and there are no visible bases in $S_0 = S \restriction \{1, \ldots, r\}$. Thus $S_0 \in MO$ and must have k base pairs.

CASE 3. The base pair $(i, n) \in S$ and $S_0 = S \restriction \{1, \ldots, i-1\}$ has k_0 base pairs, $S_1 = S \restriction \{i+1, \ldots, n-1\}$ has k_1 base pairs, where $k = k_0 + k_1 + 1$, and $S_0 \in LO$ and $S_1 \in LO$. Note that S_0 may have visible bases, which then remain visible in S.

The contribution for Case 1 is given by $L(n-1, k-1)$, while that of Case 2 is given by $M(r, k)$, for $r \in \{n-\theta-1, \ldots, n-1\}$, and that of Case 3 by $\sum_{i=2}^{n-\theta-1} \sum_{k_0=0}^{mbp_\theta(i-1)} L(i-1, k_0) \cdot L(n-i-1, k-k_0-1)$. As discussed in Figure 2, when $i = 1$, $L(i-1, k_0)$ is nonzero only when $k_0 = 0$, in which case $L(i-1, k_0) \cdot L(n-i-1, k-k_0-1) = L(0, 0) \cdot L(n-2, k-1)$. Thus Cases 1, 2, 3 are taken into account in Equation (2).

This establishes the inductive case that $LO(n, k) = L(n, k)$, and so the theorem is proved. ■

4. ASYMPTOTIC NUMBER OF SATURATED STRUCTURES

In this section, we establish a functional identity for generating functions associated with the number $LO(n,k)$ [resp. $MO(n,k)$] of saturated secondary structures on n [resp. having no visible positions] with exactly k base pairs. This then implies our main asymptotic limit result, Theorem 9. Though stated for $\theta = 1$, it nevertheless is clear how to generalize all results in this section for any fixed $\theta > 1$.

For a fixed value of θ, define $a_n(\theta)$ to be the number of distinct saturated secondary structures for a sequence of length n. Additionally, define $b_n(\theta)$ to be the number of distinct saturated secondary structures for a sequence of length n, with no visible positions. Clearly, by definition, $a_n(\theta) = \sum_k LO(n,k)$ and $b_n(\theta) = \sum_k MO(n,k)$ (Fig. 3).

Proposition 3. *For non-negative integer n, the following recurrence relation holds:*

$$a_n(\theta) = \begin{cases} 1 & \text{if } 0 \le n \le \theta+1 \\ \displaystyle\sum_{k=1}^{n-\theta-1} a_{k-1}(\theta)\cdot a_{n-k-1}(\theta) + \sum_{k=1}^{\theta+1} b_{n-k} & \text{else} \end{cases} \tag{3}$$

Proof. The proof is by induction on n. Define $a_0 = 1$. This allows us to simplify the recurrence relation in the else-clause of Equation (3).

For $1 \le n \le \theta+1$, since the threshold requirement, clause (2) of Definition 1, requires at least θ unpaired bases in a hairpin loop, the only possible secondary structure for a sequence of length n is the empty set. Due to the threshold requirement, no base pairs can be added to the empty structure for a sequence of length n, where $0 \le n \le \theta+1$, the empty structure is saturated. Thus for $1 \le n \le \theta+1$, we have $a_n = 1$.

When $n > \theta+1$, partition the collection $LO(n)$ of saturated secondary structures for a sequence of length n into two disjoint sets, $LO(n) = \mathcal{S}_0 \cup \mathcal{S}_1$, corresponding to the following cases.

CASE 1. $\mathcal{S}_0$ is the set of saturated secondary structures on a sequence of length n, where n does not base pair. Let $S \in \mathcal{S}_0$. In this case, let $k_0 \in \{0,\ldots,\theta\}$ be the largest value k such that each position $n-k, n-k+1, \ldots, n$ is *visible*. Note that if $k > \theta$ and $n-k,\ldots,n$ are all visible in S, then the base pair $(n-k+1, n)$ could be added to S, and so S would not be saturated. By choice of k_0, the restriction $S \restriction \{1,\ldots,k_0-1\}$ of S to the sequence $1,\ldots,k_0-1$, has no visible positions.

Such restrictions account for the term $b_{n-(k_0+1)}$, hence altogether Case 1 accounts for

$$\sum_{k_0=0}^{\theta} b_{n-(k_0+1)} = \sum_{k=1}^{\theta+1} b_{n-k}$$

many saturated secondary structures on $\{1,\ldots,n\}$.

CASE 2. $\mathcal{S}_1$ is the set of saturated secondary structures such that position n is base paired. Let $S \in \mathcal{S}_1$, and let $k_0 \in \{1,\ldots,n-\theta-1\}$ be such that $(k_0, n) \in S$. Since Item (3) of Definition 1 disallows pseudoknots, S consists of the base pair (k_0, n), together with $S|\{1,\ldots,k_0-1\}$ and $S|\{k_0+1,\ldots,n-1\}$. Applying the induction hypothesis to each of the latter, it follows that there are a_{k_0-1} many saturated secondary structures on $\{1,\ldots,k_0-1\}$ and $a_{n-1-(k_0+1)+1} = a_{n-k_0-1}$ many saturated secondary structures on $\{k_0+1,\ldots,n-1\}$. This accounts for the term

$$\sum_{k=1}^{n-\theta-1} a_{k-1}(\theta)\cdot a_{n-k-1}(\theta)$$

of Equation (3). ■

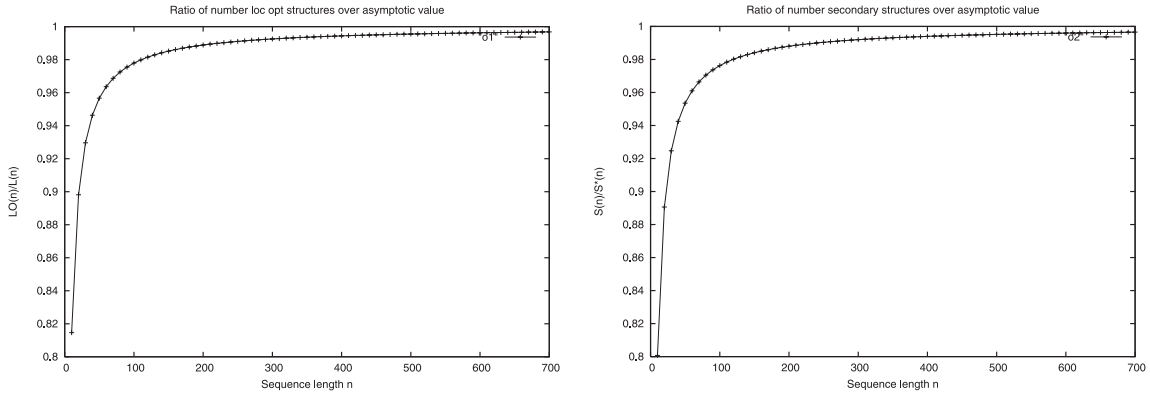

FIG. 3. ***(i) Left panel:*** Ratio of the number $LO(n)$ of saturated secondary structures on a sequence of length n divided by the asymptotic value a_n^*. ***(ii) Right panel:*** Ratio of the number $S(n)$ of secondary structures on a sequence of length n divided by the asymptotic value $S^*(n)$. Convergence is rapid. Superposition of both curves suggests the same rate of convergence of $\frac{a_n}{a_n^*} \to 1$ and $\frac{S(n)}{S^*(n)} \to 1$, although this should be considered a conjecture in the absence of information about precise rate of convergence.

Proposition 4. *For non-negative integer n, the following recurrence relation holds:*

$$b_n(\theta) = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } 1 \le n \le \theta + 1 \\ \displaystyle\sum_{k=1}^{n-\theta-1} b_{k-1}(\theta) \cdot a_{n-k-1}(\theta) & \text{else} \end{cases} \tag{4}$$

Proof. The proof is by induction on n. Define $b_0 = 1$. This allows us to simplify the recurrence relation in the else-clause of Equation (4).

For $1 \le n \le \theta + 1$, since the threshold requirement requires at least θ unpaired bases in a hairpin loop, the only possible secondary structure for a sequence of length n is the empty set. However, since n is positive, there are visible positions. It follows that there are no saturated structures on a sequence of length n, having no visible positions. Thus $b_n = 0$ for $1 \le n \le \theta + 1$.

Suppose that $n > \theta+1$, and that S is a saturated secondary structure on a sequence of length n, which has no visible positions. Let $k_0 \in \{1, \ldots, n-\theta-1\}$ be such that $(k_0, n) \in S$. By nonexistence of pseudoknots, it follows that S consists of the base pair (k_0, n) together with the restrictions $S \restriction \{1, \ldots, k_0 - 1\}$ and $S \restriction \{k_0 + 1, \ldots, n - 1\}$.

By the induction hypothesis, there are b_{k_0-1} many saturated secondary structures on $\{1, \ldots, k_0 - 1\}$ with no visible positions. Additionally, there are $a_{n-1-(k_0+1)+1} = a_{n-k_0-1}$ many saturated secondary structures on $\{k_0 + 1, \ldots, n - 1\}$, where there may or may not be any visible positions. Hence altother, there are $\sum_{k=1}^{n-\theta-2} b_{k-1}(\theta) \cdot a_{n-k-1}(\theta)$ many saturated secondary structures on $\{1, \ldots, n\}$ with no visible positions. ■

At this point, we take the minimum number θ of unpaired bases in a hairpin loop to be 1. Small modifications of the following can be undertaken for the general case of arbitrary fixed value θ. When $\theta = 1$, we write a_n resp. b_n in place of $a_n(\theta)$, resp., $b_n(\theta)$. Define the variables y, z and generating functions f, g as follows:

$$y = f(x) = \sum_{n=0}^{\infty} a_n x^n \tag{5}$$

$$z = g(x) = \sum_{n=0}^{\infty} b_n x^n. \tag{6}$$

Proposition 5. $x^2yz = z - 1 + x^2z$.

Proof.

$$xy = \sum_{n=0}^{\infty} a_n x^{n+1} = \sum_{n=1}^{\infty} a_{n-1} x^n;$$

$$xz = \sum_{n=0}^{\infty} b_n x^{n+1} = \sum_{n=1}^{\infty} b_{n-1} x^n.$$

Thus

$$\begin{aligned} x^2yz &= \left(\sum_{n=1}^{\infty} a_{n-1}x^n\right) \cdot \left(\sum_{n=1}^{\infty} b_{n-1}x^n\right) = \sum_{n=2}^{\infty} \left(\sum_{k=1}^{n-1} b_{k-1} \cdot a_{n-k-1}\right) x^n \\ &= \sum_{n=2}^{\infty} (b_n + b_{n-2} \cdot a_0)\, x^n = \sum_{n=2}^{\infty} (b_n + b_{n-2})\, x^n \\ &= \sum_{n=2}^{\infty} b_n x^n + \sum_{n=2}^{\infty} b_{n-2} x^n = (z - b_1 x - b_0) + x^2 z \\ &= z - 1 + x^2 z \end{aligned}$$

The first part of line is by definition, and the second part of line 1 follows by distributing the sums according to power of x. In the second line, we use the fact that $b_n = \sum_{k=1}^{n-\theta-1} b_{k-1}(\theta) \cdot a_{n-k-1}(\theta)$ from Proposition 4 and that $a_0 = 1$ from Proposition 3. By distributing the sums and applying the definition of variable z, the result follows. ■

Proposition 6. $x^2y^2 = y(x^2 + 1) - z(x^2 + x) - 1$.

Proof. Since

$$xy = \sum_{n=0}^{\infty} a_n x^{n+1} = \sum_{n=1}^{\infty} a_{n-1} x^n$$

we have

$$\begin{aligned} x^2y^2 &= \left(\sum_{n=1}^{\infty} a_{n-1}x^n\right) \cdot \left(\sum_{n=1}^{\infty} a_{n-1}x^n\right) = \sum_{n=2}^{\infty} \left(\sum_{k=1}^{n-1} a_{k-1} \cdot a_{n-k-1}\right) x^n \\ &= \sum_{n=2}^{\infty} \left(a_{n-2} \cdot a_0 + \sum_{k=1}^{n-2} a_{k-1} \cdot a_{n-k-1}\right) x^n \\ &= \sum_{n=2}^{\infty} \left(a_{n-2} + a_n - \sum_{k=1}^{2} b_{n-k}\right) x^n \\ &= \sum_{n=2}^{\infty} (a_n + a_{n-2} - b_{n-1} - b_{n-2})\, x^n \end{aligned}$$

$$= \sum_{n=2}^{\infty} a_n x^n + \sum_{n=2}^{\infty} a_{n-2} x^n - \sum_{n=2}^{\infty} b_{n-2} x^n - \sum_{n=2}^{\infty} b_{n-1} x^n$$

$$= (y - a_1 x - a_0) + x^2 y - x^2 z - x(z - b_0)$$

$$= (y - x - 1) + x^2 y - x^2 z - x(z - 1)$$

$$= y(x^2 + 1) - z(x^2 + x) - 1$$

The first equation follows by definition, telescoping and rearranging sums in terms of powers of x, The third line uses the facts that $a_0 = 1$ and

$$a_n = \sum_{k=1}^{n-2} a_{k-1} \cdot a_{n-k-1} + \sum_{k=1}^{2} b_{n-k}$$

from Proposition 3—here, recall that we assume $\theta = 1$. All remaining lines are straightforward to justify. It follows that $x^2 y^2 = y(x^2 + 1) - z(x^2 + x) - 1$. ■

Note that it immediately follows from Proposition 5 that

$$z = \frac{1}{x^2 - x^2 y + 1}. \tag{7}$$

Let

$$F(x, y) = x^2 y^2 - y(x^2 + 1) + \frac{x^2 + x}{x^2 - x^2 y + 1} + 1 + y. \tag{8}$$

Rewriting, we have

$$F(x, y) = 1 + \frac{x(1 + x)}{1 - x^2(y - 1)} - x^2 y + x^2 y^2. \tag{9}$$

Note that

$$\frac{\partial}{\partial y} F(x, y) = x^2 \left(-1 + \frac{x(1 + x)}{(-1 + x^2(y - 1))^2} + 2y \right). \tag{10}$$

Additionally, note that

$$\frac{\partial}{\partial x} F(x, y) = \frac{1 + x^2(y - 1) - 4x^3(y - 1)^2 y + 2x^5(y - 1)^3 y + 2x(y^2 - y + 1)}{(-1 + x^2(y - 1))^2} \tag{11}$$

and

$$\frac{\partial^2}{\partial y^2} F(x, y) = 2x^2 + \frac{2x^5(1 + x)}{(1 - x^2(y - 1))^3}. \tag{12}$$

Theorem 7. *Let $F(x, y)$ be given as in Equation (8). Then $F(x, y) = y$ holds.*

Proof. By Equation (7), $z = \frac{1}{x^2 + 1 - x^2 y}$. Substitute this expression for z in the equation $x^2 y^2 = y(x^2 + 1) - z(x^2 + x) - 1$ from Proposition 6 to obtain $x^2 y^2 = y(x^2 + 1) - \frac{x^2 + x}{x^2 - x^2 y + 1} - 1$ hence yielding $F(x, y) = y$. ■

Meir-Moon's rectification of Bender's theorem. By $A(f, F, \mathcal{S})$, we denote the following collection of assumptions.

1. The generating function $y = f(x) = \sum_{n=1}^{\infty} c_n x^n$ is a function of complex variable x, such that all coefficients c_n are real and non-negative.
2. There is a functional relation F, satisfying $F(x, y) = y$, which is analytic in complex variables x, y, and whose power series $F(x, y) = \sum_{i+j\geq 1} f_{i,j} x^i y^j$ converges absolutely in some neighborhood of $(0, 0)$.
3. For each i, there exists $j \geq 2$ such that $f_{i,j} > 0$.
4. If $k = \min\{i : f_{i,0} \neq 0\}$, then $f_{k,0} > 0$.
5. $f_{0,1} \neq 1$.

Let $\mathcal{S}$ denote the set of points (ρ, τ), where both ρ, τ are real and positive, such that the following conditions hold.

1. There exist $\epsilon, \delta > 0$, such that $F(x, y)$ converges absolutely for all x, y satisfying $|x| < |\rho| + \epsilon$, $|y| < |\tau| + \delta$.
2. $F(\rho, \tau) = \tau$.
3. $1 = \frac{\partial}{\partial y} F(\rho, \tau)$.

The following theorem is due to Meir and Moon (1989), and immediately follows from Lemma 2 of Meir and Moon (1989) and Darboux's theorem cited in that paper. First, following standard convention, we denote $\frac{\partial^2}{\partial y^2} F(x, y)$ by F_{yy}, and $\frac{\partial}{\partial x} F(x, y)$ by F_x.

Theorem 8 (Meir and Moon, 1989). *Suppose that $A(f, F, \mathcal{S})$ holds, and that there exists a point $(\rho, \tau) \in \mathcal{S}$ such that*

$$F(\rho, \tau) = \tau$$

$$F_y(\rho, \tau) = 1$$

$$F_{y,y}(r, t) > 0$$

$$F_x(r, t) > 0$$

holds for all $0 < r \leq \rho$ and $0 < t \leq \tau$. Then ρ is the radius of convergence of the power series $f(x) = \sum_{n\geq 1} c_n x^n$, $f(\rho) = \tau$, and

$$c_n \sim \sqrt{\frac{r F_x(\rho, \tau)}{2\pi F_{yy}(\rho, \tau)}} \cdot n^{-3/2} \ \rho^{-n}$$

Here, recall that $c_n \sim d_n$ means asymptotic equivalence, i.e.

$$\lim_{n\to\infty} \frac{c_n}{d_n} = 1.$$

Let (ρ_0, σ_0), with $\sigma_0 > 1$, be the simultaneous solution to $F(x, y) = y$ and $F_y(x, y) = 1$, as defined in Equations (8) and (10). Using the function `FindRoot` from *Mathematica*, with initial parameters $x_0 = 0.5$, $y_0 = 1$, we obtain the (approximate) solution $\rho_0 \approx 0.424687$ and $\sigma_0 \approx 2.656896$. A computation shows that the expression $\sqrt{\frac{r F_x(\rho_0, \sigma_0)}{2\pi F_{yy}(\rho_0, \sigma_0)}}$ approximately equals 1.0742707.

Theorem 9. *Let a_n denote the number of saturated secondary structures on a sequence of length n, as defined in Proposition 3. Let*

$$a_n^* = \sqrt{\frac{r F_x(\rho_0, \sigma_0)}{2\pi F_{yy}(\rho_0, \sigma_0)}} \cdot n^{-3/2} \cdot \rho_0^{-n}$$

Then $a_n \sim a_n^*$.

TABLE 1. RATIO $L(n)/S(n)$ OF ASYMPTOTIC NUMBER $L(n)$ OF SATURATED SECONDARY STRUCTURES WITH RESPECT TO ASYMPTOTIC NUMBER $S(n)$ OF ALL SECONDARY STRUCTURES FOR BIOLOGICALLY RELEVANT VALUES OF LENGTH n

n	$L(n)/S(n)$	n	$L(n)/S(n)$	n	$L(n)/S(n)$	n	$L(n)/S(n)$
50	0.00373174	300	1.49154e-15	550	5.96154e-28	800	2.38277e-40
100	1.23668e-05	350	4.94288e-18	600	1.97562e-30	850	7.89635e-43
150	4.09827e-08	400	1.63804e-20	650	6.54708e-33	900	2.6168e-45
200	1.35814e-10	450	5.42837e-23	700	2.16966e-35	950	8.67193e-48
250	4.50081e-13	500	1.79893e-25	750	7.19014e-38	1,000	2.87383e-50

Values obtained by P. Clote's implementation in Python.

Proof. A computation shows that $F_x(x, y) > 0$ and $F_{yy}(x, y) > 0$ holds for all $0 < x \leq \rho_0$ and $0 < y \leq \sigma_0$. Indeed, it is straightforward to see that for $0 \leq x \leq 0.5$ and $0 \leq y \leq 3$, it suffices to show that the sum of the third and fifth term in the numerator of Equation (11) is greater than 0, in order to ensure that $F_x(x, y) > 0$. For this, it suffices that

$$y^2 - y + 1 > 2x^2(y-1)^2 y. \tag{13}$$

Since $x \leq 0.5$, inequality (13) holds if $2(y^2 - y + 1) > (y-1)^2 y$; the latter is easily seen to hold, so it follows that $F_x(x, y) > 0$ for $0 \leq x \leq \rho_0$ and $0 \leq y \leq \sigma_0$.

For the case of $F_{yy}(x, y)$, it is easy to see that $x^2(y-1) < 1$ for $0 \leq x \leq 0.5$ and $0 \leq y \leq 3$, hence the expression in the denominator of $F_{yy}(x, y)$ is positive in this region. It follows that $F_{yy}(x, y) > 0$ for $0 \leq x \leq \rho_0$ and $0 \leq y \leq \sigma_0$. The result now follows from Proposition 5, Proposition 6 and and Theorem 8. Note that

$$\begin{aligned} a_n^* &\approx 1.07427068741 \cdot n^{-3/2} \cdot 0.424687310420272^{-n} \\ &= 1.07427068741/n^{-3/2} \cdot 2.35467360447^n. \end{aligned}$$

■

In a similar manner, we can attempt to compute the asymptotic limit of b_n, as defined in Proposition 4. By Proposition 5 we have $y = (z - 1 + x^2 z)/x^2 z$ and so define $G(x, z)$ to be equal to z plus the result of replacing y by $(z - 1 + x^2 z)/x^2 z$ in

$$y(x^2 + 1) - z(x^2 + x) - 1 - x^2 y^2. \tag{14}$$

Thus

$$\begin{aligned} G(x, z) &= -1 + \frac{-1 + z}{x^2 z^2} + \frac{1}{z} + z - xz - x^2 z \\ \frac{\partial}{\partial x} G(x, z) &= \frac{2 - 2z}{x^3 z^2} - z - 2xz \\ \frac{\partial}{\partial z} G(x, z) &= \frac{-2 + (1 + x^2)z + x^2(-1 + x + x^2)z^3}{x^2 z^3} \\ \frac{\partial^2}{\partial z^2} G(x, z) &= \frac{2(-3 + z + x^2 z)}{x^2 z^4} \end{aligned}$$

We find using Mathematica that a root (ρ_1, σ_1), of $G(x, z) = z$ and $\partial G/\partial z = 1$ is $\rho_1 = 0.424687$, $\sigma_1 = 1.426201$. In particular, using Mathematics, ρ_1 is not just approximately equal to ρ_1, but is exactly equal. However, we were unable to satisfy the conditions of Theorem 8, or indeed of other theorems in Meir and Moon (1989), and hence are currently unable to rigorously obtain an asymptotic limit for the number of b_n. (Note that $\rho_1 = 3.2131$, $\sigma_1 = 0.412773$ is another common solution of $G(x, z) = z$, $G_z(x, z) = 1$, for which we at present are unable to verify applicability of the method of Meir and Moon.)

By way of comparison, recall the following.

Theorem 10 (Stein and Waterman, 1979).

$$S(n) \sim \sqrt{\frac{15+7\sqrt{5}}{8\pi}} n^{-3/2} \left(\frac{3+\sqrt{5}}{2}\right)^n$$

Currently we have no information concerning the rate of convergence of a_n. By way of comparison, $\sqrt{\frac{15+7\sqrt{5}}{8\pi}} \approx 1.104366$ and $\frac{3+\sqrt{5}}{2} \approx 2.618034$, hence $S(n)$ is asymptotically approximately equal to $\frac{1.104366}{n^{3/}} \cdot 2.618034^n$.

See Table 1 for a sense of how much larger the collection of all secondary structures is than the collection of all saturated secondary structures.

5. 0- AND 1-SATURATED STRUCTURES

We now turn to the question of the number of 0-saturated and of 1-saturated secondary structures on n, for which we show that there are polynomially many 0- and 1-saturated structures. All results in this section assume that $\theta = 1$. We begin by the following observation.

A base or position i in sequence $1, \ldots, n$ is *unpaired* in secondary structure S, if there is no base pair $(x, y) \in S$, with $i \in \{x, y\}$. The number of unpaired bases is $n - 2|S|$. If S is a saturated structure, then any unpaired base is (topologically) isolared. Inductively counting 0- and 1-saturated structures comes down to distribution isolated points in certain regions of $1, \ldots, n$.

Remark 11. *Let S be a 0-saturated secondary structure on sequence $1, \ldots, n$. If n is odd, then there is exactly one unpaired base, while if n is even, then there are exactly two unpaired bases. Now, let S be a 1-saturated secondary structure on sequence $1, \ldots, n$. If n is odd, then there are 3 unpaired bases, while if n is even, then there are 4 unpaired bases.*

Theorem 12. *$LO_0(0) = LO_0(1) = LO_0(2) = LO_0(3) = 1$ and for $m \geq 2$,*

$$LO_0(2m) = LO_0(2m-2) + m \tag{15}$$

$$LO_0(2m+1) = 1 \tag{16}$$

Proof. Consider Equation (15), where $n = 2m$ is even. The collection $\mathcal{S}$ of all 0-saturated secondary structures on sequence $1, \ldots, n$ can be partitioned into $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3 \cup \mathcal{S}_4$, where $\mathcal{S}_1 = \{S \in \mathcal{S} : (1, n) \in S\}$, $\mathcal{S}_2 = \{S \in \mathcal{S} : (1, n-1) \in S\}$, $\mathcal{S}_3 = \{S \in \mathcal{S} : (1, n-2) \in S\}$, $\mathcal{S}_4 = \{S \in \mathcal{S} : \exists k[1 < k \wedge (k, n) \in S\}$. Clearly $|\mathcal{S}_1| = LO_0(2m-2)$. We claim that $|\mathcal{S}_2| = LO_0(2m-3) = 1$, since for any $S \in \mathcal{S}_2$, the restriction of S to $2, \ldots, 2m-2$ must be 0-saturated. The number of elements in $2, \ldots, 2m-2$ is odd and by Remark 11 has only one 0-saturated structure, thus $|\mathcal{S}_2| = 1$. Since positions $2m-1, 2m$ are unpaired in any structure of $\mathcal{S}_2$, and at least one additional base in $\{2, \ldots, 2m-3\}$ must be unpaired, it follows that $|\mathcal{S}_2| = 0$.

If $S \in \mathcal{S}_4$, then there is $1 < k \leq n-2$, such that the base pair $(k, n) \in S$. By Remark 11, if S is 0-saturated, then since $n = 2m$ is even, there must be 2 unpaired bases in $1, \ldots, n$, hence the restriction of S to $\{1, \ldots, k-1\}$ must be 0-saturated, as well as the restriction of S to $\{k+1, \ldots, n-1\}$. If k is odd, then $k-1$ is even and any 0-saturated secondary structure on $1, \ldots, k-1$ must have two unpaired bases. This situation is not possible, so we can only consider values $k = 2i$ which are even. In such cases, the region $k+1, \ldots, 2m-1$ contains $2m-k-1$ elements, an odd number, for which there is only one 0-saturated structure. It follows that

$$|\mathcal{S}_4| = \sum_{i=1}^{m-1} LO_0(2i-1) \cdot LO_0(2m-2i-1) = \sum_{i=1}^{m-1} 1 = m-1$$

Taking each of the preceding four cases into account, the number $|\mathcal{S}|$ of 0-saturated secondary structures on $1, \ldots, 2m$ is equal to

$$LO_0(2m) = LO_0(2m-2) + m \tag{17}$$

where $LO_0(0) = LO_0(1) = LO_0(2) = LO_0(3) = 1$. This establishes (15). Finally, Equation (16) follows immediately from Remark 11 since there is only one 0-saturated secondary structure on a sequence of odd length. ■

Corollary 13. *$LO_0(0) = LO_0(1) = LO_0(2) = LO_0(3) = 1$ and for $m \geq 2$,*

$$LO_0(n) = \begin{cases} n(n+2)/8 & \text{if } n \text{ is even} \\ 1 & \text{if } n \text{ is odd.} \end{cases}$$

Proof. Easy solution of recurrence relation (15) yields $LO_0(2m) = m(m+1)/2$. Substitution of $n = 2m$ in this case yields $n(n+2)/8$. ■

Theorem 14. *$LO_1(0) = LO_1(1) = LO_1(2) = 0$, $LO_1(3) = 1$ and for $m \geq 2$,*

$$\begin{aligned} LO_1(2m+1) &= LO_1(2m-1) + LO_0(2m-2) + LO_0(2m-3) \\ &\quad + \sum_{i=1}^{m-1} LO_0(2i-1) \cdot LO_0(2m-2i) \\ &\quad + \sum_{i=1}^{m-1} LO_0(2i) \cdot LO_0(2m-2i-1) \end{aligned} \tag{18}$$

$$\begin{aligned} LO_1(2m) &= LO_1(2m-2) + LO_1(2m-3) + LO_0(2m-4) \\ &\quad + \sum_{i=1}^{m-1} (LO_1(2i-1) + LO_1(2m-2i-1)) \\ &\quad + \sum_{i=1}^{m-2} LO_0(2i) \cdot LO_0(2m-2i-2) \end{aligned} \tag{19}$$

Proof. The collection $\mathcal{S}$ of all 1-saturated secondary structures on sequence $1, \ldots, n$ can be partitioned into $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3 \cup \mathcal{S}_4$, where $\mathcal{S}_1 = \{S \in \mathcal{S} : (1,n) \in S\}$, $\mathcal{S}_2 = \{S \in \mathcal{S} : (1,n-1) \in S\}$, $\mathcal{S}_3 = \{S \in \mathcal{S} : (1,n-2) \in S\}$, $\mathcal{S}_4 = \{S \in \mathcal{S} : \exists k[1 < k \wedge (k,n) \in S\}$. We consider $\mathcal{S}_1$ in Case 1, $\mathcal{S}_2$ in Case 2, $\mathcal{S}_3$ in Case 3, $\mathcal{S}_4$ in Case 4; additionally Case 4 has four subcases.

CASE 1. If $S \in \mathcal{S}_1$ is 1-saturated, then the restriction of S to $2, \ldots, n-1$ must be 1-saturated. The term $LO_1(2m-1)$ in Equation (18) [resp. $LO_1(2m-2)$ in Equation (19)] arises from all 1-saturated secondary structures containing the base pair $(1, 2m+1)$ [resp. $(1, 2m)$].

CASE 2. If $S \in \mathcal{S}_2$ is 1-saturated, then the restriction of S to $2, \ldots, n-1$ must be 1-saturated when n is even and 0-saturated [sic] when n is odd. This difference arising from parity of n follows from Remark 11. Indeed, if n is unpaired, and $(1, n-1)$ is base-paired, then $2, \ldots, n-2$ has an odd number of elements when n is even, and an even number of elements when n is odd. If $S \in \mathcal{S}_1$ is 1-saturated, then S has 4 unpaired bases when n is even, and 3 unpaired bases when n is odd.

The term $LO_0(2m-2)$ in Equation (18) arises from all 1-saturated secondary structures in which $n = 2m+1$ is unpaired, and which contain the base pair $(1, 2m)$. Since the region $2, \ldots, 2m-1$ enclosed within the base pair $(1, 2m)$ contains an even number of elements, the restriction $S \restriction \{2, \ldots, 2m-1\}$ must be 0-saturated. There are 2 unpaired bases in the region $2, \ldots, 2m-1$, which together with the unpaired base $n = 2m+1$ makes the required 3 unpaired bases.

The term $LO_1(2m-3)$ in Equation (19) arises from all 1-saturated secondary structures in which $n = 2m$ is unpaired and which contain the base pair $(1, 2m-1)$. Since the region $2, \ldots, 2m-2$ within the base pair

$(1, 2m-1)$ contains an odd number of elements, the restriction $S \restriction \{2, \ldots, 2m-2\}$ must be 1-saturated. There are 3 unpaired bases in the region $2, \ldots, 2m-2$, which together with the unpaired base $n = 2m$ makes the required 4 unpaired bases.

CASE 3. If $S \in \mathcal{S}_3$, then $(1, n-2) \in S$ and $n-1, n$ are unpaired. Suppose first that $n = 2m+1$ is odd. Then the inner region $2, \ldots, n-3$ is odd, and any 0-saturated secondary structure on $2, \ldots, 2m-2$ must have 1 unpaired bases. These, along with the unpaired bases $n-1, n$ makes a total of 3 unpaired bases. Thus arises a contribution of $LO_0(2m-3)$ in (18). Now suppose that $n = 2m$ is even. Similar reasoning shows that there is a contribution of $LO_0(2m-4)$ in Equation (19).

CASE 4. There exists k in $2, \ldots, n-2$, such that $(k, n) \in S$.

SUBCASE 1. Assume $n = 2m+1$ odd, $k = 2i$ even, and that S is 1-saturated on $1, \ldots, n$ and contains base pair (k, n).

$$LO_1(2m+1) = \sum_{i=1}^{m-1} LO_0(2i-1) \cdot LO_0(2m-2i)$$

By assumption, $k = 2i$ base-pairs with $n = 2m+1$, as $2i$ takes the values $2, 4, \ldots, 2m-2$. Note that there is an odd number $2i-1$ of elements in the region $1, \ldots, 2i-1$ to the left of the base pair $(2i, n)$—denote this region as the *left region*. Similarly, there is an even number $2m-2i$ of bases in the region $2i+1, \ldots, n-1$ occurring within the base pair $(2i, n)$ and to the right of $2i$—denote this region as the *right region*.

By Remark 11, any 1-saturated secondary structure S on $1, \ldots, n$ which includes base pair (k, n) must have 0-saturated restriction in the left region $1, \ldots, 2i-1$ as well as the right region $2i+1, \ldots, n-1$. This justifies the previous claim.

SUBCASE 2. Assume $n = 2m+1$ odd, $k = 2i+1 > 1$ odd, and that S is 1-saturated on $1, \ldots, n$ and contains base pair (k, n).

$$LO_1(2m+1) = \sum_{i=1}^{m-1} LO_0(2i) \cdot LO_0(2m-2i-1)$$

By assumption, $k = 2i+1$ base-pairs with $n = 2m+1$, as $2i$ takes the values $2, 4, \ldots, 2m-2$. For S to be 1-saturated, it follows by Remark 11 that the restriction of S to both the left region $1, \ldots, 2i$ and to the right region $2i+2, \ldots, n-1$ must be 0-saturated.

SUBCASE 3. Assume $n = 2m$ even, $k = 2i$ even. and that S is 1-saturated on $1, \ldots, n$ and contains base pair (k, n).

$$\begin{aligned} LO_1(2m) &= \sum_{i=1}^{m-1} LO_1(2i-1) \cdot LO_0(2m-2i-1) \\ &\quad + LO_0(2i-1) \cdot LO_1(2m-2i-1) \\ &= \sum_{i=1}^{m-1} LO_1(2i-1) + LO_1(2m-2i-1) \end{aligned}$$

By assumption, $k = 2i$ base-pairs with $n = 2m$, as $2i$ takes the values $2, 4, \ldots, 2m-2$. For S to be 1-saturated, it follows by Remark 11 that the restriction of S to the left region $1, \ldots, 2i$ must be 1-saturated and to the right region must be 0-saturated, or alternatively the restriction of S to the left region $1, \ldots, 2i$ must be 0-saturated and to the right region must be 1-saturated. This will ensure that there are 4 unpaired bases, as required in a 1-saturated secondary structure on a sequence of even length n. Finally, since $2m-2i-1$ and $2i-1$ are odd, and the number of 0-saturated secondary structures on an odd sequence is 1, it follows that $LO_0(2m-2i-1) = 1 = LO_0(2i-1)$.

SUBCASE 4. Assume $n = 2m$ even, $k = 2i + 1 > 1$ odd, and that S is 1-saturated on $1, \ldots, n$ and contains base pair (k, n).

$$LO_1(2m) = \sum_{i=1}^{m-2} LO_0(2i) \cdot LO_0(2m - 2i - 2)$$

By assumption, $k = 2i + 1$ base-pairs with $n = 2m$, as $2i$ takes the values $2, 4, \ldots, 2m - 4$. For S to be 1-saturated, it follows by Remark 11 that the restriction of S to both the left region $1, \ldots, 2i$ and to the right region $2i + 2, \ldots, n - 1$ must be 0-saturated. This establishes the theorem. ■

Corollary 15. *$LO_1(0) = LO_1(1) = LO_1(2) = 0$, $LO_1(3) = 1$ and for $m \geq 2$,*

$$\begin{aligned} LO_1(2m + 1) &= LO_1(2m - 1) + m(m - 1)/2 + 1 \\ &\quad + \sum_{i=1}^{m-1} ((m - i)(m - i + 1)/2 + i(i + 1)/2) \\ LO_1(2m) &= m(m - 1)/2 + 1 + (m - 2)(m - 1)/2 \\ &\quad + \sum_{i=1}^{m-1} (LO_1(2i - 1) + LO_1(2m - 2i - 1)) \\ &\quad + \sum_{i=1}^{m-2} \frac{i(i + 1)(m - i - 1)(m - 1)}{4}. \end{aligned}$$

Proof. Immediate from the previous theorem by replacing terms of the form $LO_0(k)$ by their corresponding value from Theorem 12. ■

Although it would be indeed tedious to solve the recurrence relation given in Corollary 15, it is nevertheless clear that $LO_1(n)$ is bounded by a polynomial in n.

6. CONCLUSION

In this paper, we have studied combinatorial problems associated with the new concept of *saturated* secondary structure. Saturated structures form natural kinetic traps in the folding process, and hence the combinatorial results in this paper shed some light on the the distribution of (local) energy minima in the energy surface for secondary structures of length n RNA sequences. We have established an exponential asymptotic limit for the number of saturated secondary structures, and shown that if n is even there are exactly $n(n + 2)/8$ many 0-saturated structures, while the number of 1-saturated structures is polynomial in n. Additionally, we have used the recurrence relation for the number of saturated secondary structures as the basis of a dynamic programming algorithm to compute the number $LO(n, k)$ of saturated secondary structures having k base pairs. The number $LO_k(n)$ of k-saturated secondary structures is then $LO(n, \lfloor \frac{n-\theta}{2} \rfloor - k)$.

Application of the (erroneous) theorem of Bender allows one to evaluate the asymptotic limit of $b_n = \sum_k MO(n, k)$, where b_n is the number of saturated stuctures on length n sequence for threshold $\theta = 1$ (see Proposition 4). If Bender's theorem were correct, then asymptotically b_n is equal to an exponential expression with the *same* exponential base as that of a_n. However, at the present time, it does not seem that the (correct) theorem of Meir and Moon (1989) can be applied. It would be of interest to rigorously derive the asymptotic limit of b_n.

An unexplored problem which may be tractable is to apply generating function theory to compute formulas, or asymptotic limits, for the number $LO(n, k)$ of saturated structures on a length n sequence with k base pairs. By using the correspondence between secondary structures and *linear trees*, given by

Schmitt and Waterman (1994) [see also Chapter 13 of Waterman (1995)], it is straightforward to define a one-one correspondence between $LO(n, k)$ and a class of linear trees with the property that are at most 2 children of each node which are leaves, and in the case that there are 2 leaf children, they are adjacent. It may be possible to develop a closed formula or asymptotic limit for this class of trees.

ACKNOWLEDGMENTS

I would like to thank R. Bundschuh and D. Mathews for suggesting to compute the number of saturated secondary structures having k base pairs, rather than directly computing the number of k-saturated structures, as done in Clote (2005a), to M. Zuker for pointing out the relevance of Zuker (1986), and especially to the referee for pointing out the error in Bender's Theorem (Bender, 1974) and (partial) rectification in Meir and Moon (1989). Research partially supported by NSF DBI-0543506.

REFERENCES

Banerjee, A.R., Jaeger, J.A., and Turner, D.H. 1993. Thermal unfolding of a group I ribozyme: the low-temperature transition is primarily disruption of tertiary structure. *Biochemistry* 32, 153–163.

Barrick, J.E., Corbino, K.A., Winkler, W.C., et al. 2004. New RNAmotifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci. USA* 101, 6421–6426.

Bender, E.A. 1974. Asymptotic methods in enumeration. *SIAM Rev.* 16, 485–515.

Böck, A., Forschhammer, K., Heider, J., et al. 1991. Selenoprotein synthesis: an expansion of the genetic code. *Trends Biochem. Sci.* 16, 463–467.

Canfield, E.R. 1984. Remarks on an asymptotic method in combinatorics. *J. Combin. Theory Ser. A* 37, 348–352.

Clote, P. 2005a. An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov-Jacobson energy model. *J. Comput. Biol.* 12, 83–101.

Clote, P. 2005b. RNALOSS: A web server for RNA locally optimal secondary structures. *Nucleic Acids Res.* 33, W600–W604.

Clote, P., Ferré, F., Kranakis, E., et al. 2005a. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA* 11, 578–591.

Clote, P., Kranakis, E., Krizanc, D., et al. 2005b. Asymptotic expected number of base pairs in optimal secondary structure for random RNA using the Nussinov-Jacobson energy model. *Discrete Appl. Math.* (in press).

Clote, P., Waldispuhl, J., Behzadi, B., et al. 2005. Energy landscape of k-point mutants of an RNA molecule. *Bioinformatics* 21, 4140–4147.

Cupal, J., Hofacker, I., and Stadler, P. 1996. Dynamic programming algorithm for the density of states of RNA secondary structures. *Comput. Sci. Biol. 96* 184–186.

Ding, Y., and Lawrence, C.E. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* 31, 7280–7301.

Evers, D.J., and Giegerich, R. 2001. Reducing the conformation space in RNA structure prediction. *German Conf. Bioinform.* 118–124.

Havgaard, J.H., Lyngsø, R., Stormo, G., et al. 2005. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics* 21, 1815–1824.

Heider, J., Baron, C., and Böck, A. 1992. Coding from a distance dissection of the mrna elements required for the incorporation of selenocysteine into protein. *EMBO J.* 11, 3759–3766.

Hofacker, I.L., Fontana, W., Stadler, P.F., et al. 1994. Fast folding and comparison of RNA secondary structures. *Monatsch. Chem.* 125, 167–188.

Hofacker, I.L., Schuster, P., and Stadler, P.F. 1998. Combinatorics of RNA secondary structures. *Discrete Appl. Math.* 88, 207–237.

Lim, L.P., Glasner, M.E., Yekta, S., et al. 2003. Vertebrate microRNA genes. *Science* 299, 1540.

Mathews, D.H., Sabina, J., Zuker, M., et al. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288, 911–940.

Mathews, D.H., and Turner, D.H. 2002. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.* 317, 191–203.

Mathews, D.H., Turner, D.H., and Zuker, M. 2000. Secondary structure prediction, 11.2.1–11.2.10, *In* S. Beaucage, D.E. Bergstrom, G.D. Glick, et al., eds., *Current Protocols in Nucleic Acid Chemistry*. John Wiley & Sons, New York.

Meir, A., and Moon, J.W. 1989. On an asymptotic method in enumeration. *J. Combin. Theory Ser. A* 51, 77–89.

Moon, S., Byun, Y., Kim, H.-J., et al. 2004. Predicting genes expressed via −1 and +1 frameshifts. *Nucleic Acids Res.* 32, 4884–4892.

Nebel, M.E. 2002. Combinatorial properties of RNA secondary structure. *J. Comput. Biol.* 9, 541–573.

Nussinov, R., and Jacobson, A.B. 1980. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proc. Natl. Acad. Sci. USA* 77, 6309–6313.

Schattner, P., Decatur, W.A., Davis, Jr., C.A., et al. 2004. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.* 32, 4281–4296.

Schmitt, W.R., and Waterman, M.S. 1994. Linear trees and RNAsecondary structure. *Discrete Appl. Math.* 51, 317–323.

Stein, P.R., and Waterman, M.S. 1979. On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Math.* 26, 261–272.

Tuschl, T. 2003. Functional genomics: RNA sets the standard. *Nature* 421, 220–221.

Voss, B., Meyer, C., and Giegerich, R. 2004. Evaluating the predictability of conformational switching in RNA. *Bioinformatics* 20, 1573–1582.

Waterman, M.S. 1995. *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman & Hall, New York.

Wuchty, S., Fontana, W., Hofacker, I.L., et al. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49, 145–164.

Xia, T., Jr., SantaLucia, J., Burkard, M.E., et al. 1999. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* 37, 14719–14735.

Zuker, M. 1986. RNA folding prediction: the continued need for interaction between biologists and mathematicians. *Lect. Math. Life Sci.* 17, 87–124.

Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415.

Zuker, M., and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9, 133–148.

Address reprint requests to:
Dr. Peter G. Clote
Department of Biology
Boston College
Higgins Hall 355
140 Commonwealth Ave.
Chestnut Hill, MA 02467

E-mail: clote@bc.edu