Classifying RNA Pseudoknotted Structures

Anne Condon*, Beth Davy
Baharak Rastegari, Shelly Zhao
The Department of Computer Science
University of British Columbia
Vancouver, BC, V6T 1Z4, Canada

Finbarr Tarrant
The Kane Building
Department of Computer Science
University College Cork
College Road, Ireland

May 26, 2004

Abstract

Computational prediction of the minimum free energy (mfe) secondary structure of an RNA molecule from its base sequence is valuable in understanding the structure and function of the molecule. Since the general problem of predicting pseudoknotted secondary structures is NP-hard, several algorithms have been proposed that find the mfe secondary structure from a restricted class of secondary structures. In this work, we order the algorithms by generality of the structure classes that they handle. We provide simple characterizations of the classes of structures handled by four algorithms, as well as linear time methods to test whether a given secondary structure is in three of these classes. We report on the percentage of biological structures from the PseudoBase and Gutell databases that are handled by these two algorithms.

1 Introduction

RNA molecules - sequences of nucleic acid bases - play diverse roles in the cell: as carriers of information, catalysts in cellular processes, and mediators in determining the expression level of genes [4]. The structure of RNA molecules is often key to their function, and so tools for computational prediction of RNA *secondary structure* - the set of base pairings present in its folded state - are widely used.

While comparative approaches are most reliable for secondary structure prediction [6], these approaches require that several homologous (i.e. evolutionarily and functionally related) sequences are available. When just a single molecule is available, computational prediction of its secondary structure from its base sequence (at fixed temperature, ionic concentration, and pressure) is based on the premise that out of the exponentially many possibilities, an RNA molecule is most likely to fold into the *minimum free energy (mfe)* structure. The free energy of a given structure for a sequence is estimated by summing thermodynamic and entropic free energy terms associated with the component loops of the secondary structure. Some of these terms have been obtained experimentally, and others are estimated based on existing databases of naturally occurring structures.

Unfortunately, finding the mfe secondary structure for a given RNA sequence is NP-hard [9]. Several polynomial time algorithms have been proposed for predicting the mfe secondary structure from restricted classes of secondary structures. The most well known such class is that of *pseudoknot free* secondary structures (see Figure 1). Many biological RNA structures are pseudoknot free and extensive experimental work has been done to determine parameters for the underlying thermodynamic model. Pseudoknot free secondary structures can be described as generalized strings of balanced parentheses. Dynamic programming algorithms for finding the mfe pseudoknot free secondary structure from the base sequence run in $\Theta(n^3)$ time, and are the basis for the well known mfold and Vienna secondary structure prediction packages [7, 10]. Moreover, there are linear time methods to test that a

*Communicating author: condon@cs.ubc.ca

secondary structure (represented as an ordered list of base pairs or stems) is pseudoknot free, and to calculate the free energy of a given secondary structure for a given sequence. The latter algorithm is quite useful in practice, and software to do it is available as part of the mfold package.

Pseudoknots occur in many natural structures [10, 11]. Recently algorithms have been designed to predict the mfe secondary structure for limited classes of pseudoknotted structures [1, 5, 8, 11, 13]. The running times of these algorithms range from $\Theta(n^4)$ to $\Theta(n^6)$, and each handles a different class of structures. However, the trade-off between the running time of the algorithms and the generality of the classes of structures they handle has been poorly understood. Rivas and Eddy [11] state that "we still lack a systematic *a priori* characterization of the class of configurations that this algorithm can solve". Lyngsø and Pedersen [9] do order several classes in terms of their generality, but rely on examples rather than formal characterizations to explain which structures cannot be handled by the classes. Moreover, other than for the algorithm of Dirks and Pierce, there has been little data to indicate whether the class of structures handled by these algorithms includes most known pseudoknotted biological structures.

To address these problems, in Section 3 we provide simple characterizations of the classes of structures handled by four algorithms: the Rivas and Eddy (R&E) class (which is the most general class known to us), the class of Akutsu and Uemura et al (A&U) [13, 1], the Dirks and Pierce (D&P) [5] class, and a simple class, modeled after that of Lyngsø and Pedersen (L&P). Using our characterizations, we provide linear time algorithms to test if an input structure is in the R&E, D&P, and L&P classes, and present results for several RNA secondary structure families. As an example of our results, in tests of 486 secondary structures with isolated base pairs removed, we found that all but three Group II Intron structures are in the R&E class.

We provide background on RNA secondary structure in Section 2, and describe our results in more detail in subsequent sections.

2 Secondary Structure Background

An RNA molecule is a chain of four types of bases, denoted by A, C, G, and U (for Adenine, Cytosine, Guanine, and Uracil). The chain has distinct $5'$ and $3'$ ends. We index the bases consecutively from the $5'$ end, starting at 1. A folded molecule is held together by hydrogen bonds between pairs of bases, with each base typically participating in at most one pair. A set R of such base pairs is called a *secondary structure* of the molecule. If bases indexed i and j are paired where $i < j$ then we write $i \cdot j \in R$. Throughout, we use n to denote the length of a molecule.

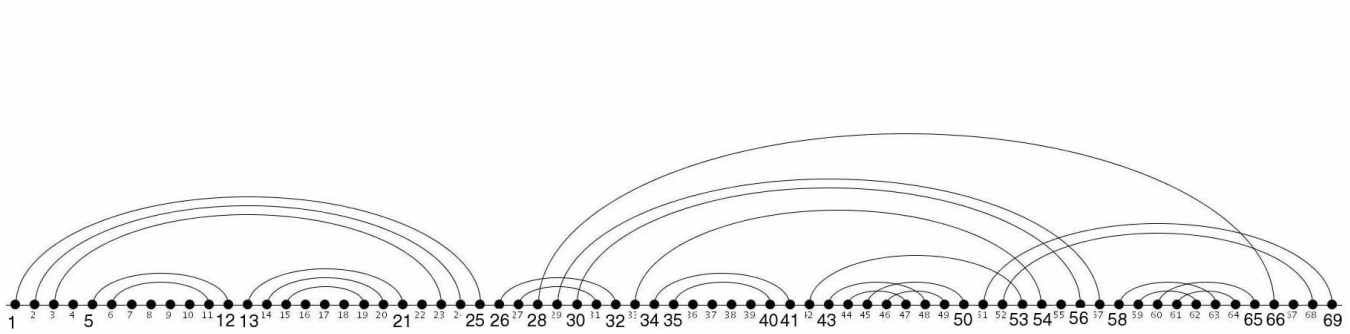

Figure 1: Arc representation of an RNA secondary structure. In the left substructure (up to index 25), arcs are hierarchically nested, thus this is a *pseudoknot free* substructure. Arcs cross in the substructure on the right, thus it is *pseudoknotted*.

Figure 1 gives an arc representation of a secondary structure. The chain of indexed bases is represented as an indexed line of dots, and arcs connect paired bases. Later we will use a linked list representation of secondary structures, where list elements are the base indices, ordered starting from the $5'$ end, with additional pointers corresponding to arcs of the structure's arc representation.

We will also use an alternative *pattern* representation of secondary structures. In a pattern, information about the base indices is lost but the pattern of nesting or overlaps among base pairs is preserved. (We note that the definition of pattern could be extended so that unpaired bases are represented using a special symbol.) To define patterns precisely, we introduce some notation. We use ϵ to denote the empty string. Let $\mathcal{N}_n$ denote the natural numbers between 1 and n (inclusive). For any string s over alphabet Σ, $s \downarrow \sigma$ denotes the string s with all occurrences of σ removed. Also, $|s|$ denotes the number of symbols in s.

Patterns: A string p (of even length) over some alphabet Σ is a *secondary structure pattern*, or simply a pattern, if every symbol of Σ occurs either exactly twice, or not at all, in p. We say that secondary structure R for a strand of length n *corresponds to* pattern p if there exists a mapping $m : \mathcal{N}_n \rightarrow \Sigma \cup \{\epsilon\}$ with the following properties: (i) if $i \cdot j \in R$ then $m(i) \in \Sigma$ and $m(i) = m(j)$, (ii) if $i \cdot j$ and $j \cdot i \notin R$ for all $j \in \mathcal{N}_n$, then $m(i) = \epsilon$, and (iii) $p = m(1)m(2)\ldots m(n)$. For example, the substructure of Figure 1 from index 5 to index 21 corresponds to pattern *ABBACDEEDC*, and the substructure from index 43 to index 50 corresponds to pattern *ABCDBADC*. It is possible to convert between commonly used representations of secondary structures, such as list of base pairs, and pattern representations, in time linear in n.

Let $R \subseteq \mathcal{N}_n^2$ correspond to pattern p over alphabet σ and let mapping m witness this correspondence. Let $m_i^{-1} : \Sigma \rightarrow \mathcal{N}_n$ be defined by $m_i^{-1}(\sigma) = min\{i \in \mathcal{N}_n \mid m(i) = \sigma\} \cup \{n+1\}$ and let $m_j^{-1}(\sigma) = max\{j \in \mathcal{N}_n \mid m(j) = \sigma\} \cup \{0\}$. Note that $m_i^{-1}(\sigma) \in \mathcal{N}_n$ if and only if $m_j^{-1}(\sigma) \in \mathcal{N}_n$, in which case $m_i^{-1}(\sigma) < m_j^{-1}(\sigma)$. Also, $(m_i^{-1}(\sigma), m_j^{-1}(\sigma)) = (n+1, 0)$ if and only if σ does not occur in pattern p. If $(m_i^{-1}(\sigma), m_j^{-1}(\sigma)) \neq (n+1, 0)$ then $m_i^{-1}(\sigma) \cdot m_j^{-1}(\sigma) \in R$ and we say that base pair $m_i^{-1}(\sigma) \cdot m_j^{-1}(\sigma)$ *corresponds to* σ.

Figure 1 illustrates a pseudoknot free substructure (up to index 25) and a pseudoknotted substructure. Formally, a secondary structure R is *pseudoknot free* if for all pairs $i \cdot j$ and $i' \cdot j'$ in R, it is not the case that $i < i' < j < j'$. Applying this definition directly to determine whether a secondary structure is pseudoknot free would require $\Theta(n^2)$ time. However, linear time tests for pseudoknot freeness are well known. For completeness, and because our characterizations of pseudoknotted secondary structure classes presented later are similar, we outline such a test here. We use the notion of pseudoknot free patterns.

Pseudoknot free patterns: We say that symbol σ is *self-adjacent* in string p if $\sigma\sigma$ is a substring of p. We write $p \underset{\text{PKF}}{\longrightarrow} p'$ if $p' = p \downarrow \sigma$, for some self-adjacent symbol σ of p, and $p \underset{\text{PKF}}{\overset{*}{\longrightarrow}} p'$ if $p = p'$ or $\exists$ patterns $p_1, \ldots, p_k$ for some k such that $p \underset{\text{PKF}}{\longrightarrow} p_1 \underset{\text{PKF}}{\longrightarrow} \ldots p_k \underset{\text{PKF}}{\longrightarrow} p'$. We say that p is a *pseudoknot free pattern* if and only if $p \underset{\text{PKF}}{\overset{*}{\longrightarrow}} \epsilon$.

It is straightforward to show that secondary structure R is pseudoknot free if and only if the pattern corresponding to R is pseudoknot free. Roughly, a linear time test that a pattern is pseudoknot free simply scans the pattern from left to right, removing self-adjacent pairs when possible. The pattern is empty after the last symbol is scanned, if and only if it is pseudoknot free.

3 Structure Classes

Dynamic programming algorithms for prediction of minimum free energy (mfe) pseudoknotted secondary structures were proposed (in chronological order) by Uemura et al. [13], Akutsu [1], Rivas and Eddy (R&E) [11], Lyngsø and Pedersen (L&P) [8], and Dirks and Pierce (D&P) [5]. (We note that the D&P algorithm is more general than the others, in that it can calculate the partition function as well as the mfe secondary structure.) Each algorithm finds the mfe structure from a limited class of secondary structures. Akutsu derived his class by simplifying that of Uemura et al.. We call the class, as described by Akutsu, the A&U class. We refer to the other classes by the initials of the authors of the algorithm. We note that our version of the L&P class is simpler than that defined in their paper – see section 3.4. We use PKF to denote the class of pseudoknot free secondary structures. In summary, the classes can be properly ordered as follows:

$$\text{PKF} \subset \text{L\&P} \subset \text{D\&P} \subset \text{A\&U} \subset \text{R\&E}.$$

The fact that the A&U class is properly contained in the R&E class was already noted by Lyngsø and Pedersen [9], by comparing the structure of the recurrences of both algorithms. Instead, to derive our containments, in the next subsections we develop formal characterizations of each of the D&P and R&E classes. Precise descriptions of the A&U and L&P classes are already in the work of Akutsu [1] and Lyngsø and Pedersen (L&P) [8], respectively, but we also characterize these classes in a manner similar to our characterizations of the D&P and R&E classes, so that the containments above can be derived.

3.1 Rivas and Eddy (R&E) Structures

The R&E structure class is defined implicitly by the recurrences of the algorithm of Rivas and Eddy [11]. We first abstract the form of the recurrences to define the class of secondary structures that their class can handle; this is done formally in the definition of R&E-algorithm patterns below. Following this, we give our characterization of the structure class, and argue that our characterization is equivalent to the class of R&E-algorithm patterns.

We need to account for "gapped regions" that play an important role in the R&E recurrences, and so we first introduce generalized patterns. A *generalized pattern* over Σ is a string p over $\Sigma \cup \{G\}$ where $G \notin \Sigma$, such that there is at most one G in p and $p \downarrow G$ is a pattern. A generalized pattern p over alphabet Σ is a *generalized R&E-algorithm pattern* if at least one of the following conditions hold.

1. $p = \epsilon$ or $p = \sigma g \sigma$, for some $\sigma \in \Sigma$ and $g \in \{G, \epsilon\}$.
2. $p = Gp'$ or $p = p'G$ where p' is a generalized R&E-algorithm pattern.
3. $p = p_1 p_2 g p_3 p_4$ where $|p| \geq 4$, $g \in [G, \epsilon]$, $p_1 p_2 p_3 p_4 \in \Sigma^*$, and either
   (a) $p_1 G p_3$ and $p_2 G p_4$ are generalized R&E-algorithm patterns, $|p_1 G p_3| < |p|$, and $|p_2 G p_4| < |p|$, or
   (b) $p_1 G p_4$ and $p_2 g p_3$ are generalized R&E-algorithm patterns, $|p_1 G p_4| < |p|$, and $|p_2 g p_3| < |p|$.
4. $p = p_1 G p_2 p_3 p_4$ or $p = p_1 p_2 p_3 G p_4$ where $|p| \geq 4$, $p_1 G p_3$ and $p_2 G p_4$ are generalized R&E-algorithm patterns, $|p_1 G p_3| < |p|$, and $|p_2 G p_4| < |p|$.

A generalized R&E-algorithm pattern p is a R&E-algorithm pattern if p does not contain G. We say that secondary structure R is an *R&E secondary structure* it corresponds to a R&E-algorithm pattern.

We next define the class of R&E patterns, which is a simple generalization of the pseudoknot free patterns. In Theorem 1 we will show that the set of R&E-algorithm patterns is equal to the set of R&E patterns. Let p be a string over alphabet Σ. Symbol σ is *directly adjacent* to symbol τ in p if and only if either $\sigma\tau\sigma$ is a substring of p or there are two disjoint substrings x, y of p, both of length 2, such that τ and σ are both in x and τ and σ are both in y. (The direct adjacency relation is not necessarily symmetric.) If σ is directly adjacent to some symbol in pattern p, we say σ is directly adjacent in p.

Let p be a pattern. We say that $p \xrightarrow[\text{R\&E}]{} p'$ if $p' = p \downarrow \sigma$ for some σ that is either self-adjacent or directly adjacent in p. Also, $p \xrightarrow[\text{R\&E}]{*} p'$ if $p = p'$ or $\exists$ patterns $p_1, \ldots, p_k$ for some k such that $p \xrightarrow[\text{R\&E}]{} p_1 \xrightarrow[\text{R\&E}]{} \ldots p_k \xrightarrow[\text{R\&E}]{} p'$. Pattern p is *R&E* if $p \xrightarrow[\text{R\&E}]{*} \epsilon$.

A generalized pattern p over $\Sigma \cup \{G\}$ is a generalized R&E pattern if either p is R&E, $p \xrightarrow[\text{R\&E}]{*} G$ or $p \xrightarrow[\text{R\&E}]{*} \sigma G \sigma$ for some $\sigma \in \Sigma$.

Theorem 1 *The R&E secondary structures are exactly the structures corresponding to R&E patterns.*

Proof We show one direction, that if p is a generalized R&E-algorithm pattern, then p is a generalized R&E pattern. The proof is by induction on $|p|$. The base case is straightforward. For the inductive step, let $|p| > 3$, and suppose that all generalized R&E-algorithm patterns of length $< |p|$ are R&E.

In the most interesting case, $p = p_1p_2gp_3p_4$ where case 3(b) of the definition of R&E-algorithm patterns holds for p_1, p_2, p_3 and p_4. Since $|p_1Gp_3| < |p|$ and $|p_2gp_4| < |p|$, it follows from the induction hypothesis that p_1Gp_3 and p_2gp_4 are generalized R&E patterns. The result therefore follows from the following claim.

Claim. If p_1Gp_4 and p_2gp_3 are generalized R&E patterns, where $g \in [G, \epsilon]$ then $p_1p_2gp_3p_4$ is a generalized R&E pattern.

The Claim can be also be proved in a straightforward way, by induction on $|p_1p_2p_3p_4|$. One base case is when for some σ and τ, $\sigma = p_1 = p_4$ and $\tau = p_2 = p_3$, so that $p = \sigma\tau g\tau\sigma \underset{\text{R\&E}}{\longrightarrow} \sigma g\sigma$, which is a generalized R&E pattern by the base case of the definition of R&E patterns.

For the inductive step, let $|p_1p_2p_3p_4| > 4$ and suppose that the Claim holds for all p'_1, p'_2, p'_3, p'_4 with $|p'_1p'_2p'_3p'_4| < |p_1p_2p_3p_4|$. We consider the case where $|p_1p_4| > 2$; the case where $|p_2p_3| > 2$ is similar. Suppose that $p_1Gp_4 \underset{\text{R\&E}}{\longrightarrow} p'_1Gp'_4$, where $p'_1Gp'_4$ is a generalized R&E pattern. Then, by the induction hypothesis, since $|p'_1Gp'_4| < |p_1Gp_4|$, $p'_1p_2gp_3p'_4$ is a generalized R&E pattern. Also, $p_1p_2gp_3p_4 \underset{\text{R\&E}}{\longrightarrow} p'_1p_2gp_3p'_4$, since if σ is self-adjacent or directly adjacent in p_1gp_4 then σ must also be self-adjacent or directly adjacent in $p_1p_2gp_3p_4$. Therefore,

$$p_1p_2gp_3p_4 \underset{\text{R\&E}}{\longrightarrow} p'_1p_2gp_3p'_4 \overset{*}{\underset{\text{R\&E}}{\longrightarrow}} \epsilon,$$

from which the Claim follows. □

3.2 Akutsu and Uemura et al. (A&U) Structures

We first define the A&U structure class, following the definition of Akutsu [1]. A secondary structure R is called a *simple pseudoknot* if there exist $j'_0, j_0 \in \mathcal{N}_n$ with $j'_0 < j_0$ for which the following conditions are satisfied.

1. Each $i \cdot j \in R$ satisfies either $i < j'_0 \leq j < j_0$ or $j'_0 \leq i < j_0 \leq j$.
2. If $i \cdot j$ and $i' \cdot j'$ are in R with either $i < i' < j'_0$ or $j'_0 \leq i < i'$, then $j' < j$.

R is an A&U secondary structure if either R is a simple pseudoknot or a pseudoknot free secondary structure, or for some $i_0, k_0, 1 \leq i_0 < k_0 \leq n, R = R' \cup R''$ where $R' \subseteq (\mathcal{N}_n - [i_0, k_0])^2$, $R'' \subseteq [i_0, k_0]^2$, R' is an A&U structure and R'' is a nonempty simple pseudoknot or pseudoknot free structure.

Our characterization of A&U structures is less elegant than that obtained for the R&E class in the previous section, but is nevertheless useful in order to compare the classes.

Let p be a string. We say that σ is *directly nested in p, with respect to τ*, if disjoint substrings $\tau\sigma$ and $\sigma\tau$ appear in p in that order. If σ is directly nested in p, with respect to some τ, we say that σ is *directly nested* in p. We say that σ is *A&U-adjacent* in p, with respect to τ if σ is directly nested in p or if substring τ is followed (not necessarily contiguously) by substring $\sigma\tau\sigma$ in p.

Given τ, we say $p \underset{\text{A\&U},\tau}{\longrightarrow} p'$ if $p' = p \downarrow \sigma$ for some σ that is A&U-adjacent in p with respect to τ. We say $p \overset{*}{\underset{\text{A\&U},\tau}{\longrightarrow}} p'$ if for some $k > 0$, $\exists$ patterns $p_1, \ldots, p_k$ such that $p \underset{\text{A\&U},\tau}{\longrightarrow} p_1 \underset{\text{A\&U},\tau}{\longrightarrow} \ldots p_k \underset{\text{A\&U},\tau}{\longrightarrow} p'$, and moreover, τ is self-adjacent in p'.

We say that $p \underset{\text{A\&U}}{\longrightarrow} p'$ if $p' = p \downarrow \sigma$ for some σ that is self-adjacent or directly nested in p or if $p \overset{*}{\underset{\text{A\&U},\tau}{\longrightarrow}} p'$ for some τ. We say that $p \overset{*}{\underset{\text{A\&U}}{\longrightarrow}} p'$ if $p = p'$ or $\exists$ patterns $p_1, \ldots, p_k$ such that $p \underset{\text{A\&U}}{\longrightarrow} p_1 \underset{\text{A\&U}}{\longrightarrow} \ldots p_k \underset{\text{A\&U}}{\longrightarrow} p'$. Pattern p is *A&U* if $p \overset{*}{\underset{\text{A\&U}}{\longrightarrow}} \epsilon$.

Theorem 2 *The A&U secondary structures are exactly the structures corresponding to A&U patterns.*

Proof We include the direction that every A&U secondary structure corresponds to an A&U pattern. Let R be an A&U secondary structure. Firstly, if R is pseudoknot free, then its corresponding pattern is a pseudoknot free pattern, and therefore an A&U pattern, since the $\underset{\text{A\&U}}{\longrightarrow}$ relation generalizes the $\underset{\text{PKF}}{\longrightarrow}$ relation.

Secondly, suppose that R is a simple pseudoknot, but is not pseudoknot free. Let p be the pattern corresponding to R. Let p' be obtained from p by repeatedly removing any self-adjacent symbols and let R' be the substructure of R obtained by removing the base pairs corresponding to these self-adjacent symbols. Let τ be the first symbol in p'. Let $i \cdot j$ be the base pair of R' corresponding to τ. Then, since R' is not pseudoknot free, by condition 1 of the definition of a simple pseudoknot it must be that $i \leq j'_0 \leq j < j_0$. Moreover, condition 1, together with the fact that p' contains no self-adjacent base pairs, implies that for all other base pairs $i' \cdot j'$, either (i) $i < i' < j' < j < j_0$ or (ii) $j'_0 \leq i' < j < j_0 \leq j'$.

Let σ be the symbol just to the left of the second occurrence of τ in p', and let $i' \cdot j'$ be the base pair of R' corresponding to σ. Then, either $i' \cdot j'$ satisfies case (i) of the last paragraph, in which case σ is directly nested in p' with respect to τ, or $i' \cdot j'$ satisfies case (ii) of the last paragraph, in which case $\sigma\tau\sigma$ is a substring of p'. In either case, σ is A&U-adjacent to τ in p', and so $p' \xrightarrow[\text{A\&U},\tau]{} p' \downarrow \sigma$. Let $p'' = p' \downarrow \sigma$ and let σ' be the symbol just to the left of the second occurrence of τ in p''. If $\sigma' \neq \tau$, then by the same reasoning as for σ, it must be that $p'' \xrightarrow[\text{A\&U},\tau]{} p'' \downarrow \sigma'$. Continuing in this way, we conclude that $p' \xrightarrow[\text{A\&U},\tau]{*} \tau\tau$.

Therefore, $p \xrightarrow[\text{A\&U}]{*} \epsilon$ by a series of steps in which first self-adjacent symbols are removed, then symbols that are A&U adjacent to the first symbol τ of p are removed, and finally $\tau\tau \xrightarrow[\text{A\&U}]{} \epsilon$. Therefore, p is an A&U pattern.

Finally, suppose that R is an A&U secondary structure that is neither a simple pseudoknot nor pseudoknot free. Then, for some $i_0, k_0, 1 \leq i_0 < k_0 \leq n, R = R' \cup R''$ where $R' \subseteq (\mathcal{N}_n - [i_0, k_0])^2$, $R'' \subseteq [i_0, k_0]^2$, R' is an A&U structure and R'' is a nonempty simple pseudoknot or pseudoknot free structure. Let p, p', and p'' be the patterns corresponding to R, R', and R'' respectively. A proof by induction can be used to show that if p' and p'' are A&U patterns, then so is p; in fact $p \xrightarrow[\text{A\&U}]{*} p' \xrightarrow[\text{A\&U}]{*} \epsilon$. Therefore R corresponds to an A&U pattern. □

3.3 Dirks and Pierce (D&P) Structures

As with the R&E structure class, the D&P structure class is also defined implicitly, in this case by the recurrences of the algorithm of Dirks and Pierce [5]. We abstract the form of the recurrences to define the class of secondary structures that their class can handle; this is done formally in the definition of D&P-algorithm patterns below, which is in fact a restriction of the recurrences of Rivas and Eddy. (Our abstraction does not capture certain features of their algorithm, that are important in the context of their work, but not important in terms of defining the class of structures handled.) Following this, we give our characterization of the D&P structure class.

A generalized pattern p over alphabet Σ is a *generalized D&P-algorithm pattern* if at least one of the following conditions hold.

1. $p = \epsilon$ or $\sigma g \sigma$, for some $\sigma \in \Sigma$ and $g \in \{G, \epsilon\}$.
2. $p = Gp'$ or $p'G$ where p' is a generalized D&P-algorithm pattern.
3. $p = \sigma\tau p_1 \tau\sigma$ for some $\sigma, \tau \in \Sigma$, where $\tau p_1 \tau$ is a generalized D&P-algorithm pattern.
4. $p = p_1p_2p_3p_4$ where $|p| \geq 4$, $p \in \Sigma^*$, p_1Gp_3 and p_2Gp_4 are generalized D&P-algorithm patterns, $|p_1Gp_3| < |p|$, and $|p_2Gp_4| < |p|$.
5. $p = p_1p_2Gp_3p_4$ where $|p| \geq 4$ and either
   (a) $p_2Gp_3p_4$ and p_1 are generalized D&P-algorithm patterns, $|p_2Gp_3p_4| < |p|$ and $|p_1| < |p|$, or
   (b) $p_1Gp_3p_4$ and p_2 are generalized D&P-algorithm patterns, $|p_1Gp_3p_4| < |p|$ and $|p_2| < |p|$, or
   (c) $p_1p_2Gp_4$ and p_3 are generalized D&P-algorithm patterns, $|p_1p_2Gp_4| < |p|$ and $|p_3| < |p|$, or
   (d) $p_1p_2Gp_3$ and p_4 are generalized D&P-algorithm patterns, $|p_1p_2Gp_3| < |p|$ and $|p_4| < |p|$.

A generalized D&P-algorithm pattern p is a D&P-algorithm pattern if p does not contain G. We say that secondary structure R is an *D&P secondary structure* it corresponds to a D&P-algorithm pattern.

We next define the class of D&P patterns. Let p be a string. We say that $p \xrightarrow[\text{D\&P}]{} p'$ if $p' = p \downarrow \sigma$ for some σ that is self-adjacent or directly nested in p, or if $p' = (p \downarrow \sigma) \downarrow \tau$ for some symbols σ and τ such that the substring $\sigma\tau\sigma\tau$ is in p. $p \xrightarrow[\text{D\&P}]{*} p'$ if $p = p'$ or $\exists$ patterns $p_1, \ldots, p_k$ such that $p \xrightarrow[\text{D\&P}]{} p_1 \xrightarrow[\text{D\&P}]{} \ldots p_k \xrightarrow[\text{D\&P}]{} p'$. Pattern p is *D&P* if $p \xrightarrow[\text{D\&P}]{*} \epsilon$.

Theorem 3 *The D&P secondary structures are exactly the structures corresponding to D&P patterns.*

The proof of Theorem 3 is similar in spirit to that of Theorem 1.

3.4 Lyngsø and Pedersen (L&P) Structures

Lyngsø and Pedersen [8] outline a dynamic programming algorithm for a restricted class of structures. The class includes structures of the form $s_1 s_2 s_1' s_2'$ where both $s_1 s_1'$ and $s_2 s_2'$ are pseudoknot free. We call such structures L&P structures. Similar to the characterizations above, we can also describe the L&P structure class as follows.

Let p be a string. We say that $p \xrightarrow[\text{D\&P}]{} p'$ if $p' = p \downarrow \sigma$ for some σ that is self-adjacent or directly nested in p, or if $p = \sigma\tau\sigma\tau$ and $p' = \epsilon$. $p \xrightarrow[\text{L\&P}]{*} p'$ if $p = p'$ or $\exists$ patterns $p_1, \ldots, p_k$ such that $p \xrightarrow[\text{L\&P}]{} p_1 \xrightarrow[\text{L\&P}]{} \ldots p_k \xrightarrow[\text{L\&P}]{} p'$. Pattern p is *L&P* if $p \xrightarrow[\text{L\&P}]{*} \epsilon$. Secondary structure R is *a L&P secondary structure* if it corresponds to a L&P pattern p.

The following theorem follows easily from the above definitions.

Theorem 4 *The L&P structure class is exactly the set of L&P secondary structures.*

The algorithm outlined by Lyngsø and Pedersen can also handle structures of the form $s_1 s_2 s_1' s_2' s_1''$ where both $s_1 s_1' s_1''$ and $s_2 s_2'$ are pseudoknot free. We call this class L&P$^+$. Lyngso and Pedersen [9] note that PKF $\subset$ L&P$^+$ $\subset$ R&E. However, L&P$^+$ is incomparable with the D&P and A&U classes, since $ABACBC$ is in L&P$^+$ but not in A&U, while $ABCDCDAB$ is in D&P but not in L&P$^+$. In what follows, we work with the simpler L&P class.

3.5 Containments Between the Classes

We can now prove the following theorem:

Theorem 5

$$PKF \subset L\&P \subset D\&P \subset A\&U \subset R\&E.$$

Proof Consider each of $\xrightarrow[\text{PKF}]{}$, $\xrightarrow[\text{L\&P}]{}$, $\xrightarrow[\text{D\&P}]{}$, $\xrightarrow[\text{A\&U}]{}$ and $\xrightarrow[\text{R\&E}]{}$ as relations. Each relation is specified using rules that generalize that relation earlier in the above list. Therefore,

$$p \xrightarrow[\text{PKF}]{*} p' \Rightarrow p \xrightarrow[\text{L\&P}]{*} p' \Rightarrow p \xrightarrow[\text{D\&P}]{*} p' \Rightarrow p \xrightarrow[\text{A\&U}]{*} p' \Rightarrow p \xrightarrow[\text{R\&E}]{*} p'.$$

From our characterizations, it follows that the following patterns separate the classes (details omitted): (i) *ABAB* is in L&P - PKF, (ii) *ABCBCA* is D&P - L&P, (iii) *ABCBDADC* is A&U - D&P [5], and (iv) *ABCABC* is R&E - A&U [5]. Also, *ABCADBECDE* is not R&E [11]. □

4 Testing Membership in Structure Classes

We have developed linear-time tests for membership in the R&E, D&P, and the L&P classes. Here, we describe the algorithms, and report on the results of applying the algorithms on several biological structures.

4.1 A Linear Time Algorithm to Recognize R&E Structures

Algorithm 1 tests if a pattern over some fixed alphabet Σ is a R&E pattern. The pattern is scanned from the left and the $\underset{\text{R\&E}}{\longrightarrow}$ operation is applied when possible. In the algorithm, τ is a symbol variable over $\Sigma \cup \{\epsilon\}$ and p_L and p_R are string variables over Σ^*. Let s and s' be string variables over Σ^*. If $s' = \sigma'_1\sigma'_2 \ldots \sigma'_k$ with all $\sigma'_i \in \Sigma$ and $k \geq 1$, then we define the operation $\tau s \leftarrow s'$ to set τ to σ'_1 and s to $\sigma'_2 \ldots \sigma'_k$. Similarly, the operation $s\tau \leftarrow s'$ sets τ to σ'_k and s to $\sigma'_1 \ldots \sigma'_{k-1}$. If $s' = \epsilon$ then the operations $\tau s \leftarrow s'$ and $s\tau \leftarrow s'$ set $\tau = s = \epsilon$.

algorithm R&E-Pattern-Test
input: pattern $p = \sigma_1\sigma_2 \ldots \sigma_k \in \Sigma^k$ with $k \geq 2$
output: yes, if p is an R&E pattern and no otherwise

$p_L \leftarrow \epsilon$; $\tau \leftarrow \sigma_1$; $p_R \leftarrow \sigma_2 \ldots \sigma_k$;
repeat
  if some σ is directly adjacent to τ **then**
    arbitrarily choose any such σ;
    $p_L \leftarrow p_L \downarrow \sigma$; $p_R \leftarrow p_R \downarrow \sigma$;

  elseif τ is self-adjacent or directly adjacent **then**
    $p'_L \leftarrow p_L \downarrow \tau$; $p'_R \leftarrow p_R \downarrow \tau$;
    if $p'_L \neq \epsilon$ then $p_L\tau \leftarrow p'_L$; $p_R \leftarrow p'_R$
    else $\tau p_R \leftarrow p'_R$; $p_L \leftarrow p'_L$;

  else $p_L \leftarrow p_L\tau$; $\tau p_R \leftarrow p_R$;
until $\tau = \epsilon$;
if $p_L = \epsilon$ **then return** yes **else return** no

Algorithm 1: A test for R&E patterns.

If the pattern is stored as a doubly linked list of symbols, with additional links between the two instances of each symbol, then each iteration of the repeat loop can be implemented in $O(1)$ time. Thus, the total time is $O(k)$ on an input pattern of length k. In Theorem 6 we prove that algorithm *R&E-Pattern-Test* recognizes exactly the R&E patterns. The following lemma is key to the proof:

Lemma 1 *Let p be a R&E pattern and let $p \underset{R\&E}{\longrightarrow} p \downarrow \sigma$. Then $p \downarrow \sigma$ is R&E.*

Proof Suppose that $p = p_0 \underset{\text{R\&E}}{\longrightarrow} p_1 \ldots \underset{\text{R\&E}}{\longrightarrow} p_k \underset{\text{R\&E}}{\longrightarrow} \epsilon$. Let i be such that $p_i = p_{i-1} \downarrow \sigma$. The proof is given in two cases.

First, suppose that σ is self-adjacent in p, or that σ is directly adjacent to ρ in p, where $\rho \in p_i$. Then we claim that

$$p \downarrow \sigma = p_0 \downarrow \sigma \underset{\text{R\&E}}{\longrightarrow} p_1 \downarrow \sigma \underset{\text{R\&E}}{\longrightarrow} \ldots p_{i-1} \downarrow \sigma = p_i \underset{\text{R\&E}}{\longrightarrow} \ldots p_k \downarrow \sigma \underset{\text{R\&E}}{\longrightarrow} \epsilon, \tag{1}$$

from which the lemma follows. To see why (1) is true, fix any $j, 1 \leq j < i$ and let $p_j = p_{j-1} \downarrow \tau$. We need to show that $p_{j-1} \downarrow \sigma \underset{\text{R\&E}}{\longrightarrow} p_j \downarrow \sigma$. If τ is self-adjacent or directly adjacent to some $\rho \neq \sigma$ in p_{j-1}, then it is also the case that τ is self-adjacent or directly adjacent to some $\rho \neq \sigma$ in $p_{j-1} \downarrow \sigma$, and so $p_{j-1} \downarrow \sigma \underset{\text{R\&E}}{\longrightarrow} p_j \downarrow \sigma$. Otherwise, it must be that τ is directly adjacent to σ in p_{j-1}. Then if σ is self-adjacent in p, $\tau\sigma\sigma\tau$ is a substring of p_{j-1}, in which case τ is self-adjacent in $p_{j-1} \downarrow \sigma$. If $\sigma\rho\sigma$ is a substring of p for some ρ, then $\tau\rho\tau$ is a substring of $p_{j-1} \downarrow \sigma$, in which case τ is directly adjacent in $p_{j-1} \downarrow \sigma$. Finally, if p contains two disjoint substrings x, y, both of length 2, such that for some ρ, ρ and σ are both in x and ρ and σ are both in y, and $\rho \in p_i$ then $p_{j-1} \downarrow \sigma$ must contain two

substrings x, y, both of length 2, such that τ and ρ are both in x and τ and ρ are both in y, in which case again τ is directly adjacent to ρ in $p_{j-1} \downarrow \sigma$. In all cases, we can conclude that $p_{j-1} \downarrow \sigma \xrightarrow[\text{R\&E}]{} p_j \downarrow \sigma$.

Second, suppose that p contains two disjoint substrings x, y, both of length 2, such that for some ρ, ρ and σ are both in x, ρ and σ are both in y, and ρ is not in p_i. Let h be such that $p_h = p_{h-1} \downarrow \rho$, where $h < i$. For $h \leq l \leq i-1$, let p'_h be obtained by replacing σ with ρ. Note that $p_{h-1} \downarrow \sigma = p'_h$ and $p'_{i-1} \xrightarrow[\text{R\&E}]{} p'_{i-1} \downarrow \rho = p_i$. In this case we claim that

$$p \downarrow \sigma = p_0 \downarrow \sigma \xrightarrow[\text{R\&E}]{} \ldots p_{h-1} \downarrow \sigma = p'_h \xrightarrow[\text{R\&E}]{} p'_{h+1} \xrightarrow[\text{R\&E}]{} \ldots p'_{i-1} \xrightarrow[\text{R\&E}]{} p_i \xrightarrow[\text{R\&E}]{} p_{i+1} \xrightarrow[\text{R\&E}]{} \ldots p_k \downarrow \sigma \xrightarrow[\text{R\&E}]{} \epsilon.$$

To see why, for $j < h$, the argument that $p_{j-1} \downarrow \sigma \xrightarrow[\text{R\&E}]{} p_j \downarrow \sigma$ is as in the first case above. It remains to show that for $h < j \leq i-1$, $p'_{j-1} \xrightarrow[\text{R\&E}]{} p'_j$. Fix any such j. Let $p_j = p_{j-1} \downarrow \tau$, (so that also $p'_j = p'_{j-1} \downarrow \tau$). If τ is directly adjacent to σ in p_{j-1}, then τ is directly adjacent to ρ in p'_{j-1}, in which case $p'_{j-1} \xrightarrow[\text{R\&E}]{} p'_j$. □

Theorem 6 *Algorithm R&E-Pattern-Test outputs yes on input p if and only if p is R&E.*

Proof Let $p_L^{(i)}$, $\tau^{(i)}$, and $p_R^{(i)}$ be the values of variables p_L, τ, and p_R at the end of the ith iteration of the repeat loop. Let l be the total number of iterations of the repeat loop (since at every iteration, either $|p_R|$ or $|p_L|$ decreases, the algorithm must halt). It is straightforward to show that for each i, $1 \leq i \leq l$, either $p_L^{(i-1)}\tau^{(i-1)}p_R^{(i-1)} = p_L^{(i)}\tau^{(i)}p_R^{(i)}$ or $p_L^{(i-1)}\tau^{(i-1)}p_R^{(i-1)} \xrightarrow[\text{R\&E}]{} p_L^{(i)}\tau^{(i)}p_R^{(i)}$.
Also, if the algorithm returns yes, then $p_L^{(l)}\tau^{(l)}p_R^{(l)} = \epsilon$. Therefore, if the algorithm returns yes, p is R&E.

If the algorithm returns no, then $|p_L^{(l)}| > 0$ and also $\tau^{(l)}p_R^{(l)} = \epsilon$. Therefore, $p \xrightarrow[\text{R\&E}]{*} p_L^{(l)}$. From Lemma 1 it follows that if $p_L^{(l)}$ is not R&E, then p is not R&E. To show that $p_L^{(l)}$ is not R&E we show that the following invariant is true for each $i, 0 \leq i \leq l$.

Invariant: No symbol is self-adjacent in $p_L^{(i)}$, and if any symbol σ is directly adjacent in $p_L^{(i)}$ then $\sigma\tau^{(i)}\sigma$ is a substring of $p_L^{(i)}$.

Since $p_L^{(0)} = \epsilon$, the invariant is true for $i = 0$. Suppose the invariant is true for $i-1$. We show that it is true for i. It is straightforward to show that if $p_L^{(i)}$ is a prefix of $p_L^{(i-1)}$, then the invariant holds for i. Otherwise, it must be the case that for some σ, either $\sigma\tau^{(i-1)}$ or $\tau^{(i-1)}\sigma$ (or both) is in $p_L^{(i-1)}$, $p_L^{(i)} = p_L^{(i-1)} \downarrow \sigma$, and $\tau^{(i)} = \tau^{(i-1)}$.

Let α and β be the symbols (if any) just before and just after $\tau^{(i-1)}$ in $p_L^{(i)}$. Then α and β are the only symbols which could possibly be self-adjacent or directly adjacent in $p_L^{(i)}$ but not in $p_L^{(i-1)}$. Since the only change to α and β is that now $\tau^{(i-1)}$ is their neighbour, and since neither α nor β can equal $\tau^{(i-1)}$, they cannot be self-adjacent. However, if $\alpha = \beta$ then α is directly adjacent to $\tau^{(i-1)}$, because $\alpha\tau^{i-1}\alpha$ is a substring of $p_L^{(i)}$. Therefore, since $\tau^{(i)} = \tau^{(i-1)}$, the invariant holds for i. From the Invariant, we conclude that no symbol in $p_L^{(l)}$ is self-adjacent or directly adjacent. Therefore, $p_L^{(l)}$ is not R&E. □

4.2 Linear Time Algorithms to Recognize D&P and L&P Structures

Algorithms 2 and 3 test for membership in the D&P and L&P structure classes, respectively. They are very similar to Algorithm 1, and their correctness proofs are also similar (details omitted).

4.3 Classification of Biological Structures

We applied our algorithms to classify secondary structures from PseudoBase (PBase) [3], the Nucleic Acids Database (NDB) [2], 16S and 23S ribosomal RNA and Group I and Group II Introns from the Gutell Database [6]. (We also considered 5S RNA secondary structures, and all were pseudoknot free.) We considered only secondary structures

algorithm *D&P-Pattern-Test*
input: pattern $p = \sigma_1\sigma_2 \ldots \sigma_k \in \Sigma^k$ with $k \geq 2$
output: yes, if p is a D&P pattern and no otherwise

$p_L \leftarrow \epsilon$; $\tau \leftarrow \sigma_1$; $p_R \leftarrow \sigma_2 \ldots \sigma_k$;
repeat
  if some σ is directly nested with respect to τ **then**
    $p_L \leftarrow p_L \downarrow \sigma$; $p_R \leftarrow p_R \downarrow \sigma$;

  elseif τ is self-adjacent **then**
    $p'_L \leftarrow p_L \downarrow \tau$; $p'_R \leftarrow p_R \downarrow \tau$;
    if $p'_L \neq \epsilon$ then $p_L\tau \leftarrow p'_L$; $p_R \leftarrow p'_R$
    else $\tau p_R \leftarrow p'_R$; $p_L \leftarrow p'_L$;

  else if $\sigma\tau\sigma$ is a suffix of p_L for some σ **then**
    $p'_L \leftarrow (p_L \downarrow \sigma) \downarrow \tau$;
    if $p'_L \neq \epsilon$ then $p_L\tau \leftarrow p'_L$; $p_R \leftarrow p'_R$
    else $\tau p_R \leftarrow p'_R$; $p_L \leftarrow p'_L$;

  else $p_L \leftarrow p_L\tau$; $\tau p_R \leftarrow p_R$;
until $\tau = \epsilon$;
if $p_L = \epsilon$ **then return** yes **else return** no.

Algorithm 2: A test for D&P patterns.

algorithm *L&P-Pattern-Test*
input: pattern $p = \sigma_1\sigma_2 \ldots \sigma_k \in \Sigma^k$ with $k \geq 2$
output: yes, if p is an L&P pattern and no otherwise

$p_L \leftarrow \epsilon$; $\tau \leftarrow \sigma_1$; $p_R \leftarrow \sigma_2 \ldots \sigma_k$;
repeat
  if some σ is directly nested with respect to τ **then**
    $p_L \leftarrow p_L \downarrow \sigma$; $p_R \leftarrow p_R \downarrow \sigma$;

  elseif τ is self-adjacent **then**
    $p'_L \leftarrow p_L \downarrow \tau$; $p'_R \leftarrow p_R \downarrow \tau$;
    if $p'_L \neq \epsilon$ then $p_L\tau \leftarrow p'_L$; $p_R \leftarrow p'_R$
    else $\tau p_R \leftarrow p'_R$; $p_L \leftarrow p'_L$;

  else $p_L \leftarrow p_L\tau$; $\tau p_R \leftarrow p_R$;
until $\tau = \epsilon$;
if $p_L = \epsilon$ or $|p_L| = 4$ **then return** yes **else return** no.

Algorithm 3: A test for L&P patterns.

with no occurrences of triple base stacking. Structures in these data sets may contain *isolated base pairs*, namely base pairs $i \cdot j$ such that neither $(i+1) \cdot (j-1)$ nor $(i-1) \cdot (j+1)$ is in the structure. Since isolated base pairs are sometimes considered to be tertiary rather than secondary structure, we classified the structures before and after removal of the isolated base pairs. Our results are presented in Table 1.

(a)

	PBase	16S	23S	Gp I Intron	Gp II Intron	NDB
# Strs	240	152	69	10	3	12
Avg. #Bps	14.2	466	763.1	128.9	209	312.4
PKF	0	0	14	0	0	1
L&P	231	12	14	10	0	1
D&P	232	150	14	10	0	5
R&E	240	152	25	10	0	7

(b)

	PBase	16S	23S	Gp I Intron	Gp II Intron	NDB
# Strs	240	152	69	10	3	12
Avg. #Bps	14.1	455.6	733	126.1	207	268
PKF	0	0	21	0	0	6
L&P	231	12	21	10	0	6
D&P	232	152	21	10	0	11
R&E	240	152	69	10	0	12

Table 1: Structure classification. Part (a) is for structures with isolated base pairs not removed and part (b) is for structures with isolated base pairs removed. In each part, columns 2-7 present data for each RNA data set. For each data set (column), the entry in first row lists the number of structures in the data set. The second row lists the average number of base pairs in the structures once isolated base pairs are removed. The remaining rows list the number of structures of the data set that are in the PKF, L&P, D&P, and R&E classes, respectively.

The R&E structure class is indeed very general, containing all of the secondary structures with isolated base pairs removed except for three (long) Group II Intron sequences. The D&P class does not contain most of the 23S rRNA structures, and contains 8 fewer PseudoBase structures than the R&E class, but otherwise compares well with the R&E class. The L&P class additionally misses almost all of the 16S rRNA structures, yet still contains almost all of the structures in PseudoBase.

5 Conclusions

Our characterizations of structure classes handled by RNA secondary structure prediction algorithms, and our tests for membership in these classes, provide the first means for evaluating the generality of current algorithms. The results show that current algorithms do in fact handle a wide range of known biological structures, though not all such structures.

There is a trade-off between algorithm complexity and the generality of the class of structures that can be handled by the algorithm. An interesting question is whether faster algorithms can be found for any of the classes L&P, D&P, A&U, or R&E, or whether algorithms with comparable running times but that handle a more general (and biologically interesting) class of structures can be obtained.

In future work, we will develop a linear time algorithm for characterizing the A&U structure class.

6 Acknowledgements

We thank Mirela Andronescu, Matthew Cook, Robert Dirks, Holger Hoos, Niles Pierce, Joseph Schaeffer, Dan Tulpan, and Erik Winfree for valuable feedback on earlier versions of this work.

References

[1] Akutsu, T. (2000). Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots, *Discrete Applied Mathematics*, 104:45–62.

[2] Berman, H. M. et al. (1992). The Nucleic Acid Database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, 63:751–759.

[3] Batenburg, F. H. D. van, A. P. Gultyaev, C. W. A. Pleij, J. Ng, and J. Oliehoek (2000). Pseudobase: a database with RNA pseudoknots. *Nucl. Acids Res.* 28(1):201–204.

[4] Dennis, C. (2002). The brave new world of RNA, *Nature*, 418(11):122–124.

[5] Dirks, R. M. and N. A. Pierce (2003). A partition function algorithm for nucleic acid secondary structure including pseudoknots, *J. Comput. Chem.*, 24(13):1664-1677.

[6] Cannone J. J. et al. (2002). The Comparative RNA Web (CRW) Site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BioMed Central Bioinformatics*, 3:2. [Correction: BioMed Central Bioinformatics. 3:15.]

[7] Hofacker, I. L., Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster (1994). Fast folding and comparison of RNA secondary structures, *Monatsh. Chem.* 125:167–188.

[8] Lyngsø R. B. and C. N. Pedersen (2000). Pseudoknots in RNA secondary structures, Proc. 4th Annual International Conference on Computational Molecular Biology (RECOMB), 201 – 209.

[9] Lyngsø R. B. and C. N. Pedersen (2000). RNA pseudoknot prediction in energy-based models, *J. Computational Biology*, 7(3):409–427.

[10] Mathews, D. H., J. Sabina, M. Zuker and D.H. Turner (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, *J. Mol. Biol.*, 288:911–940.

[11] Rivas E. and S. R. Eddy (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots, *J. Molecular Biology*, 285:2053–2068.

[12] Rivas E. and S. R. Eddy (2000). The language of RNA: A formal grammar that includes pseudoknots, *Bioinformatics*, 16:334-340.

[13] Uemura Y., A. Hasegawa, S. Kobayashi and T. Yokomori (1999). Tree adjoining grammars for RNA structure prediction, *Theoretical Computer Science*, 210:277–303.

[14] Zuker M. and P. Steigler (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, *Nucleic Acids Res.*, 9:1330–1348.