Statistical RNA Secondary Structure Sampling Based on a Length-Dependent SCFG Model

Anika Scheid* and Markus E. Nebel
Department of Computer Science, University of Kaiserslautern,
P.O. Box 3049, D-67653 Kaiserslautern, Germany
{a_scheid,nebel}@cs.uni-kl.de

Abstract

One of the fundamental problems in computational structural biology is the prediction of RNA secondary structures from a single sequence. To solve this problem, mainly two different approaches have been used over the past decades: the free energy minimization (MFE) approach which is still considered the most popular and successful method and the competing stochastic context-free grammar (SCFG) approach. While the accuracy of the MFE based algorithms is limited by the quality of underlying thermodynamic models, the SCFG method abstracts from free energies and instead tries to learn about the structural behavior of the molecules by training the grammars on known real RNA structures, making it highly dependent on the availability of a rich high quality training set. However, due to the respective problems associated with both methods, new statistics based approaches towards RNA structure prediction have become increasingly appreciated. For instance, over the last years, several statistical sampling methods and clustering techniques have been invented that are based on the computation of partition functions (PFs) and base pair probabilities according to thermodynamic models. A corresponding SCFG based statistical sampling algorithm for RNA secondary structures has been studied just recently in [NSar]. Notably, this probabilistic method is capable of producing accurate (prediction) results, where its worst-case time and space requirements are equal to those of common RNA folding algorithms for single sequences.

The aim of this work is to present a comprehensive study on how enriching the underlying SCFG by additional information on the lengths of generated substructures (i.e. by incorporating *length-dependencies* into the SCFG based sampling algorithm, which is actually possible without significant losses in performance) affects the reliability of the induced RNA model and the accuracy of sampled secondary structures. As we will see, significant differences with respect to the overall quality of generated sample sets and the resulting predictive accuracy are typically implied. In principle, when considering the more specialized length-dependent SCFG model as basis for statistical sampling, a higher accuracy of predicted foldings can be reached at the price of a lower diversity of generated candidate structures (compared to the more general traditional SCFG variant or sampling based on PFs that rely on free energies).

1 Introduction

The function of an RNA molecule in the cell's processes is often to a large extend determined by its complete 3D structure, called its *tertiary structure*. As most of the tertiary structure is given by the intramolecular base pairings in the plane, it has proven convenient to first search for its 2D structure, called the *secondary structure* of the molecule.

To date, the most common and still most appreciated approach for computationally predicting the RNA secondary structure of a single sequence is based on the free energy minimization paradigm; it will be called MFE approach in the sequel. Over the last decades, the most successful and popular method for energy minimization has been the use of dynamic programming (DP) algorithms. While early methods, like [NPGK78, NJ80, ZS81], computed only one (optimal) structure, the minimum free energy structure of the molecule, several efficient algorithms have been developed over the years for generating a set of suboptimal foldings (see, e.g., [WFHS99, Zuk89]). Some of the corresponding implementations, such as for example the Mfold software [Zuk03] or the Vienna package [Hof03], have become widely used tools to predict RNA structure.

*Corresponding author. The research of this author has been supported by the Carl-Zeiss-Stiftung.

However, one major problem of MFE based DP algorithms for RNA secondary structure prediction is that the predicted set of suboptimal foldings often contains many structures that are not significantly different. To obtain more variation in the set of suboptimal structures, as well as to address the problem of a statistical representation of probable foldings, a corresponding physics-based statistical sampling algorithm has been introduced [DL03]. This algorithm can be seen as a sampling extension of the *partition function* (PF) approach for computing base pair probabilities as introduced in [McC90], which then described a novel (albeit still physics-based) alternative to the common MFE approach and laid the foundation for statistical characterizations of the equilibrium ensemble of RNA secondary structures. Similar to its MFE counterpart, the algorithm for calculating the PFs and base pair probabilities is realized by DP routines that take cubic time and require quadratic storage.

Another well-established approach towards the computational prediction of the secondary structure of a single RNA molecule is based on *stochastic context-free grammars* (SCFGs). Briefly, SCFGs are an extension to the concept of traditional context-free grammars (CFGs) in a sense that they do not only model the class of objects (language) to be generated, but also define a (usually non-uniform) probability distribution on its elements. Just like MFE methods, the corresponding algorithms are traditionally realized by DP routines that run in cubic time and require quadratic storage, but in contrast to physics-based methods, the main focus of attention is laid on the typical structural composition of foldings. Examples for successful applications of this alternative methodology can be found, for instance, in [DE04], where an efficient implementation is given by the popular Pfold tool [KH99, KH03].

Notably, SCFGs try to learn about the typical behavior of a particular class of RNAs from statistical grounds. This is realized by employing appropriate training procedures for estimating probabilities for the distinct production rules (i.e. for calculating estimates for the respective grammar parameters). In fact, as there is no *lab-based* prior to the grammar parameters like the standard Turner model [MSZT99] for MFE and PF approaches, the corresponding distribution has to be derived from a collection of real-life RNA data (RNA sequences with annotated secondary structures) when considering such *probabilistic* prediction methods. Note that in the case of SCFGs, one actually uses *generative* parameter training, whereas for more complex probabilistic models, like for instance *conditional log-linear model (CLLMs)*, which usually have the power to represent more complex scoring schemes (e.g., a Mfold-like energy based one as considered in [DWB06]), one needs to employ *discriminative* training methods.

Principally, generative (SCFG based) training can easily be realized by counting the observed frequencies of applications of the distinct production rules of an unambiguous SCFG (yielding a maximum likelihood estimator), by expectation maximization or similar methods from machine-learning. That way, the resulting estimates of the grammar parameters are adapted to a particularly considered data set. As for these data, we again have two different choices: First, we may consider a training set where only structures of a single biological class (e.g., tRNA) are contained. Then, we may expect that all structural properties (including aspects which are caused by interaction with proteins or by other non-energetic details of RNA folding) that are typical to this class are trained into the respective parameter values. For a general model of RNA folding, this has to be assumed some sort of "over-specialization", since we cannot expect the model to generalize well to new data from a different class. Second, we may use a rich training set of mixed biological classes. In that case, the danger of a potential lack of generalization is much smaller, but we lose the chance to capture some class-specific properties of the structures within our model.

In both cases, the main problem that comes inherently with the SCFG approach for modeling RNA structures and limits the performance of the corresponding computational prediction methods is that it is obviously highly dependent on the availability of a rich, reliable training set in order to minimize the danger of *overfitting*. In statistics, the term overfitting is used to express that a statistical model describes random error or noise instead of the underlying relationship, caused by the fact that there are too many parameters relative to the number of observations. Hence, in our context overfitting potentially occurs when the number of grammar parameters to be estimated is too large for the training data, that is if the considered RNA data are not rich enough to reliably estimate the distinct parameters. Obviously, this might especially be the case when using an excessively complex SCFG design that distinguishes between all different features in RNA structure aiming at a highly realistic model (for which a large number of parameters needs to be determined).

Nevertheless, it should be mentioned that the dependence of the resulting accuracy on the used model parameters is not only a major problem in case of probabilistic approaches, but it indeed also limits the performance of physics-based methods. In fact, the (usually many hundreds of) thermodynamic parameters used in standard state-of-the-art energy models are mostly estimated from experimental results (on the basis of diverse structural RNAs), but folding processes of RNA molecules are usually to a large part controlled by a number of additional non-energetic effects. The corresponding needed information on folding kinetics, that is certain important chemical aspects (like for example the influence of proteins/enzymes or the effect of co-transcriptional folding) can simply not be incorporated into physics-based models, since energy parameters are actually measured in vitro. Therefore, one increasingly accepted solution to these problems is to estimate the thermodynamic parameters from pure RNA structure databases (of one particular type of RNAs, respectively) via Bayesian statistical inference (where the experimentally derived Turner parameter values can be used for prior specification), see for example [Din06]. Obviously, such a Bayesian inference approach makes it possible to derive energy parameters that are suited for structure prediction (if the original type of the input sequence is known, the corresponding estimates ought to be used, in analogy to probabilistic methods). However, the accuracy of the estimated parameters unsurprisingly also strongly depends on the quality of the employed data.

Finally, note that for a long time SCFGs for RNA secondary structures have seemingly been chosen rather arbitrarily, whereas recently a number of more sophisticated[1] SCFG designs have been presented (see, e.g., [NS11, NSW11, NSar]). Particularly, the one from [NSar] has been constructed as an exact probabilistic counterpart to the standard energy model employed for example in the Sfold software, such that it could adequately be used in order to directly and reliably compare probabilistic (generative), discriminative and thermodynamic approaches. Note that aiming at more informative investigation results, the considered SCFG has been used as basis for a probabilistic statistical sampling algorithm that incorporates only comprehensive structural features and – instead of the recent thermodynamic parameters (as done in Sfold) – additional information obtained from known databases of RNA structures. In particular, the strategy studied in [NSar] relies on a probabilistic approach in order to compute sampling probabilities corresponding to those defined in [DL03] that are used for Sfold's stochastic traceback step. The sampling of base pairs and unpaired base(s) basically works in the same way, that is structures are sampled rigorously from a particular distribution of all feasible foldings for a given sequence as induced by the appropriately trained SCFG. Notably, the probabilities needed for sampling are calculated using only the considered grammar parameters (trained beforehand on a suitable database) and a collection of inside and outside probabilities (computed for a given input sequence).

Just like its PF counterpart, the SCFG based method generates a sample that is guaranteed to be statistically reproducible and representative, where both approaches have the same worst-case time and space requirements. However, the SCFG method can be used with less restrictions (one can allow hairpin loops of sizes less than three, non-canonical base pairs and bulge / interior loops of arbitrary length, due to the departure from thermodynamic models). Moreover, when comparing the results of both sampling methods, significant differences can be observed, as shown in [NSar]. One of the potentially most interesting ones is that the accuracy of shape predictions and the diversity within sample sets can be significantly improved by using the SCFG method instead of the PF variant.

Note that one basic fact in connection with SCFG approaches is that at any point, the probability for generating a particular structure motif (as modeled by the grammar) is given by the corresponding estimated parameter value (of the corresponding production rule), which does actually not depend on the *length* of the generated substructure, although in reality it often does. For example, the probability for leaving a particular fragment unpaired instead of folding at least one additional base pair on it (resulting for instance in a simple hairpin loop instead of a more stable paired substructure) is identical for any fragment length, but in nature short fragments are much more likely to be left unpaired than longer ones (as it is usually energetically more favorable to fold additional base pairs if possible).

In order to model this native behavior of RNA molecules, it seems reasonable to additionally include *length-dependencies* into traditional SCFG models. To the best of our knowledge, this idea has first been applied in [NE07] in connection with database similarity searching based on *covariance models (CMs)*. Briefly, CMs are profile SCFGs (a particular SCFG architecture) for cleanly describing both the secondary structure and the primary sequence consensus of an RNA (see, e.g., [RD94]). Principally, in [NE07], it is described how to accelerate CM searches by using a *banded* DP strategy (a standard approach in many areas of sequence analysis), which actually for each node calculates the probability of generating a subsequence of a particular length.

However, with respect to traditional probabilistic RNA secondary structure prediction methods, one might easily consider an appropriate *length-dependent* stochastic context-free grammar (LSCFG), as formally introduced in [WN11]. Basically, LSCFGs exactly address the problem sketched above, that is they are defined as an extension to the concept of conventional SCFGs such that the probabilities of the

[1] By means of ability to model the diverse structural motifs by different productions with corresponding distinct parameters, on a par with modern thermodynamic models.

productions depend on the length of the generated subword. One important aspect is that in general, LSCFG based algorithms can be implemented to have the same worst-case time and space requirements as their length-independent counterparts. Furthermore, due to the larger number of grammar parameters implied, algorithms implementing LSCFG models are obviously not only more explicit (due to the higher level of specialization), but unfortunately also more prone to overfitting than the corresponding traditional SCFG variants.
Nevertheless, motivated by the idea discussed in [Mai07] of improving the SCFG approach for RNA secondary structure prediction by explicitly considering the lengths of particular substructures, the main objectives of this paper are given as follows: First, we want to investigate to which extend the additional incorporation of length-dependencies into a sophisticated SCFG changes the quality of the induced probabilistic RNA model. Our second aim is to quantify the differences in resulting accuracy that can be observed when applying both probabilistic models (length-dependent and traditional one) to identical inputs. For our examinations, we decided to rely on the elaborate SCFG design from [NSar] and analogously use it as the basis for a probabilistic statistical sampling algorithm, since this effectively makes it possible to perform comprehensive comparisons of both variants (SCFG and LSCFG based) with respect to different meaningful applications that can immediately be considered in connection with sampling approaches. In fact, we will present a fundamental analysis of the resulting sample sets from different relevant perspectives in order to see if the incorporation of additional length information into SCFGs eventually yields a quality improvement.

The plan for the rest of the paper is given as follows: Section 2 formally introduces the considered (L)SCFG model for RNA secondary structures. The needed modifications of the original sampling algorithm from [NSar] to manage the additional length-dependencies are described in Section 3. Section 4 discusses the potentials and possible drawbacks of extending the underlying sophisticated SCFG model to a length-dependent one. Particularly, Section 4 contains important results concerning the quality of the underlying probabilistic model with respect to both overfitting and lack of generalization. Furthermore, it examines if adding length-dependency actually improves the accuracy of predictions obtained from statistical sampling and the overall quality of generated sample sets (with respect to the produced shapes). For this purpose, corresponding results obtained by the length-dependent and the traditional version of the probabilistic sampling approach are opposed to each other. Additionally, all results are compared to corresponding ones obtained with the competing PF approach implemented in the well-established Sfold program for further judgements. Finally, Section 5 concludes this paper.

2 Considered (L)SCFG Model

The aim of this section is to introduce the (length-dependent) SCFG model that will be used as foundation of the (extended) probabilistic statistical sampling method studied within this article.

2.1 Underlying Traditional SCFG Model

First, note that in this work, we decided to not recall all basic definitions and concepts regarding (stochastic) context-free grammars and languages. For a fundamental introduction on stochastic context-free languages, see for example [HF71]. However, a formal definition is given as follows:

Definition 2.1 ([FH72])**.** A *stochastic context-free grammar (SCFG)* is a 5-tuple $G = (I, T, R, S, \mathrm{Pr})$, where I (respectively T) is an alphabet (finite set) of intermediate (respectively terminal) symbols (I and T are disjoint), $S \in I$ is a distinguished intermediate symbol called *axiom*, $R \subset I \times (I \cup T)^*$ is a finite set of production rules and Pr is a mapping from R to $[0, 1]$ such that each rule $f \in R$ is equipped with a probability $p_f := \mathrm{Pr}(f)$. The probabilities are chosen in such a way that for all $A \in I$ the equality $\sum_{f \in R} p_f \cdot \delta_{Q(f),A} = 1$ holds. Here, δ is Kronecker's delta and $Q(f)$ denotes the source of the production f, i.e. the first component A of a production rule $(A, \alpha) \in R$. In the sequel, we will write $p_f : A \to \alpha$ instead of $f = (A, \alpha) \in R$, $p_f = \mathrm{Pr}(f)$.

It is worth mentioning that if a formal language is modeled by a *consistent* SCFG, then the probability distribution on the production rules of the SCFG implies a probability distribution on the words of the generated language and thus on the modeled structures[2].

[2] To ensure that a SCFG gets consistent, one can for example assign relative frequencies to the productions, which are computed by counting the production rules used in the leftmost derivations of a finite training set of words from the generated language [CG98].

As already mentioned, in [NSar], a sophisticated SCFG mirror of the thermodynamic model applied in the Sfold package has been constructed. This elaborate SCFG design serves as foundation for the corresponding SCFG based sampling method and actually differentiates between all mutually and exclusive cases that have to be considered for the derivation of the needed sampling probabilities (by employing distinct production rules for generating the diverse motifs). Actually, these conditional sampling probabilities directly correspond to those used in [DL03] for the PF approach, expect for a slight difference as regards bulge and interior loops such that the restriction of limiting their lengths[3] can be dropped. Formally, that sophisticated grammar models (a subset of) the formal language of all correctly bracketed words (according to two structural parameters $\min_{HL}$ and $\min_{\text{hel}}$) over the alphabet $\{(,),\circ\}$, where $(\,)$ and $\circ$ represents a base pair and unpaired base, respectively (see [VC85]). It is actually given as follows:

Definition 2.2 ([NSar])**.** The (unambiguous) SCFG $\mathcal{G}_s$ generating exactly all feasible secondary structures[4] is given by $\mathcal{G}_s = (\mathcal{I}_{\mathcal{G}_s}, \Sigma_{\mathcal{G}_s}, \mathcal{R}_{\mathcal{G}_s}, S)$, where $\mathcal{I}_{\mathcal{G}_s} = \{S, T, C, A, P, L, F, H, G, B, M, O, N, U, Z\}$, $\Sigma_{\mathcal{G}_s} = \{(,),\circ\}$ and for $m_h := \min_{HL} \geq 1$ and $m_s := \min_{\text{hel}} \geq 1$, $\mathcal{R}_{\mathcal{G}_s}$ contains exactly the following rules:

$p_1 : S \to T,$ ⇝ initiate exterior loop

$p_2 : T \to C,\quad p_3 : T \to A,\quad p_4 : T \to CA,\quad p_5 : T \to AT,\quad p_6 : T \to CAT,$ ⇝ shape of exterior loop

$p_7 : C \to ZC,\quad p_8 : C \to Z,$ ⇝ strands in exterior loop

$p_9 : A \to (^{m_s}L)^{m_s},$ ⇝ initiate helix

$p_{10} : P \to (L),$ ⇝ extend helix

$p_{11} : L \to F,\quad p_{12} : L \to P,\quad p_{13} : L \to G,\quad p_{14} : L \to M,$ ⇝ initiate any loop

$p_{15} : F \to Z^{m_h-1}H,$ ⇝ start hairpin loop

$p_{16} : H \to ZH,\quad p_{17} : H \to Z,$ ⇝ extend hairpin loop

$p_{18} : G \to BA,\quad p_{19} : G \to AB,\quad p_{20} : G \to BAB,$ ⇝ shape of bulge/interior loop

$p_{21} : B \to ZB,\quad p_{22} : B \to Z,$ ⇝ strands in bulge/interior loop

$p_{23} : M \to UAO,$ ⇝ first substructure of multiple loop

$p_{24} : O \to UAN,$ ⇝ second substructure of multiple loop

$p_{25} : N \to UAN,\quad p_{26} : N \to U,$ ⇝ kth substructure of multiple loop, $k \geq 3$

$p_{27} : U \to ZU,\quad p_{28} : U \to \epsilon,$ ⇝ strands in multiple loop

$p_{29} : Z \to \circ.$ ⇝ unpaired base

Note that the unambiguity of such rather complex grammars can readily be proven by describing the construction of their rule sets, as done for example in [NSW11]. Briefly, one starts with a rather simple and small (so-called *lightweight*) grammar that models only the basic structure motifs and specializes it (by replacing single productions that model one particular type of substructure by a bunch of corresponding new productions for generating the respective special types of substructures to be considered) until all substructures that need to be distinguished are represented by separate rules (and parameters). In order to avoid ambiguity, we only have to take care that at any point (where a more general old rule is replaced by a set of more specialized new ones), none of the considered alternative structure motifs can be constructed from more than one production.

Finally, it should be mentioned that in [NSar], the parameters for the corresponding SCFG model (for secondary structures on RNA sequences) are given by products of *transition probabilities* for the productions in $\mathcal{R}_{\mathcal{G}_s}$ and *emission probabilities* for the four possible choices of unpaired bases (here represented by $\circ$) and for the 16 different base pairs (that are represented by $(\,)$). Note that this separation into rule and emission probabilities eventually allows us to only consider the productions of the unambiguous grammar $\mathcal{G}_s$ modeling the class of all feasible secondary structures, although we actually had to deal with the larger set of productions of a corresponding ambiguous grammar $\mathcal{G}_r$ generating any possible RNA sequence (where the derivation trees uniquely correspond to the different secondary structures for that

[3] When applying the PF approach, one has to choose a constant value for the parameter $\max_{BL}$ which defines the maximum allowed size of single-stranded regions in bulge and interior loops, as this ensures a cubic worst-case time complexity. For applications, $\max_{BL} = 30$ is a common choice.

[4] Feasible structures contain neither hairpin loops consisting of less than $\min_{HL}$ unpaired nucleotides nor helices formed by less than $\min_{\text{hel}}$ consecutive base pairs. In literature, commonly used choices for these parameters are given by $\min_{HL} \in \{1, 3\}$ and $\min_{\text{hel}} \in \{1, 2\}$.

sequence). Notably, if all emissions (for unpaired bases and base pairs, respectively) come from the same distribution (i.e., for any considered loop type, one uses the same emission probabilities for unpaired bases located within and base pairs closing a corresponding loop), this separation usually reduces the number of free parameters that need to be estimated by corresponding training procedures in a very significant way. Hence, under that assumption (which by the way has become custom in connection with RNA modeling in order to minimize the danger of overfitting), the number of free parameters that have to be trained in our case is limited by $\text{card}(\mathcal{R}_{\mathcal{G}_s}) - \text{card}(\mathcal{I}_{\mathcal{G}_s}) + \text{card}(\Sigma_{\mathcal{G}_r}) + \text{card}(\Sigma_{\mathcal{G}_r})^2 = 29 - 15 + 4 + 16 = 34$.

2.2 Incorporation of Length-Dependencies

As already indicated in Section 1, in an attempt to improve the ability of the underlying stochastic model to capture typical structural features of a particular RNA family within its parameters, we want to additionally incorporate length-dependencies according to the following definition:

Definition 2.3 ([WN11])**.** A *length-dependent stochastic context-free grammar (LSCFG)* is defined as a SCFG $G = (I, T, R, S, \text{Pr})$ with the following exceptions:

- $\text{Pr}\ : R \times \mathbb{N} \to [0, 1]$ now takes a second argument (length of subword generated).
- The constraint on the probabilities changes to:
  $\forall A \in I\ \forall n \in \mathbb{N} : \sum_{A \to \alpha \in R} \text{Pr}(A \to \alpha, n) \in \{0, 1\}$.
- Additionally, we introduce a probability distribution $\text{Pr}(n)$ on the lengths of the words in $\mathcal{L}(G)$, i.e. $\sum_{n \in \mathbb{N}\ :\ T^n \cap \mathcal{L}(G) \neq \emptyset} \text{Pr}(n) = 1$.
- Let $\text{len}(A \to \alpha)$ denote the length of a specific rule application $A \to \alpha$ in a parse tree, which is defined as the length of the (terminal) subword finally generated from $A \to \alpha$. Furthermore, for $\alpha \in (I \cup T)^*$ and $n \in \mathbb{N}$, we denote by $c_{\alpha,n}$ the number of different assignments of lengths to the symbols of α that satisfy:
  - Terminals are always assigned a length of 1.
  - A nonterminal B can be assigned any length l for which there is $w \in T^l$ such that $\text{Pr}(B \Rightarrow w) > 0$.
  - The assigned lengths add up to n.

  The probability of a parse tree for a word of length n is then $\text{Pr}(n)$ times the product of the probabilities of all rule applications $A \to \alpha$ in the tree multiplied by ${}^1/_{c_{\alpha,\text{len}(A \to \alpha)}}$.

Note that the factors ${}^1/_{c_{\alpha,\text{len}(A \to \alpha)}}$ and $\text{Pr}(n)$ are necessary to ensure a probability distribution on the language that is generated by the LSCFG (see [WN11] for details). In fact, considering a conventional SCFG, the probability of a parse tree δ is given by

$$\prod\nolimits_{A \to \alpha \text{ applied in } \delta} \text{Pr}(A \to \alpha),$$

whereas for the corresponding LSCFG, the probability of a parse tree δ for a terminal word $w \in T^n$ is defined by

$$\text{Pr}(n) \cdot \prod\nolimits_{A \to \alpha \text{ applied in } \delta} \text{Pr}(A \to \alpha, \text{len}(A \to \alpha)) \cdot {}^1/_{c_{\alpha,\text{len}(A \to \alpha)}}.$$

As proposed in [WN11], we will confine ourselves with grouping the lengths together in several intervals which allows us to store the needed transition probabilities as a vector and thus makes it possible to retrieve them in algorithms and applications without further computational efforts. Obviously, the needed emission probabilities for unpaired bases and base pairs, respectively, can be stored and retrieved in the same way. As we will see later, when choosing appropriate intervals, this restriction is – under certain circumstances – still powerful enough to yield a significant improvement over traditional SCFGs with respect to the prediction accuracy of statistical sampling and the quality of the generated sample sets. However, if lengths are to be grouped into intervals, we have to deal with the fact that not all such groupings yield a consistent grammar. Nevertheless, the following definition of consistency offers a sufficient condition that they do (see [WN11] for details).

Definition 2.4 ([WN11])**.** Let $G = (I, T, R, S)$ a CFG and Q a partitioning of $\mathbb{N}$. We call Q *consistent with* G if it satisfies $\forall q \in Q,\ i, j \in q : \exists A \to \alpha \in R,\ w_i \in T^i : \alpha \Rightarrow^* w_i \curvearrowright \exists w_j \in T^j : \alpha \Rightarrow^* w_j$.

This means to satisfy that condition, we may not group lengths together for which different subsets of the considered rule set can eventually yield a terminal word of the respective length; the partitioning into sets of one element each (which corresponds to not grouping lengths into intervals) is trivially always consistent.

It should be clear that the danger of overfitting of the induced model is much more present if length-dependent probabilities rather than their traditional length-independent counterparts are considered, since then the observations made for a particular structural motif have to be splitted into distinct subsets of observations according to the corresponding lengths (or length intervals). In fact, LSCFGs typically imply significantly greater numbers of free parameters than the corresponding conventional SCFGs, where the actual numbers indeed increase with growing complexity (by means of number of distinguished length intervals) of the considered partitioning of $\mathbb{N}$.

2.3 Finding Appropriate Length Intervals

In order to find appropriate length intervals for applications based on grammar $\mathcal{G}_s$, we first partition the productions in $R_{\mathcal{G}_s}$ into subsets, where each subset generates terminal words of different lengths. Let $m_p := (2 \cdot m_s + m_h)$ denote the minimum allowed size of a paired substructure. Furthermore, recall that in this paper, we will only consider the common choices of $m_h = \min_{HL} \in \{1, 3\}$ and $m_s = \min_{\text{hel}} \in \{1, 2\}$. Then, we obtain the following:

Lengths		Rules that can actually produce a terminal word of these lengths		
$= 0$		$p_{28} : U \rightarrow \epsilon,$		
≥ 0		$p_{26} : N \rightarrow U,$		
$= 1$		$p_8 : C \rightarrow Z,$	$p_{17} : H \rightarrow Z,$	$p_{22} : B \rightarrow Z,$
		$p_{29} : Z \rightarrow \circ,$		
≥ 1		$p_1 : S \rightarrow T,$	$p_2 : T \rightarrow C,$	$p_{27} : U \rightarrow ZU,$
≥ 2		$p_7 : C \rightarrow ZC,$	$p_{16} : H \rightarrow ZH,$	$p_{21} : B \rightarrow ZB,$
$\geq m_h$	$\in [1; 3]$	$p_{11} : L \rightarrow F,$	$p_{15} : F \rightarrow Z^{m_h - 1}H,$	
$\geq m_h + 2$	$\in [3; 5]$	$p_{10} : P \rightarrow \textbf{(}L\textbf{)},$	$p_{12} : L \rightarrow P,$	
$\geq m_p$	$\in [3; 7]$	$p_3 : T \rightarrow A,$	$p_9 : A \rightarrow \textbf{(}^{m_s}L\textbf{)}^{m_s},$	
		$p_{24} : O \rightarrow UAN,$	$p_{25} : N \rightarrow UAN,$	
$\geq m_p + 1$	$\in [4; 8]$	$p_4 : T \rightarrow CA,$	$p_5 : T \rightarrow AT,$	
		$p_{13} : L \rightarrow G,$	$p_{18} : G \rightarrow BA,$	$p_{19} : G \rightarrow AB,$
$\geq m_p + 2$	$\in [5; 9]$	$p_6 : T \rightarrow CAT,$	$p_{20} : G \rightarrow BAB,$	
$\geq 2 \cdot m_p$	$\in [6; 14]$	$p_{14} : L \rightarrow M,$	$p_{23} : M \rightarrow UAO.$	

Consequently, a partitioning Q of $\mathbb{N}$ that is consistent with $\mathcal{G}_s$ (according to Definition 2.4) can be given as follows:

$$\begin{aligned} Q = &\{[0; 0], [1; 1], [2; \max(2, m_h - 1)]\} \cup \\ &\{[m_h; m_h], [m_h + 1; m_h + 1], [m_h + 2; \max(m_h + 2, m_p - 1)]\} \cup \\ &\{[m_p; m_p], [m_p + 1; m_p + 1], [m_p + 2; 2 \cdot m_p - 1], [2 \cdot m_p; \infty]\}. \end{aligned}$$

However, for the sake of simplicity, it would be more convenient to consider only a grouping of lengths into intervals that is appropriate for all our different structural parameter choices. In order to keep the estimated probabilities accurate, we decided to make the intervals longer as the considered subwords get longer, for the following two reasons: First, since typically any training set contains fewer data points per length as the length gets longer and second, since the influence a change in length has on the probabilities of productions most likely depends on the relative change rather than the absolute one. Therefore, we decided to use the successively increasing intervals $[i; i]$ for $0 \leq i \leq 40$, $[i; i+1]$ for $41 \leq i \leq 59$, $[i; i+2]$ for $61 \leq i \leq 82$, $[i; i+3]$ for $85 \leq i \leq 97$, $[i; i+4]$ for $101 \leq i \leq 136$, $[i; i+9]$ for $141 \leq i \leq 191$ and $[i; i+19]$ for $201 \leq i \leq 281$, together with the longer intervals $[301; 330]$, $[331; 360]$, $[361; 390]$, $[391; 430]$, $[431; 470]$, $[471; 510]$, $[511; 560]$, $[561; 610]$, $[611; 670]$, $[671; 730]$, $[731; 800]$, $[801; 900]$, $[901; 1000]$, and $[1001; \infty]$[5]. Obviously, this partitioning of $\mathbb{N}$ (into 105 distinct length intervals) is consistent with $\mathcal{G}_s$ for all considered structural parameter choices and thus in any case yields a consistent LSCFG $\mathcal{G}_s$.

Finally, note that under the common assumption that all emissions come from the same distribution, there are accordingly at most $(\text{card}(\mathcal{R}_{\mathcal{G}_s}) - \text{card}(\mathcal{I}_{\mathcal{G}_s})) \cdot 105 + \text{card}(\Sigma_{\mathcal{G}_r}) \cdot 1 + \text{card}(\Sigma_{\mathcal{G}_r})^2 \cdot 105 = (29 - 15) \cdot 105 + 4 + 16 \cdot 105 = 1470 + 4 + 1680 = 3154$ free parameters that need to be estimated for the LSCFG $\mathcal{G}_s$ when

[5] Note that these are basically the same intervals as used in [WN11] for the length-dependent grammars and have thus proven convenient.

considering these 105 appropriate length intervals. This number is obviously indeed to a large extend greater than the corresponding parameter number (of at most 34) implied in case of the conventional SCFG $\mathcal{G}_s$. However, it should be mentioned that the actual number of relevant (i.e., being greater than 0) free parameters will usually be much smaller, since a potentially significant amount of the length-dependent probabilities will inevitably always be equal to zero (independent on the used training data). This is due to the partitioning of data points according to different lengths and the constraints imposed by the structural parameters $\min_{\text{hel}}$ and $\min_{HL}$. For instance, as regards multiloops, we might only obtain $\Pr(L \to M, l) \neq 0$ for $l \geq 2 \cdot m_p$, whereas for $l < 2 \cdot m_p$, $\Pr(L \to M, l) = 0$ must always hold.

3 Algorithm

In this section, we will describe how to modify the routines and formal definitions proposed in [NSar] in order to obtain a corresponding statistical sampling method for RNA secondary structures according to the length-dependent SCFG model defined in the last section. Therefore, recall that in accordance with the popular PF variant presented in [DL03], the SCFG based sampling method has two basic steps. Its first step (preprocessing) computes the inside and outside probabilities for all substrings of an RNA sequence based on the considered SCFG. In the second step (structure sampling), base pairs (and unpaired bases) are randomly drawn according to the conditional sampling probabilities for the considered fragment (that are calculated by using only the inside and outside values derived in step one and the probabilities of the grammar rules) in order to sample complete secondary structures.

3.1 Computation of Inside and Outside Probabilities

In [NSar], all inside and outside probabilities have been computed based on an Earley-style parser[6]. Notably, if grammar parameters are separated into transitions and emissions, then probabilistic Earley parsing can easily be applied to work for all SCFGs (length-dependent or not) by a few simple modifications of the corresponding subroutines. Basically, instead of considering both the transition and emission probabilities already in the initial prediction steps, one has to multiply in the right rule probabilities (multiplied by corresponding factors) in the completion steps and the corresponding emission probabilities in the scanner steps, respectively. Under the assumption that the lengths are grouped together in several intervals, these modifications do not influence the run-time significantly (we only need an additional parameter for probability lookup).

A formal and more detailed description on how the inside and outside variables for a given input sequence can be computed with a special variant of an Earley-style parser based on the (L)SCFG $\mathcal{G}_s$ can be found in Section Sm-I[7].

3.2 Computation of Sampling Probabilities and Structure Sampling

Furthermore, note that the equations defining the needed sampling probabilities for all considered cases as presented in [NSar] depend not only on inside and outside values for the given RNA sequence, but also on probabilities $\Pr(\mathit{rule})$ of production rules $\mathit{rule} \in \mathcal{R}_{\mathcal{G}_s}$ of the underlying SCFG. Thus, in order to obtain the respective sampling probabilities based on the corresponding LSCFG model, besides computing the inside and outside probabilities in a slightly different way as described previously, we additionally have to consider length-dependent rule probabilities (multiplied by corresponding factors) instead of their traditional length-independent counterparts in the respective definitions.

The corresponding sampling algorithm and the use of the diverse sampling probabilities (derived length-dependently or conventionally) remain the same as proposed in [NSar]. Thus, all in one, by using the LSCFG approach instead of the corresponding length-independent variant, we can produce a statistical sample of the complete ensemble of all possible structures for a given sequence without significant losses in performance. However, when comparing the results of both SCFG methods, significant differences can be observed, as we will see in the next section.

[6] The authors actually relied on the formalism presented in [Goo98, Goo99] for describing parsers which is called *semiring parsing* for the inside outside calculations, as this approach also works for SCFGs like $\mathcal{G}_s$ that are not in *Chomsky normal form (CNF)*.

[7] All references starting with Sm are references to the supplementary material to this paper, available at `http:///wwwagak.cs.uni-kl.de/publications/`.

4 Applications and Discussion

The purpose of this section is to explore the benefits and potential drawbacks of enriching the sophisticated SCFG design introduced in [NSar] with additional information on the lengths of generated subwords (corresponding to particular RNA substructures).
In principle, the main question is to what extend are the sampling quality and the predictive power of the corresponding sampling variants affected by relying on the more elaborate LSCFG model (with a resulting comparatively huge number of more specific parameters) instead of on the conventional SCFG model (that implies only a rather moderate parameter number). It would also be interesting to see how much the performances of the traditional and the length-dependent SCFG variant (with the more generalized and specialized transition and emission probabilities, respectively) differ from that of the popular PF approach (that employs many hundreds of mostly experimentally obtained thermodynamic parameters). Therefore, we decided to consider a number of meaningful applications in connection with sampling approaches to the generated samples and for any of those applications oppose the results obtained with the proposed LSCFG based sampling approach to corresponding outputs of the simple SCFG variant from [NSar] and the PF method as described in [DL03][8].

4.1 Considered RNA Data and Probabilistic Parameters

In order to obtain an adequate basis for the investigations that will be performed within this section, we took the same sets of real world RNA data as were used for the corresponding applications in [NSar]: First, a (very rich) tRNA database (of 2163 distinct structures with lengths in $[64, 93]$) obtained from [SHB$^+$98]. Second, a (not quite so rich) 5S rRNA database (of 1149 distinct sequences with lengths in $[102, 135]$) retrieved from [SBEB02]. And last but not least, a (rather sparse) mixed structural database (of 151 distinct RNA molecules with lengths in $[23, 568]$) as collected in [DWB06]. Note that the latter will be denoted by *S-151Rfam database* in the sequel and is ought to illustrate quality differences of the corresponding results compared to the rich (and pure) tRNA and 5S rRNA data sets.

Training data	Model	Structural Constraints	num_{tr}	num_{em}^{unp}	num_{em}^{bp}
tRNA	SCFG	$\min_{HL} \in \{1,3\}, \min_{\text{hel}} = 1$	28	4	15
		$\min_{HL} \in \{1,3\}, \min_{\text{hel}} = 2$	27	4	14
	LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	334	4	162
		$\min_{HL} = 1, \min_{\text{hel}} = 2$	281	4	155
		$\min_{HL} = 3, \min_{\text{hel}} = 1$	332	4	162
		$\min_{HL} = 3, \min_{\text{hel}} = 2$	279	4	155
5S rRNA	SCFG	$\min_{HL} \in \{1,3\}, \min_{\text{hel}} \in \{1,2\}$	28	4	16
	LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	392	4	572
		$\min_{HL} = 1, \min_{\text{hel}} = 2$	357	4	565
		$\min_{HL} = 3, \min_{\text{hel}} = 1$	390	4	572
		$\min_{HL} = 3, \min_{\text{hel}} = 2$	355	4	565
S-151Rfam	SCFG	$\min_{HL} \in \{1,3\}, \min_{\text{hel}} \in \{1,2\}$	29	4	6
	LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	1171	4	477
		$\min_{HL} = 1, \min_{\text{hel}} = 2$	1055	4	446
		$\min_{HL} = 3, \min_{\text{hel}} = 1$	1155	4	470
		$\min_{HL} = 3, \min_{\text{hel}} = 2$	1022	4	434

Table 1: Numbers of relevant parameters (transition and emission probabilities being greater than zero) that are obtained from training the respective database (in the traditional or length-dependent way). Here, num_{tr} denotes the number of relevant transition probabilities. Accordingly, num_{em}^{unp} and num_{em}^{bp} denote the numbers of relevant emission probabilities of unpaired bases and base pairs, respectively.

As already mentioned, we implemented length-dependency by grouping the lengths into distinct intervals, such that the probabilities change only from one interval to the other but not within them. In fact, we used the 105 reasonable intervals[9] presented at the end of Section 2. Table 1 shows that quite different numbers

[8] It should be mentioned that for our examinations, we have implemented our own version of Sfold's sampling procedure. For this implementation, we decided to use the thermodynamic parameters from Mathews et al. [MSZT99], which were also used for version 3.0 of the Mfold software [Zuk03]

[9] Note that since the interval lengths grow with increasing subword length, we can hope for accurate estimated probabilities. Furthermore, as all molecules in the considered benchmark sets are shorter than 1000 nucleotides, the probabilities of the last interval do not influence our results.

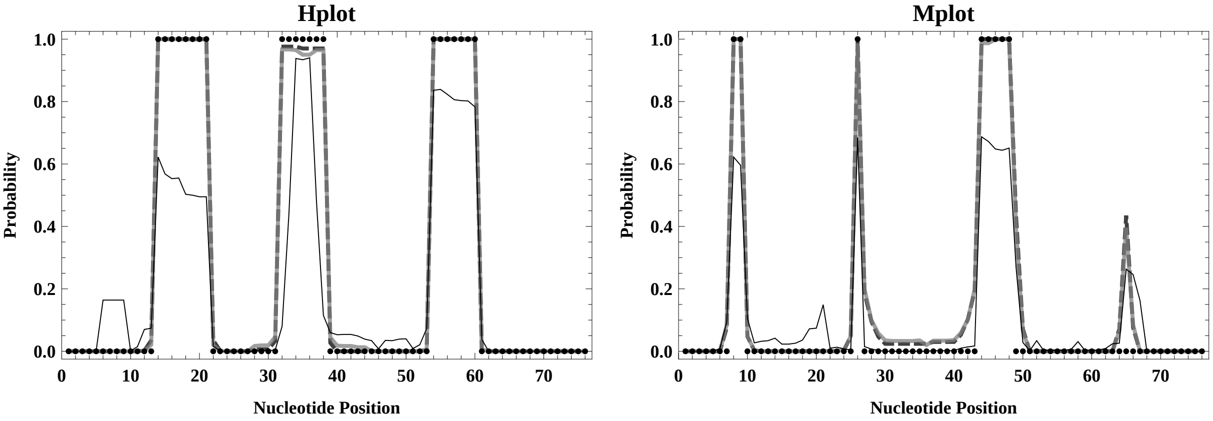


Figure 1: Loop profiles for *E.coli* tRNA^{Ala}, obtained with the PF approach and the length-dependent SCFG variant. Hplot and Mplot display the probability that an unpaired base lies in a hairpin and multibranched loop, respectively. Results for the PF approach (for $\max_{BL} = 30$) are displayed by the thin black lines. For the SCFG approach, we chose $\min_{\text{hel}} = 1$ (thick gray lines) and $\min_{\text{hel}} = 2$ (thick dashed darker gray lines), combined with $\min_{HL} = 3$, respectively. The corresponding probabilities for the correct structure of *E.coli* tRNA^{Ala} are also displayed (by black points).

of relevant length-dependent (rule and emission) probabilities result when training our grammar on the three different data sets, respectively. However, although the numbers of relevant grammar parameters are unsurprisingly to a large extend greater when considering the LSCFG model rather than the traditional length-independent variant, they are indeed of a considerably smaller (in case of tRNAs and 5S rRNAs) or at least only of a similar (in case of the S-151Rfam data) order of magnitude than the numbers of energy parameters employed in standard thermodynamic models. Hence, if statistical parameter learning makes sense in connection with free energy approaches (which it obviously does, since it has become increasingly appreciated), then it should also yield reasonable results in case of LSCFG based probabilistic methods. Motivated by this assumption, we decided to start our examinations in the next section by considering one of the most intuitive applications in connection with statistical sampling methods that is of great practical interest.

4.2 Probability Profiling for Specific Loop Types

As starting point for our examinations, a statistical sample of all possible secondary structures for a given RNA sequence shall be used for sampling estimates of the probabilities of any structural motifs. In particular, we will consider *probability profiles* of unpaired bases in each specific loop type of RNA secondary structure. This means for each nucleotide position i, $1 \leq i \leq n$, of a given RNA sequence of length n, we compute the probabilities that i is an unpaired base within a specific loop type; these probabilities are given by the observed frequency in a sample set of secondary structures for the given sequence.

To compare the sampling results obtained with the presented LSCFG approach to those for the PF variant, we decided to consider the corresponding probability profiles for *Escherichia coli* tRNA^{Ala}, which are shown in Figure 3 of Section Sm-II. The probably most interesting ones are also displayed in Figure 1 which perfectly exhibit the cloverleaf structure of tRNAs, enhancing the corresponding profiles for the length-independent SCFG variant (these are presented in Figure 2 and Figure 4 of Section Sm-II for convenience).

It is obvious that the statistical samples generated by the (L)SCFG approach are significantly more accurate than those obtained with the PFs. Furthermore, comparing the plots in Figure 1 to those in Figure 2 that were computed without considering length-dependency, we see that the sampling results can indeed be improved by incorporating additional length information into the underlying SCFG model; the correct cloverleaf structure of the considered tRNA is almost exactly reached in all sampled structures.

Nevertheless, before we proceed with applications of the considered sampling approaches to RNA structure prediction, we first want to discuss some important results with respect to the quality of particular probabilistic structure models (induced by the three considered RNA classes, respectively) underlying the proposed LSCFG based sampling method.

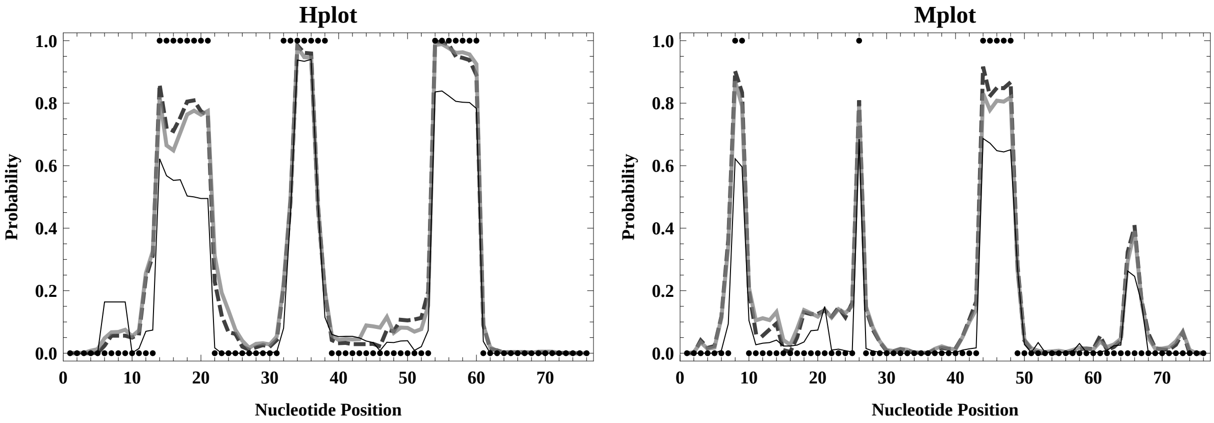


Figure 2: Loop profiles for *E.coli* tRNA^{Ala} corresponding to those presented in Figure 1, obtained with the PF approach and the traditional SCFG variant that does not incorporate length-dependencies.

4.3 The Problem of Overfitting and the Lack of Generalization

Analogous to [NSar], we will address two possible issues of our sophisticated LSCFG in connection with this study: the problem of overfitting and the lack of generalization. With respect to the latter, it seems important to mention that the profiling results for *Escherichia coli* tRNA^{Ala} eventually validate two obvious assumptions: First, if our LSCFG is trained on trusted tRNAs only, it should inevitably produce the typical tRNA cloverleaf shape more often than the alternative PF variant that is not suited to a specific class of RNA structures. Second, as the additional consideration of length-dependencies yields more specialized probabilities for the distinct structural motifs and tRNA molecules naturally show a low structural variety, the LSCFG based profiles should inherently show the cloverleaf structure more explicitly.

Consequently, it is likely that the higher accuracy reached by the LSCFG model could be an artefact caused by lack of generalization of the underlying stochastic structure model. To get evidence of the correctness of this assumption, we took the random sequence sets from [NSar][10] and applied the different sampling approaches. The results are collected in Table 2.

Undoubtably, the presented statistics demonstrate that the LSCFG variant mainly samples cloverleaf structures, even if the signal towards cloverleaf is low or does actually not exist ($\min_{\text{hel}} = 0$ means completely random sequences, that is no signal). This yields the assumption that incorporating length-dependencies into the underlying sophisticated SCFG model in fact causes lack of generalization, as the cloverleaf shape is always preferred over others, regardless of the signal induced by the actual sequence composition. Hence, there is some reason to believe that the accuracy gain of the LSCFG sampling approach (at least for tRNA profiling) is due to the high degree of specialization of the underlying stochastic structure model (very explicitly tailored to a certain shape), which bares an undesirable lack of generalization (to possible but usually less likely other shapes). Since we most likely observe such effects in connection with tRNA and its invariant cloverleaf shape, we skipped similar investigations for the other cases.

Nevertheless, in order to investigate if overfitting may be a problem for the subsequent examinations, i.e. to see if our different data sets are rich enough to reliably derive the parameters of our grammar even in the length-dependent case, we performed the following experiments (similar to [NSar]): For each of the three considered structural RNA databases, we selected a random 90% portion of the original database (such that the resulting sample size equals that of training sets used for 10-fold cross-validation experiments which will be intensively performed in the sequel) and re-estimated the probabilities of all grammar rules (for any of the previously chosen length intervals, respectively). Since the number of feasible structures that can be considered for training is reduced by prohibiting small hairpin loops and isolated base pairs, we decided to rely on the most realistic restrictions of $\min_{\text{hel}} = 2$ and $\min_{HL} = 3$ for our SCFG $\mathcal{G}_s$ in order to obtain the potentially most meaningful results.

The corresponding re-estimation process was iterated 100 times for any database, resulting in a sample of 100 parameter sets, respectively, each of them consisting of exactly $\text{card}(\mathcal{R}_{\mathcal{G}_s})$ sets of length-dependent

[10] For any fixed value of $\min_{\text{hel}}$, the corresponding set has been generated by randomly creating secondary structures (with corresponding sequences) having the cloverleaf shape, where all four helices (the stem and the three adjacent helices of the multiloop) are formed by exactly $\min_{\text{hel}}$ consecutive (canonical) base pairs.

Approach	$\mathrm{min}_{\mathrm{hel}}$	$\mathrm{num}_{\mathrm{d}}$	c_{d}	c_{MF}	c_{CL}	$\mathrm{num}_{\mathrm{MF}}$
PF	0	36	8333.33	94085	3331	6
	1	34	8823.53	87785	5338	6
	2	35	8571.43	96083	2745	6
	3	37	8108.11	95332	4492	6
	4	30	10000.	107881	9967	6
	5	29	10344.8	111716	20875	3
	6	33	9090.91	102788	49733	2
	7	27	11111.1	94859	94859	0
SCFG	0	858	349.65	26341	14114	5
	1	916	327.511	22643	15596	4
	2	915	327.869	21258	13912	4
	3	895	335.196	20175	16207	2
	4	914	328.228	19828	17784	2
	5	844	355.45	20560	20560	0
	6	747	401.606	34753	34753	0
	7	658	455.927	59644	59644	0
LSCFG	0	28	10714.3	92727	92727	0
	1	28	10714.3	91276	91276	0
	2	25	12000.	88660	88660	0
	3	27	11111.1	94323	94323	0
	4	27	11111.1	94536	94536	0
	5	27	11111.1	100720	100720	0
	6	27	11111.1	107157	107157	0
	7	26	11538.5	115788	115788	0

Table 2: Results derived from random data sets, where $\mathrm{min}_{\mathrm{hel}}$ (defining the minimum allowed length of helical regions) has been used as structural constraint for the generation of random sequences with corresponding (more or less strong) signals towards a cloverleaf structure. $\mathrm{num}_{\mathrm{d}}$ denotes the number of distinct shapes (here, *abstract shapes* of level 5 according to [JRG08]) in all samples and c_{d} the average count of one of these distinct shapes. Furthermore, c_{MF} and c_{CL} represent the count of the most frequent and cloverleaf shape in all samples, whereas $\mathrm{num}_{\mathrm{MF}}$ denotes the number of distinct shapes that are observed more frequently than the cloverleaf. For any setting of $\mathrm{min}_{\mathrm{hel}}$, all tabulated values were computed from a corresponding random data set of cardinality 300 (containing 10 random sequences for any length $n \in \{64, \ldots, 93\}$ according to the length range observed from our tRNA database), respectively. A sample size of 1000 structures and $\max_{BL} = 30$ has been chosen for either approach.

probabilities $p_i(I)$ for the distinct length intervals I rather than of one single (conventional, i.e. length-independent) probability value p_i, $1 \leq i \leq \mathrm{card}(\mathcal{R}_{\mathcal{G}_s})$. Therefore, for each of the distinguished length-dependent grammar parameters $p_i(I)$, we determined its variance along the constructed sample of size 100 and subsequently computed the maximum variance (observed for a particular length interval I) among all variances $\mathbb{V}[p_i(I)]$ implied by production rule f_i, for $1 \leq i \leq \mathrm{card}(\mathcal{R}_{\mathcal{G}_s})$. Formally, for each set of length-dependent probabilities corresponding to grammar parameter p_i, $1 \leq i \leq \mathrm{card}(\mathcal{R}_{\mathcal{G}_s})$, we calculated $\max_I \mathbb{V}[p_i(I)]$. The resulting values are collected in Table 3.

Note that the variances 0 in most cases result for production rules finishing the generation of unpaired regions (for example $p_8 : C \to Z$ or $p_{28} : U \to \epsilon$), since those can only produce words of one particular length (1 or 0), whereas longer words (unpaired regions) are generated by the corresponding alternative productions with same left-hand side (for example $p_7 : C \to ZC$ or $p_{27} : U \to ZU$), and the weights on the production rules must indeed sum up to unity for any considered length interval. Thus, since we use unary intervals for lengths 0 and 1, respectively, for any production ending a run of unpaired bases, a probability of 1 is predetermined, yielding variance 0. For basically the same reason, there must result a variance of 0 for production $p_{29} : Z \to \circ$, i.e. this observation is due to the fact that this rule unexceptionally generates words of length 1 (an arbitrary unpaired base) and there exist no other alternatives for the corresponding premise implying words of that particular length.

However, all the other (maximum) variances presented in Table 3 (at least for tRNAs and 5S rRNAs) are rather small, too. Therefore, we may assume that overfitting is not really an issue in connection with our sophisticated SCFG and the training sets used (at least for the rich tRNA and 5S rRNA data), even in the case of length-dependent parameter estimation procedures. It remains to mention that the tabulated

$\max_I \mathbb{V}[\cdot(I)]$	tRNA	5S rRNA	S-151Rfam
p_1	1.1372×10^{-4}	4.6414×10^{-5}	2.1343×10^{-2}
p_2	7.1533×10^{-5}	7.0953×10^{-5}	4.8821×10^{-3}
p_3	7.0888×10^{-7}	2.4405×10^{-5}	1.8768×10^{-3}
p_4	2.0229×10^{-7}	7.7689×10^{-6}	1.0272×10^{-3}
p_5	3.5269×10^{-5}	2.5606×10^{-5}	2.4133×10^{-3}
p_6	4.9101×10^{-6}	6.4274×10^{-6}	4.9589×10^{-3}
p_7	1.1616×10^{-5}	3.0681×10^{-5}	1.7759×10^{-4}
p_8	0	0	0
p_9	4.9153×10^{-6}	3.5895×10^{-5}	6.0461×10^{-3}
p_{10}	5.5523×10^{-6}	1.5978×10^{-5}	2.9525×10^{-3}
p_{11}	2.6480×10^{-6}	5.7427×10^{-6}	8.8191×10^{-4}
p_{12}	6.1203×10^{-6}	1.5467×10^{-5}	1.7275×10^{-3}
p_{13}	1.6234×10^{-7}	6.3548×10^{-6}	3.1334×10^{-4}
p_{14}	2.9152×10^{-6}	3.2344×10^{-6}	6.5392×10^{-5}
p_{15}	3.1928×10^{-6}	1.0465×10^{-4}	2.0547×10^{-3}
p_{16}	1.9113×10^{-6}	6.4819×10^{-6}	1.1604×10^{-4}
p_{17}	0	0	0
p_{18}	0	1.0346×10^{-4}	1.6601×10^{-3}
p_{19}	0	8.9041×10^{-5}	2.0498×10^{-3}
p_{20}	1.8388×10^{-3}	1.1285×10^{-4}	9.3347×10^{-3}
p_{21}	4.1771×10^{-5}	6.9182×10^{-7}	7.3819×10^{-5}
p_{22}	0	0	0
p_{23}	9.5068×10^{-5}	4.1479×10^{-5}	3.6034×10^{-2}
p_{24}	5.1666×10^{-5}	6.1313×10^{-4}	5.1346×10^{-2}
p_{25}	1.6458×10^{-5}	0	1.7848×10^{-3}
p_{26}	1.2797×10^{-6}	1.7441×10^{-4}	1.6096×10^{-2}
p_{27}	8.0792×10^{-7}	4.0028×10^{-6}	6.0669×10^{-4}
p_{28}	0	0	0
p_{29}	0	0	0

Table 3: Truncated maximum variances of any set of grammar parameters (transition probabilities) for different length intervals, derived from 100 iterations of (length-dependently) training our SCFG $\mathcal{G}_s$ on random subsets containing 90 percent of the original data, respectively, under the assumption of $\min_{HL} = 3$ and $\min_{\text{hel}} = 2$.

(maximum) variances derived for the considered length-dependent grammar parameters are in most cases indeed larger than the corresponding variances for the conventional parameters which do not depend on the lengths of generated subwords (as can be observed by comparing Table 3 to Table 13 of Section Sm-II). This to some extend proves the fact that length-dependent training procedures generally require richer training sets in order to estimate the grammar parameters (for any considered length interval) as reliably as in the traditional length-independent case.

4.4 Prediction Accuracy – Sensitivity and PPV

In order to investigate how the quality of predictions changes when using the (length-dependent) SCFG approach for computing the sampling probabilities, we decided to consider the common accuracy measures *sensitivity* and *positive predictive value*[11]. In the context of RNA secondary structure prediction, they are usually defined as follows (see e.g. [BBC+00]):

- sensitivity (Sens.) is the relative frequency of correctly predicted pairs among all position pairs that are actually paired in a stem of native foldings, whereas
- the positive predictive value (PPV) is defined as the relative frequency of correctly predicted pairs among all position pairs that were predicted to be paired with each other.

Formally, these measures are given by Sens. $= TP \cdot (TP + FN)^{-1}$ and PPV $= TP \cdot (TP + FP)^{-1}$, where TP is the number of correctly predicted base pairs (*true positives*), FN is the number of base pairs in the native structure that were not predicted (*false negatives*) and FP is the number of incorrectly predicted base pairs (*false positives*).

For assessing the differences in the predictive accuracy of sample sets generated according to either approach, we decided to perform a suitable k-fold cross-validation for any of our three different RNA databases. Actually, we used the same partitions of the comprehensive tRNA and 5S rRNA databases into

[11] Note that the positive predictive value is often called *specificity*, although this measure formally obeys to a slightly different definition

$k = 10$ and of the mixed S-151Rfam database into $k = 2$ approximately equal-sized folds as in [NSar] and derived the corresponding k-fold cross-validations results, respectively. In particular, for any sequence, we sampled a set of 1000 structures and then applied different principles to obtain a corresponding structure prediction:

First, we assumed that the prediction for a given sequence is equal to the most frequently sampled secondary structure, which will be called *most frequent (MF)* structure in the sequel. This is convenient, since the sampling algorithm produces a statistically representative sample of secondary structures for a given RNA sequence, and thus, if the sample size is large enough, the most frequently sampled structure can be assumed to be highly probable among all structures for this sequence. For this reason, given the case that there is more than one most frequently sampled structure, we always chose one with the highest probability (according to the probability distribution implied by the respective approach).

As an alternative, we decided to additionally consider a *maximum expected accuracy (MEA)* structure of the generated sample set as prediction. Basically, the MEA structures for a given sequence are the ones among all candidate structures that maximize the number of correctly unpaired and paired positions with respect to the true folding of that sequence. However, contrary to this traditional definition (as employed in common DP approaches for predicting a single folding, like for instance Pfold [KH03] and CONTRAfold [DWB06]), we rely on the slightly modified version as proposed in [NSar], where the base pairing probabilities $p_{i,j}$ reflect the distribution observed in the sample set and not the distribution implied by the complete structure ensemble for the given input sequence.

Finally, we took the unique *centroid* structure of the generated sample set as predicted folding. Briefly, a centroid is defined as the single structure in the entire ensemble that has the minimum total base-pair distance to all other structures and thus best represents the central tendency of the structure set. This choice can thus be seen as purposive for sampling approaches like the ones opposed in this study, as the centroid reflects the overall behavior of the structures in a given sample set.

Approach	Parameters	MF struct.		MEA struct.		Centroid	
		Sens.	PPV	Sens.	PPV	Sens.	PPV
PF	$\max_{BL} = 30$	0.6565	0.5890	0.6434	0.6035	0.6159	0.6344
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.7791	0.8445	0.7324	0.8939	0.6754	0.9158
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.8004	0.8457	0.7685	0.8878	0.7113	0.9123
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.8545	0.8517	0.7848	0.9021	0.7304	0.9213
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.8677	0.8593	0.8182	0.8953	0.7713	0.9168
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.8542	0.9535	0.8335	0.9736	0.8250	0.9783
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.8530	0.9502	0.8518	0.9613	0.8435	0.9657
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.8602	0.9526	0.8371	0.9733	0.8278	0.9775
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.8575	0.9494	0.8562	0.9609	0.8477	0.9651

Table 4: Sensitivity and PPV values for our tRNA database (computed by 10-fold cross-validation procedures, using sample size 1000).

Approach	Parameters	MF struct.		MEA struct.		Centroid	
		Sens.	PPV	Sens.	PPV	Sens.	PPV
PF	$\max_{BL} = 30$	0.5897	0.5806	0.6015	0.6191	0.5789	0.6508
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.4251	0.5362	0.3403	0.6967	0.2689	0.8044
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.4542	0.5435	0.3638	0.6901	0.2727	0.8069
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.4728	0.5290	0.3544	0.7033	0.2764	0.8091
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.5167	0.5577	0.3860	0.7010	0.2846	0.8140
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.8996	0.9408	0.8959	0.9513	0.8873	0.9574
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.8726	0.9239	0.8714	0.9280	0.8673	0.9333
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.8992	0.9405	0.8958	0.9509	0.8863	0.9568
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.8721	0.9231	0.8712	0.9276	0.8667	0.9330

Table 5: Sensitivity and PPV values for our 5S rRNA database (computed by 10-fold cross-validation procedures, using sample size 1000).

All corresponding sensitivity and PPV measures are collected in Tables 4, 5 and 6. Obviously, the results for tRNAs and 5S rRNAs lead to the conclusion that by using the LSCFG approach for statistical sampling, a significantly higher predictive accuracy can be reached than by sampling based on PFs.

Approach	Parameters	MF struct.		MEA struct.		Centroid	
		Sens.	PPV	Sens.	PPV	Sens.	PPV
PF	$\max_{BL} = 30$	0.6652	0.5188	0.6633	0.5450	0.6437	0.5799
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.4433	0.5447	0.3815	0.7386	0.3235	0.7749
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.4894	0.5551	0.4263	0.7181	0.3474	0.7743
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.4852	0.5948	0.3935	0.7426	0.3352	0.7825
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.5171	0.5661	0.4342	0.7228	0.3588	0.7683
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.1815	0.5422	0.1390	0.7523	0.1251	0.8003
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.1646	0.5322	0.1276	0.6706	0.1099	0.7114
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.1761	0.5354	0.1396	0.7614	0.1238	0.8039
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.1528	0.5023	0.1230	0.6634	0.1094	0.7118

Table 6: Sensitivity and PPV values for the mixed S-151Rfam database (computed by two-fold cross-validation procedures, using the same folds as in [DWB06] and sample size 1000).

Moreover, we immediately observe that eventually the incorporation of length-dependencies into the SCFG approach can have a positive impact on the resulting prediction accuracy: Although the predictions for the longer and thus more variant 5S rRNAs are less accurate than for the shorter tRNAs when using the conventional SCFG approach, the consideration of length-dependent probabilities for the production rules (obtained by training them on real world data) makes the underlying SCFG model explicit enough to handle the larger variety of structure motifs and guarantees high quality prediction results.

However, as we expected, for the S-151Rfam database, the more specialized LSCFG approach yields the worst prediction results, whereas the highest accuracy for this mixed data set is reached with PF sampling that relies on thermodynamic parameters and is not suited for a particular RNA type. In fact, this observation is strongly related to the fact that the S-151Rfam data set is rather sparse and additionally contains structures that belong to distinct RNA types that obey to different structural properties, such that it can not be considered an optimal training basis. This problem is considerably increased by the partitioning of (the already rather few) data points according to the various interval lengths for our LSCFG variant (which is actually in accordance with the worse results for the S-151Rfam set compared to the rich and pure tRNA and 5S rRNA sets as presented in Table 3 of Section 4.3).

Altogether, we can assume that if a reasonable RNA secondary structure database (containing a sufficiently large number of known structures that are of the same or similar RNA types) can be used for estimating the parameters of the underlying LSCFG model, then even for RNA molecules with a high variability of typical structural features (for which the traditional SCFG method lacks the ability to identify the typical shape of the respective family by considering the estimated length-independent parameters), the predictive results might be of high quality and potentially manage to outperform predictions obtained with the PF variant that is based on the competing free energy approach.

Approach	Parameters	MEA struct.	Centroid
PF	$\max_{BL} = 30$	0.482435	0.526743
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.828522	0.833894
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.830787	0.839843
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.855406	0.861640
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.857251	0.867135
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.936285	0.919736
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.916900	0.910218
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.936337	0.920387
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.916641	0.910321

Table 7: AUC values for our tRNA database (computed by 10-fold cross-validation procedures, using sample size 1000).

All these observations may be affirmed on the basis of more reliable accuracy results which can be readily obtained by computing a collection of additional predictions from each of the generated sample sets along the following lines: According to [NSar] (inspired by [DWB06]), a trade-off parameter $\gamma_{\text{t-o}} \in [0, \infty)$ that manages to control the balance between the sensitivity and PPV of the predicted foldings can easily be incorporated into the procedures for calculating the MEA and centroid structures of a given sample set.

Approach	Parameters	MEA struct.	Centroid
PF	$\max_{BL} = 30$	0.481019	0.520171
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.409278	0.408549
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.417286	0.418584
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.419116	0.417095
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.433954	0.431642
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.914801	0.918933
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.854520	0.863009
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.914114	0.918600
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.853399	0.862744

Table 8: AUC values for our 5S rRNA database (computed by 10-fold cross-validation procedures, using sample size 1000).

Approach	Parameters	MEA struct.	Centroid
PF	$\max_{BL} = 30$	0.450688	0.497350
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.499491	0.507125
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.506602	0.509403
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.507454	0.512327
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.508762	0.514958
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.270606	0.269354
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.206630	0.208092
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.271388	0.266478
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.205790	0.209557

Table 9: AUC values for the mixed S-151Rfam database (computed by two-fold cross-validation procedures, using the same folds as in [DWB06] and sample size 1000).

The corresponding results are referred to as $\gamma_{\mathrm{t-o}}$-MEA and $\gamma_{\mathrm{t-o}}$-centroid structures, respectively, where the default choice $\gamma_{\mathrm{t-o}} = 1$ has no effect and thus yields the previously described conventional results. Hence, by allowing $\gamma_{\mathrm{t-o}}$ to vary, it effectively becomes possible to find corresponding *receiver operating characteristic (ROC)* curves for MEA and centroid predictions, yielding much more meaningful accuracy measures as the common sensitivity and PPV values for one particular choice of $\gamma_{\mathrm{t-o}}$.
In fact, for $\gamma_{\mathrm{t-o}} \in \{1.25^k \mid -12 \leq k \leq -1\} \cup \{2^k \mid 0 \leq k \leq 12\}$, the respective estimated *area under the curve (AUC)* values observed for any of the considered databases are reported in Tables 7, 8 and 9. Plots of some of the respective ROC curves can be found in Figures 5, 6 and 7 of Section Sm-II). As intended, the provided AUC values allow for a more reliable comparison of the accuracies that can be reached by either approach on the basis of MEA and centroid structures for the produced samples, respectively, but eventually yield basically the same conclusions.
Finally, it remains to mention that according to the definitions of sensitivity and PPV, these two accuracy measures depend only on the numbers of correctly and incorrectly predicted base pairs (compared to the native structure). For biologists, however, it is usually much more important to get the correct *shape* of the native folding than to obtain high sensitivity and PPV when using computational prediction methods. For this reason, a corresponding discussion will follow in the next section.

4.5 Sampling Quality – Specific Values Related to Shapes

The proclaimed aim of this section is to compare sampling results generated by the PF, SCFG and LSCFG approaches with respect to an abstraction level (shapes of generated structures) that is of great relevance for biologists. Particularly, we will consider a number of specific values related to the *abstract shapes* of sampled structures to obtain further proof of the high quality of sample sets generated by the proposed LSCFG approach.
Principally, abstract shapes (see e.g. [GVR04]) are morphic images of secondary structures, where each shape comprises a class of similar structures. There are five shape types for five different levels of abstraction, where the succeeding shape types are supposed to gradually increase abstraction by disregarding certain unpaired regions or combining nested helices. For the shape abstraction types as defined informally in [JRG08] (and for secondary structures as additional type 0 shapes), it has been proven that this is the case indeed [NS09].

In order to explore the sampling qualities that can be reached by the distinct sampling approaches, we decided to consider the following four specific values related to the shapes of sampled secondary structures:

- Frequency of prediction of correct shape (CSP_{freq}): In how many cases is the predicted shape (on different levels) equal to the correct one?
- Frequency of correct shape occurring in a sample (CSO_{freq}): In how many cases can the correct shape be found in the generated sample?
- Number of occurrences of correct shape in a sample (CS_{num}): How many times can the correct shape be found in the generated sample?
- Number of different shapes in a sample (DS_{num}): How many different shapes can be found in the generated sample?

Value	Approach	Parameters	Shape Level					
			0	1	2	3	4	5
CSP_{freq} (MF struct.)	PF	$\max_{BL} = 30$	0.0633	0.1216	0.2071	0.2117	0.2639	0.3694
	SCFG	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.2450	0.4448	0.6417	0.6417	0.6422	0.7356
	LSCFG	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.3440	0.5137	0.6805	0.6805	0.6810	0.7628
CSP_{freq} (MEA struct.)	PF	$\max_{BL} = 30$	0.0416	0.1049	0.1923	0.1960	0.2496	0.3559
	SCFG	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.1008	0.2917	0.5525	0.5525	0.5543	0.6241
	LSCFG	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.2053	0.4115	0.6958	0.6958	0.6963	0.7869
CSP_{freq} (Centroid)	PF	$\max_{BL} = 30$	0.0264	0.0800	0.1595	0.1627	0.1932	0.2677
	SCFG	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.0758	0.2150	0.4563	0.4563	0.4568	0.5003
	LSCFG	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.1956	0.3824	0.6426	0.6426	0.6431	0.7240
CSO_{freq}	PF	$\max_{BL} = 30$	0.5196	0.6740	0.8160	0.8239	0.8798	0.9556
	SCFG	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.7148	0.9459	0.9875	0.9880	0.9885	0.9991
	LSCFG	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.8391	0.9441	0.9778	0.9783	0.9783	0.9986
CS_{num}	PF	$\max_{BL} = 30$	21.073	58.200	136.67	140.63	205.54	328.56
	SCFG	$\min_{HL} = 3, \min_{\text{hel}} = 2$	34.898	173.73	513.05	513.06	513.08	595.26
	LSCFG	$\min_{HL} = 3, \min_{\text{hel}} = 2$	104.09	300.04	730.09	730.09	730.54	826.21
DS_{num}	PF	$\max_{BL} = 30$	355.32	130.22	81.796	33.125	22.585	4.8848
	SCFG	$\min_{HL} = 3, \min_{\text{hel}} = 2$	592.84	103.04	18.921	18.921	18.921	12.053
	LSCFG	$\min_{HL} = 3, \min_{\text{hel}} = 2$	126.84	8.2815	2.7296	2.7296	2.7296	2.3869

Table 10: Results related to the shapes of selected predictions and sampled structures, obtained from our tRNA database (by 10-fold cross-validation procedures, using sample size 1000).

The respective results are collected in Tables 14 to 19 in Section Sm-II. Some of the most interesting ones are also displayed in Tables 10, 11 and 12. Note that all these specific values have been calculated from the predicted structures and the corresponding sample sets that were derived for the calculation of the sensitivity and PPV measures in the last section.

As regards the considered tRNAs and 5S rRNAs, the predicted shape is in most cases significantly more often equal to the correct one when using the SCFG approach (length-dependent or not) instead of the PF variant. That is, the frequency of correct structure predictions (CSP_{freq}) is often higher when using the sophisticated SCFG instead of PFs, especially when length-dependence is considered. Moreover, the statistical samples generated with either of the two different SCFG approaches generally contain the correct shapes considerably more often than those obtained with the PF method, i.e. are more accurate as regards the frequency of correct structure occurrences (CSO_{freq}). Notably, again the best results are obtained with the LSCFG presented in this work (see Tables 10 and 11).

Furthermore, for tRNAs and 5Sr RNAs, the observed averaged number of correct shapes in a sample set (CS_{num}) is in all cases to a large extend greater when using the LSCFG approach than when using the length-independent variant or the PF method. However, due to these observations it is not surprising that the observed averaged number of different shapes in a sample (DS_{num}) is always significantly smaller when using the LSCFG approach rather than the PF and especially the traditional length-independent SCFG variant (for which the by far highest diversity within the sample set can be reached). This means by incorporating additional information on fragment lengths into the underlying sophisticated SCFG model, a higher predictive accuracy with respect to the shapes of generated structures (on all abstraction levels) can be reached, at the cost of a lower variability of the generated samples.

Value	Approach	Parameters	Shape Level					
			0	1	2	3	4	5
$\mathrm{CSP}_{\mathrm{freq}}$ (MF struct.)	PF	$\max_{BL} = 30$	0.0000	0.0009	0.0078	0.0513	0.0261	0.6353
	SCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0009	0.0096	0.0244	0.0609	0.1027	0.8207
	LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.2002	0.4239	0.4700	0.4857	0.9426	0.9861
$\mathrm{CSP}_{\mathrm{freq}}$ (MEA struct.)	PF	$\max_{BL} = 30$	0.0000	0.0052	0.0139	0.0835	0.0696	0.6640
	SCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0000	0.0009	0.0009	0.0035	0.0557	0.5387
	LSCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.1062	0.4065	0.4456	0.4535	0.8990	0.9835
$\mathrm{CSP}_{\mathrm{freq}}$ (Centroid)	PF	$\max_{BL} = 30$	0.0000	0.0026	0.0104	0.0775	0.0731	0.7214
	SCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0000	0.0000	0.0000	0.0009	0.0139	0.1549
	LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.0966	0.2916	0.3238	0.3316	0.8703	0.9686
$\mathrm{CSO}_{\mathrm{freq}}$	PF	$\max_{BL} = 30$	0.0009	0.1662	0.3063	0.7580	0.6883	0.9817
	SCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0026	0.4509	0.6372	0.9904	0.9974	0.9991
	LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.6258	0.8912	0.9295	0.9504	0.9948	1.0000
$\mathrm{CS}_{\mathrm{num}}$	PF	$\max_{BL} = 30$	0.0009	0.7571	3.4207	36.641	30.288	600.35
	SCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0026	1.3795	3.1949	36.673	71.080	609.58
	LSCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	42.962	347.97	422.19	457.71	875.13	983.67
$\mathrm{DS}_{\mathrm{num}}$	PF	$\max_{BL} = 30$	710.75	333.72	237.71	93.335	63.661	7.0951
	SCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	999.68	885.81	762.67	239.28	123.91	13.558
	LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	148.01	10.076	8.5355	4.4627	3.5160	1.1297

Table 11: Results related to the shapes of selected predictions and sampled structures, obtained from our 5S rRNA database (by 10-fold cross-validation procedures, using sample size 1000).

Value	Approach	Parameters	Shape Level					
			0	1	2	3	4	5
$\mathrm{CSP}_{\mathrm{freq}}$ (MF struct.)	PF	$\max_{BL} = 30$	0.0661	0.1255	0.1586	0.2050	0.2183	0.4834
	SCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0530	0.1258	0.1522	0.1788	0.1985	0.4240
	LSCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.0199	0.0532	0.0664	0.0730	0.0995	0.3179
$\mathrm{CSP}_{\mathrm{freq}}$ (MEA struct.)	PF	$\max_{BL} = 30$	0.0660	0.1123	0.1453	0.1984	0.2051	0.4902
	SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0264	0.1193	0.1391	0.1523	0.1789	0.4239
	LSCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.0132	0.0397	0.0530	0.0596	0.0794	0.2118
$\mathrm{CSP}_{\mathrm{freq}}$ (Centroid)	PF	$\max_{BL} = 30$	0.0793	0.1321	0.1653	0.1917	0.2449	0.5100
	SCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0197	0.0927	0.1125	0.1390	0.1391	0.3577
	LSCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.0066	0.0397	0.0530	0.0596	0.0728	0.1722
$\mathrm{CSO}_{\mathrm{freq}}$	PF	$\max_{BL} = 30$	0.3638	0.4433	0.4766	0.5231	0.6488	0.7947
	SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.2717	0.5630	0.6158	0.7284	0.8079	0.9605
	LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.0463	0.2518	0.4041	0.5496	0.5960	0.8408
$\mathrm{CS}_{\mathrm{num}}$	PF	$\max_{BL} = 30$	40.390	88.886	121.55	158.32	195.83	453.58
	SCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	15.059	63.707	83.965	125.82	142.99	391.39
	LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	4.6818	30.691	44.362	62.552	92.031	305.66
$\mathrm{DS}_{\mathrm{num}}$	PF	$\max_{BL} = 30$	540.74	304.36	255.40	150.89	117.24	18.795
	SCFG	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	840.03	522.53	452.04	307.61	273.92	77.536
	LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	568.66	172.46	143.60	72.662	57.327	9.5317

Table 12: Results related to the shapes of selected predictions and sampled structures, obtained from the S-151Rfam database (by 2-fold cross-validation procedures, using sample size 1000).

However, the results for the mixed S-151Rfam data set presented in Table 12 show a completely different picture. Most importantly, the considered specific values related to shapes are basically in all cases better when length-dependencies are *not* considered, i.e. when sticking to the simple SCFG model. This actually resembles the observations made in the last section for the sensitivity and PPV measures and hence provides additional evidence that incorporating length-dependency into a SCFG model for RNA secondary structures results in a much stronger dependence on the availability of a rich and pure training set.

5 Conclusions

In this article, we described how to extend the SCFG based statistical sampling method studied in [NSar] to additionally incorporate length-dependencies, yielding a corresponding LSCFG variant that samples possible foldings of a given RNA molecule recursively from the induced probability distribution. In particular, we evaluated a LSCFG based algorithm capable of producing a statistically representative sample of secondary structures for a given RNA sequence in proportion to the distribution on the entire ensemble of feasible foldings, where the corresponding distribution is immediately implied by the learned length-dependent grammar parameters. Just like the conventional variant originated from [NSar], this LSCFG method represents a probabilistic counterpart to the energy-based PF variant of Sfold (where structures are sampled in proportion to their Boltzmann weights, guaranteeing a statistical representation of the Boltzmann-weighted ensemble).

By performing a comprehensive comparative study of results obtained with the LSCFG, the traditional SCFG and the PF variant, respectively, we showed that significant differences with respect to both predictive accuracy and overall quality of generated sample sets are implied. Actually, we can conclude that the ensemble distribution induced by the considered LSCFG approach is much more centered than that induced by the conventional SCFG variant, and even seems to be slightly more centered than the Boltzmann-distribution of possible structures. This effectively yields less variability during the sampling process, resulting in a less diverse sample set that might contain typical structures significantly more often than others. In principle, a higher prediction accuracy can be reached at the price of a lower diversity of structures within generated sample sets. This is due to the higher explicitness of the underlying SCFG model implied by training the probabilities of the production rules in a length-dependent way. However, since the prediction accuracy is extremely high, the low variety within generated samples allows for the usage of rather small sample sizes to obtain meaningful structure predictions for a given RNA sequence. This indeed means that only a few candidate structures need to be sampled in order to derive high quality predictions, in contrast to the traditional SCFG (and also to competing PF) approach where a comparatively large number of structures needs to be generated in order to guarantee that the proposed folding is sufficiently accurate (and reproducible).
A further positive aspect is that existing algorithms for calculating all inside and outside values and the formulae for computing the needed sampling probabilities for statistical sampling as proposed in [NSar] can easily be modified by a few simple changes to cope with the extended SCFG model without significant losses in performance. Consequently, for particular RNA types, the extended LSCFG approach studied in this work might be able to improve the sampling quality (with respect to the investigated applications) over the conventional one, and especially over the PF variant, while the worst-case time and space complexities remain the same.

Taking all observations made throughout Section 4 into account, we may conclude that by adding length-dependency to the corresponding stochastic structure model (i.e. taking the lengths of generated substructures into account when learning the grammar parameters) can make a particular SCFG model (for a specific class of RNAs) more explicit and thus more powerful. As a consequence, the quality of generated samples with respect to the diverse applications investigated within this article becomes more independent of (the complexity of) the specified RNA type than when employing the simple SCFG variant (where the sampling quality seems to strongly depend on the structural variety and typical molecule length of the considered type of RNA). This overcomes the major drawback of probabilistic over energy-based statistical sampling techniques already formulated in [NSar], namely that the extend of improvement that can be reached by SCFG based sampling over the sampling with PFs seems to strongly depend on the considered RNA type.
In contrast to this benefit, however, there are also a number of undesirable pitfalls that come with the additional incorporation of length-dependencies. In fact, a potential overfitting and lack of generalization of the probabilistic structure model seems to become more likely, the first mainly for rather sparse training sets that are subject to high structural diversity and the latter at least for low invariant RNA types like tRNAs that obey to a single typical shape like the cloverleaf structure. Furthermore, the higher dependence on the availability of a rich training set caused by extending the underlying sophisticated SCFG model to a more explicit length-dependent one reduces the applicability of the corresponding probabilistic sampling approach in practice, especially for molecules where there exists hardly knowledge on the typical structural behavior of their family (in the form of trusted RNA databases).

References

[BBC$^+$00] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[CG98] T. Chi and S. Geman. Estimation of probabilistic context-free grammars. *Computational Linguistics*, 24(2):299–305, 1998.

[DE04] Robin D. Dowell and Sean R. Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5:71, 2004.

[Din06] Ye Ding. Statistical and bayesian approaches to RNA secondary structure prediction. *RNA*, 12:323–331, 2006.

[DL03] Ye Ding and Charles E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003.

[DWB06] Chuong B. Do, Daniel A. Woods, and Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.

[FH72] K. S. Fu and T. Huang. Stochastic grammars and languages. *International Journal of Computer and Information Sciences*, 1(2):135–170, 1972.

[Goo98] Joshua T. Goodman. *Parsing Inside-Out.* PhD thesis, Harvard University, Cambridge, Massachusetts, May 1998.

[Goo99] Joshua Goodman. Semiring parsing. *Computational Linguistics*, 25(4):573–605, 1999.

[GVR04] Robert Giegerich, Björn Voß, and Marc Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Research*, 32(16):4843–4851, 2004.

[HF71] T. Huang and K. S. Fu. On stochastic context-free languages. *Information Sciences*, 3:201–224, 1971.

[Hof03] Ivo L. Hofacker. The vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, 2003.

[JRG08] Stefan Janssen, Jens Reeder, and Robert Giegerich. Shape based indexing for faster search of RNA family databases. *BMC Bioinformatics*, 9(131), 2008.

[KH99] B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454, 1999.

[KH03] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31(13):3423–3428, 2003.

[Mai07] Robert S. Maier. Parametrized stochastic grammars for RNA secondary structure prediction. *Information Theory and Applications Workshop*, pages 256–260, 2007.

[McC90] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[MSZT99] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[NE07] Eric P. Nawrocki and Sean R. Eddy. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Computational Biology*, 3(3):e56, 2007.

[NJ80] R. Nussinov and A. B. Jacobson. Fast algorithms for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Science of the USA*, 77(11):6309–6313, 1980.

[NPGK78] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35:68–82, 1978.

[NS09] Markus E. Nebel and Anika Scheid. On quantitative effects of RNA shape abstraction. *Theory in Biosciences*, 128(4):211–225, 2009.

[NS11] Markus E. Nebel and Anika Scheid. Analysis of the free energy in a stochastic RNA secondary structure model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6):1468–1482, 2011.

[NSar] Markus E. Nebel and Anika Scheid. Evaluation of a sophisticated SCFG design for RNA secondary structure prediction. *Theory in Biosciences*, to appear.

[NSW11] Markus E. Nebel, Anika Scheid, and Frank Weinberg. Random generation of RNA secondary structures according to native distributions. *Algorithms for Molecular Biology*, 6(24), 2011.

[RD94] Sean R.Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 2(11):2079–2088, 1994.

[SBEB02] Maciej Szymanski, Miroslawa Z. Barciszewska, Volker A. Erdmann, and Jan Barciszewski. 5s ribosomal RNA database. *Nucleic Acids Res.*, 30:176–178, 2002.

[SHB$^+$98] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, 26:148–153, 1998.

[VC85] G. Viennot and M. Vauchaussade De Chaumont. Enumeration of RNA secondary structures by complexity. *Mathematics in medicine and biology, Lecture Notes in Biomathematics*, 57:360–365, 1985.

[WFHS99] S. Wuchty, W. Fontana, I. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999.

[WN11] Frank Weinberg and Markus E. Nebel. Applying length-dependent stochastic context-free grammars to RNA secondary structure prediction. *Algorithms*, 4(4):223–238, 2011.

[ZS81] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.

[Zuk89] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

[Zuk03] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.

Supplementary Material

Sm-I Computing Inside and Outside Probabilities

In order the determine all inside and outside variables for a given sequence r, we decided to use the LSCFG $\mathcal{G}_s$ as the basis for a special version of Earley's algorithm. In particular, we chose to rely on the formalism presented in [Goo98, Goo99] for describing parsers, which is called *semiring parsing*.

Sm-I.1 Notations

Let $\mathcal{R}_{\mathcal{G}_s,\bullet}$ denote the set containing exactly the so-called *dotted* rules that are considered by Earley's algorithm for grammar $\mathcal{G}_s$. Accordingly, by $\mathcal{R}_{\mathcal{G}_s,\bullet}(\mathit{rule})$ we denote the set of all dotted rules corresponding to the original rule $\mathit{rule} \in \mathcal{R}_{\mathcal{G}_s}$. Moreover, for any $A \in \mathcal{I}_{\mathcal{G}_s}$, let $\mathcal{R}_A = \{\mathit{rule} \in \mathcal{R}_{\mathcal{G}_s,\bullet} \mid \mathit{rule} = A \to \gamma\bullet\}$.
Briefly, the corresponding semiring parser computes inside and outside values for so-called *items* $[i, \mathrm{ind}(\mathit{rule}), j]$, where for a given input word r of length n, i and j, $1 \leq i, j \leq n+1$, define positions in r (from in front of the first character to after the last character) and $\mathrm{ind}(\mathit{rule})$ denotes the index of production $\mathit{rule} \in \mathcal{R}_{\mathcal{G}_s,\bullet}$ in an appropriate ordering of $\mathcal{R}_{\mathcal{G}_s,\bullet}$. Particularly, an item of the form $[i, \mathrm{ind}(A \to \alpha \bullet \beta), j]$ asserts that $A \Rightarrow \alpha\beta \overset{*}{\Rightarrow} r_i \ldots r_{j-1}\beta$.
Note that eventually, the needed inside and outside values $\alpha_A(i,j) = \Pr(A \Rightarrow_{lm}^{*} r_i \ldots r_j)$ and $\beta_A(i,j) = \Pr(S \Rightarrow_{lm}^{*} r_1 \ldots r_{i-1}\ A\ r_{j+1} \ldots r_n)$, $A \in \mathcal{I}_{\mathcal{G}_s}$ and $1 \leq i, j \leq n$, can easily be derived from the corresponding results for items $[i, \mathrm{ind}(A \to \gamma\bullet), j]$.

Sm-I.2 Ordering of Items

In order to guarantee the correctness of the respective algorithms for calculating the inside and outside values of all items $[i, \mathrm{ind}(\mathit{rule}), j]$, $i, j \in \{1, \ldots, n+1\}$ and $\mathit{rule} \in \mathcal{R}_{\mathcal{G}_s}$, the items have to be ordered such that no item precedes any other item on which it depends. Obviously, we can use the same ordering as defined in [NSar], that is:

- First and last parameters $i, j \in \{1, \ldots, n+1\}$ are ordered according to the consideration of items induced by Earley's algorithm and
- an appropriate ordering of the considered rule set $\mathcal{R}_{\mathcal{G}_s,\bullet}$ is given by indices (p,q), where $p \in \{1, \ldots, \mathrm{card}(\mathcal{R}_{\mathcal{G}_s})\}$ and $q \in \{0, \ldots, k(p)\}$ with $k(p)$ denoting the conclusion length of the production $\mathit{rule} \in \mathcal{R}_{\mathcal{G}_s}$ indexed by p.

Particularly, the grammar rules are ordered by first index $p \in \{1, \ldots, \mathrm{card}(\mathcal{R}_{\mathcal{G}_s})\}$ as follows:

Index p	Rule r	Index p	Rule r	Index p	Rule r	Index p	Rule r
1	$Z \to \circ$,	2	$A \to (^{m_s}L)^{m_s}$,	3	$P \to (L)$,		
4	$C \to ZC$,	5	$C \to Z$,	6	$H \to ZH$,	7	$H \to Z$,
8	$B \to ZB$,	9	$B \to Z$,	10	$U \to ZU$,	11	$U \to \epsilon$,
12	$T \to C$,	13	$T \to A$,	14	$T \to CA$,		
15	$T \to AT$,	16	$T \to CAT$,	17	$F \to Z^{m_h-1}H$,		
18	$G \to BA$,	19	$G \to AB$,	20	$G \to BAB$,		
21	$M \to UAO$,	22	$O \to UAN$,	23	$N \to UAN$,	24	$N \to U$,
25	$L \to F$,	26	$L \to P$,	27	$L \to G$,	28	$L \to M$,
29	$S \to T$.						

For any $\mathit{rule} \in \mathcal{R}_{\mathcal{G}_s}$ with first index $p \in \{1, \ldots, \mathrm{card}(\mathcal{R}_{\mathcal{G}_s})\}$, the corresponding $k(p)+1$ dotted rules in $\mathcal{R}_{\mathcal{G}_s,\bullet}(\mathit{rule})$ are ordered according to the actual positions of symbol $\bullet$, such that $q \in \{0, \ldots, k(p)\}$ corresponds to the dotted rule $\mathit{rule} \in \mathcal{R}_{\mathcal{G}_s,\bullet}(\mathit{rule})$ in which symbol $\bullet$ occurs after the qth symbol in the conclusion.

Sm-I.3 Inside and Outside Values of Items

The corresponding modified versions of the inside and outside algorithms from [NSar] are given by Algorithms 1 and 2. These two modified algorithms show how to perform the complete inside computation and – once the inside values are computed – how to calculate the corresponding outside values of all items.

Algorithm 1 Computation of Inside Values

```
Require: RNA sequence $r$ of length $n \geq 1$,
         set $\mathcal{R}_{\mathcal{G}_s,\bullet}$ of production rules used by Earley's algorithm, and
         rule probabilities $\mathrm{P}(rule, l)$ of the productions $rule \in \mathcal{R}_{\mathcal{G}_s}$, as well as
         emission probabilities $\mathrm{P}(x, l)$ of unpaired bases $x \in \{a, c, g, u\}$ and
         emission probabilities $\mathrm{P}(x_1x_2, l)$ of base pair $x_1x_2 \in \{a, c, g, u\}^2$,
         all trained on the same RNA structure data.
  for $j = 1 \dots n + 1$ do
    for $i = j \dots 1$ do
      for $p = 1 \dots \mathrm{card}(\mathcal{R}_{\mathcal{G}_s})$ do
        for $q = 0 \dots k(p)$ do
          $rule = \mathrm{ind}^{-1}(p, q)$ /*$rule \in \mathcal{R}_{\mathcal{G}_s,\bullet}$ is the rule having index $(p, q)$ in our ordering.*/
          if $rule = A \to \alpha w_{j-1} \bullet \beta$ then
            /* Scanning: */
            if $w_{j-1} =' \circ'$ then
              $\mathrm{IN}[i, (p, q), j] = \mathrm{P}(r_{j-1}, 1) \cdot \mathrm{IN}[i, (p, q - 1), j - 1]$
            else if $w_{j-1} =' ('$ then
              $\mathrm{IN}[i, (p, q), j] = \mathrm{IN}[i, (p, q - 1), j - 1]$
            else if $w_{j-1} =' )'$ then
              $\mathrm{IN}[i, (p, q), j] = \mathrm{P}(r_i r_{j-1}, (j - 1) - i + 1) \cdot \mathrm{IN}[i, (p, q - 1), j - 1]$
            end if
            if $q = k(p)$ /*$rule = A \to \alpha w_{j-1}\bullet$, i.e. rule is completed in this scanning step.*/ then
              $\mathrm{IN}[i, (p, q), j] = \mathrm{IN}[i, (p, q), j] \cdot \mathrm{P}(A \to \alpha w_{j-1}, \mathrm{len}(A \to \alpha w_{j-1}) = (j - 1) - i + 1)$
            end if
          else if $rule = B \to \bullet\gamma$ then
            /* Prediction: */
            if $\gamma = \epsilon$ /*$rule$ is $\epsilon$-rule.*/ then
              $\mathrm{IN}[j, (p, q), j] = \mathrm{P}(B \to \gamma, 0)$
            else
              $\mathrm{IN}[j, (p, q), j] = 1$
            end if
          else if $rule = A \to \alpha B \bullet \beta$ then
            /* Completion: */
            $\mathrm{IN}[i, (p, q), j] = \sum_{i \leq k \leq j} \left(\mathrm{IN}[i, (p, q - 1), k] \cdot \left(\sum_{rule_B \in \mathcal{R}_B} \mathrm{IN}[k, \mathrm{ind}(rule_B), j]\right)\right)$
            if $q = k(p)$ /*$rule = A \to \alpha B\bullet$, i.e. rule is completed.*/ then
              $\mathrm{IN}[i, (p, q), j] = \mathrm{IN}[i, (p, q), j] \cdot \mathrm{P}(A \to \alpha B, \mathrm{len}(A \to \alpha B) = (j - 1) - i + 1)$
            end if
          end if
        end for
      end for
    end for
  end for
```

Note that for the sake of simplicity and in order to demonstrate that both algorithms work in either case (length-dependent or not), we used the following notation:

$$\mathrm{P}(A \to \alpha, l) := \begin{cases} \mathrm{Pr}(A \to \alpha, l) \cdot {}^{1}\!/_{c_{\alpha,l}} & \text{if length-dependent,} \\ \mathrm{Pr}(A \to \alpha) & \text{else.} \end{cases}$$

$\mathrm{P}(x, l)$ and $\mathrm{P}(x_1x_2, l)$ are defined accordingly. However, it should be mentioned that in our algorithms, we are actually using the probability $\mathrm{P}(A \to (^{m_s}L)^{m_s}, j - i + 1) \cdot \mathrm{P}(x_i x_j, j - i + 1) \cdot \mathrm{P}(x_{i+1}x_{j-1}, j - i + 1 - 2) \cdot \mathrm{P}(x_{i+2}x_{j-2}, j - i + 1 - 4) \cdots \mathrm{P}(x_{i+(m_s-1)}x_{j-(m_s-1)}, j - i + 1 - 2 \cdot (m_s - 1))$ for the initialization of a helix (of minimum allowed size $m_s := \mathrm{min}_{\mathrm{hel}}$) with first base pair $i.j$, which is not quite right. In fact, going strictly with the formal definition, we would have to consider the term $\mathrm{P}(A \to (^{m_s}L)^{m_s}, j - i + 1) \cdot \mathrm{P}(x_i x_{i+1} x_{i+2} \dots x_{i+(m_s-1)} x_{j-(m_s-1)} \dots x_{j-2} x_{j-1} x_j, j - i + 1)$ which means if $\mathrm{min}_{\mathrm{hel}} > 1$ is chosen, we would have to derive and use an additional set of emission probabilities for any possible combination of $\mathrm{min}_{\mathrm{hel}}$ consecutive base pairs. Nevertheless, this inaccuracy can easily be corrected by modifying our grammar definition such that production $p_9 : A \to (^{m_s}L)^{m_s}$ can be simulated by the composition of new

Algorithm 2 Computation of Outside Values

```
Require: RNA sequence $r$ of length $n \geq 1$,
         set $\mathcal{R}_{\mathcal{G}_s,\bullet}$ of production rules used by Earley's algorithm, and
         rule probabilities $\mathrm{P}(rule, l)$ of the productions $rule \in \mathcal{R}_{\mathcal{G}_s}$, as well as
         emission probabilities $\mathrm{P}(x, l)$ of unpaired bases $x \in \{a, c, g, u\}$ and
         emission probabilities $\mathrm{P}(x_1x_2, l)$ of base pair $x_1x_2 \in \{a, c, g, u\}^2$,
         all trained on the same RNA structure data, and also
         the corresponding inside values (computed by Algorithm 1).
  $\mathrm{OUT}[1, \mathrm{ind}(S \to T\bullet), n+1] = 1$
  for $j = n+1 \ldots 1$ do
    for $i = 1 \ldots j$ do
      for $p = \mathrm{card}(\mathcal{R}_{\mathcal{G}_s}) \ldots 1$ do
        for $q = k(p) \ldots 0$ do
          $rule = \mathrm{ind}^{-1}(p, q)$ /*$rule \in \mathcal{R}_{\mathcal{G}_s,\bullet}$ is the rule having index $(p, q)$ in our ordering.*/
          if $rule = A \to \alpha w_j \bullet \beta$ then
            /* Scanning (reverse): */
            if $w_j =' \circ'$ then
              $\mathrm{OUT}[i, (p, q-1), j] = \mathrm{P}(r_j, 1) \cdot \mathrm{OUT}[i, (p, q), j+1]$
            else if $w_j =' ('$ then
              $\mathrm{OUT}[i, (p, q-1), j] = \mathrm{OUT}[i, (p, q), j+1]$
            else if $w_j =' )'$ then
              $\mathrm{OUT}[i, (p, q-1), j] = \mathrm{P}(r_i r_j, j-i+1) \cdot \mathrm{OUT}[i, (p, q), j+1]$
            end if
            if $q = k(p)$ /*$rule = A \to \alpha w_j\bullet$, i.e. rule is completed in this scanning step.*/ then
              $\mathrm{OUT}[i, (p, q-1), j] = \mathrm{OUT}[i, (p, q-1), j] \cdot \mathrm{P}(A \to \alpha w_j, \mathrm{len}(A \to \alpha w_j) = j-i+1)$
            end if
          else if $rule = B \to \bullet\gamma$ then
            /* Prediction (reverse): */
            do nothing
          else if $rule = A \to \alpha B \bullet \beta$ then
            /* Completion (reverse): */
            if $q = k(p)$ /*$rule = A \to \alpha B\bullet$, i.e. rule is completed.*/ then
              $\mathrm{fact} = \mathrm{P}(A \to \alpha B, \mathrm{len}(A \to \alpha B) = (j-1) - i + 1)$
            else
              $\mathrm{fact} = 1$
            end if
            for $k = i \ldots j$ do
              $\mathrm{OUT}[i, (p, q-1), k] =$
                  $\mathrm{OUT}[i, (p, q-1), k] + \big(\mathrm{OUT}[i, (p, q), j] \cdot \big(\sum_{rule_B \in \mathcal{R}_B} \mathrm{IN}[k, \mathrm{ind}(rule_B), j]\big)\big) \cdot \mathrm{fact}$
              for $rule_B \in \mathcal{R}_B$ do
                $\mathrm{OUT}[k, \mathrm{ind}(rule_B), j] =$
                    $\mathrm{OUT}[k, \mathrm{ind}(rule_B), j] + (\mathrm{OUT}[i, (p, q), j] \cdot \mathrm{IN}[i, (p, q-1), k]) \cdot \mathrm{fact}$
              end for
            end for
          end if
        end for
      end for
    end for
  end for
```

productions $p_9 : A \to \mathbf{(}A_1\mathbf{)}$, $1 : A_1 \to \mathbf{(}A_2\mathbf{)}$, $1 : A_2 \to \mathbf{(}A_3\mathbf{)}$, $\ldots$, $1 : A_{m_s-1} \to \mathbf{(}L\mathbf{)}$. Then, our algorithms work conform with the formal definition.

Finally, note that by factoring in the rule probability $\mathrm{P}(A \to \alpha, l)$ of production $A \to \alpha \in \mathcal{R}_{\mathcal{G}_s}$ in the last scanning or completion steps of the corresponding items $[i, \mathrm{ind}(A \to \alpha\bullet), j]$, $1 \leq i, j \leq n+1$, instead of as usually done initially in the prediction steps of the corresponding items $[i, \mathrm{ind}(A \to \bullet\alpha), j]$, this rule probability $\mathrm{P}(A \to \alpha, l)$ is not incorporated as a factor into the corresponding inside values $[i, \mathrm{ind}(A \to \alpha \bullet \beta), j]$, $1 \leq i, j \leq n+1$, if $\beta \neq \epsilon$. This means these values are not correctly computed. However, for $\beta = \epsilon$, the inside values of items $[i, \mathrm{ind}(A \to \alpha \bullet \beta), j] = [i, \mathrm{ind}(A \to \alpha\bullet), j]$, $1 \leq i, j \leq n+1$, are correctly calculated.

Sm-I.4 Deriving the Needed Inside and Outside Probabilities

Since for a given sequence r of length n, an item of the form $[i, \mathrm{ind}(A \to \alpha\bullet), j+1]$, $1 \leq i, j \leq n$, asserts that $A \Rightarrow \alpha \overset{*}{\Rightarrow} r_i \dots r_j$, we have

$$\alpha_A(i,j) = \sum_{rule \in \mathcal{R}_A} \mathrm{IN}[i, \mathrm{ind}(rule), j+1]$$

and

$$\beta_A(i,j) = \max_{rule \in \mathcal{R}_A} \mathrm{OUT}[i, \mathrm{ind}(rule), j+1].$$

For details, we refer to [NSar]. It should be noted, however, that for any given RNA sequence r of size n, the considered inside values of items of the form $[i, \mathrm{ind}(A \to \gamma\bullet), j+1]$, that is of items $[i, \mathrm{ind}(rule \in \mathcal{R}_A), j+1]$, for each intermediate symbol $A \in \mathcal{I}_{\mathcal{G}_s}$ and $1 \leq i, j \leq n$, are always correctly calculated (see above). Hence, the corresponding traditional inside probabilities $\alpha_A(i,j)$ are accurately derived. The same holds for the corresponding outside values.

Finally, we observe that the modifications that led to Algorithms 1 and 2 do not imply a significant additional computation effort. Therefore, for a sequence r of size n, there still results cubic time complexity and quadratic memory requirement in the worst case for the computation of all inside and outside probabilities $\alpha_A(i,j)$ and $\beta_A(i,j)$, $A \in \mathcal{I}_{\mathcal{G}_s}$ and $1 \leq i, j \leq n$.

Sm-II Tables and Figures

$\mathbb{V}[\cdot]$	tRNA	5S rRNA	S-151Rfam
p_1	0	0	0
p_2	5.747×10^{-8}	2.232×10^{-6}	1.613×10^{-5}
p_3	1.223×10^{-7}	6.635×10^{-6}	8.673×10^{-6}
p_4	3.745×10^{-8}	2.718×10^{-6}	1.012×10^{-5}
p_5	9.954×10^{-7}	3.437×10^{-6}	1.983×10^{-5}
p_6	9.579×10^{-7}	1.697×10^{-6}	4.120×10^{-5}
p_7	8.853×10^{-6}	2.849×10^{-5}	7.766×10^{-6}
p_8	8.853×10^{-6}	2.849×10^{-5}	7.766×10^{-6}
p_9	0	0	0
p_{10}	0	0	0
p_{11}	4.541×10^{-9}	1.385×10^{-9}	1.362×10^{-6}
p_{12}	2.645×10^{-8}	8.330×10^{-8}	2.264×10^{-6}
p_{13}	8.500×10^{-9}	6.674×10^{-8}	4.074×10^{-6}
p_{14}	6.762×10^{-10}	3.464×10^{-10}	3.270×10^{-7}
p_{15}	0	0	0
p_{16}	1.234×10^{-8}	7.211×10^{-9}	5.812×10^{-6}
p_{17}	1.234×10^{-8}	7.211×10^{-9}	5.812×10^{-6}
p_{18}	0	1.152×10^{-6}	5.352×10^{-5}
p_{19}	0	3.919×10^{-7}	2.957×10^{-5}
p_{20}	0	4.502×10^{-7}	8.094×10^{-5}
p_{21}	2.695×10^{-3}	2.997×10^{-8}	4.429×10^{-5}
p_{22}	2.695×10^{-3}	2.997×10^{-8}	4.429×10^{-5}
p_{23}	0	0	0
p_{24}	0	0	0
p_{25}	0	0	1.333×10^{-4}
p_{26}	0	0	1.333×10^{-4}
p_{27}	4.052×10^{-7}	1.561×10^{-7}	1.347×10^{-4}
p_{28}	4.052×10^{-7}	1.561×10^{-7}	1.347×10^{-4}
p_{29}	0	0	0

Table 13: Truncated variances of grammar parameters (transition probabilities), derived from 100 iterations of training the traditional (length-independent) SCFG $\mathcal{G}_s$ on random subsets containing 90 percent of the original data, respectively, under the assumption of $\min_{HL} = 3$ and $\min_{\text{hel}} = 2$.

CSP$_{\text{freq}}$ (selection principle MF struct.):

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0633	0.1216	0.2071	0.2117	0.2639	0.3694
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.2099	0.3699	0.5594	0.5594	0.5599	0.6302
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.2187	0.3833	0.5830	0.5830	0.5835	0.6607
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.2450	0.4448	0.6417	0.6417	0.6422	0.7356
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.2409	0.4364	0.6399	0.6399	0.6403	0.7379
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.3278	0.4896	0.6565	0.6570	0.6570	0.7341
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.2936	0.4018	0.6792	0.6792	0.6796	0.7642
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.3440	0.5137	0.6805	0.6805	0.6810	0.7628
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.2982	0.4124	0.6963	0.6963	0.6967	0.7873

CSP$_{\text{freq}}$ (selection principle MEA struct.):

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0416	0.1049	0.1923	0.1960	0.2496	0.3559
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.0555	0.2094	0.4193	0.4193	0.4207	0.4679
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.0656	0.2446	0.4961	0.4961	0.4984	0.5613
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.0772	0.2510	0.4928	0.4928	0.4942	0.5497
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.1008	0.2917	0.5525	0.5525	0.5543	0.6241
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.1854	0.3574	0.4919	0.4919	0.4919	0.5465
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.1956	0.4013	0.6824	0.6824	0.6829	0.7712
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.1951	0.3676	0.4979	0.4984	0.4979	0.5552
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.2053	0.4115	0.6958	0.6958	0.6963	0.7869

CSP$_{\text{freq}}$ (selection principle Centroid):

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0264	0.0800	0.1595	0.1627	0.1932	0.2677
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.0374	0.1276	0.2973	0.2973	0.2978	0.3130
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.0485	0.1623	0.3791	0.3791	0.3800	0.4097
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.0536	0.1665	0.3773	0.3773	0.3778	0.4060
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.0758	0.2150	0.4563	0.4563	0.4568	0.5003
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.1729	0.3158	0.4300	0.4300	0.4300	0.4762
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.1877	0.3768	0.6362	0.6362	0.6366	0.7157
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.1812	0.3199	0.4304	0.4304	0.4304	0.4780
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.1956	0.3824	0.6426	0.6426	0.6431	0.7240

Table 14: Results related to the shapes of selected predictions, obtained from our tRNA database (by 10-fold cross-validation procedures, using sample size 1000).

$\mathrm{CSO}_{\mathrm{freq}}$:

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.5196	0.6740	0.8160	0.8239	0.8798	0.9556
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.6838	0.9459	0.9903	0.9903	0.9908	0.9995
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.6806	0.9006	0.9630	0.9635	0.9640	0.9991
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.7148	0.9459	0.9875	0.9880	0.9885	0.9991
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.7111	0.8997	0.9677	0.9681	0.9686	0.9995
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.8234	0.9288	0.9723	0.9750	0.9727	0.9986
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.5479	0.8100	0.9006	0.9011	0.9011	0.9963
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.8391	0.9441	0.9778	0.9783	0.9783	0.9986
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.5479	0.8160	0.9015	0.9015	0.9020	0.9963

$\mathrm{CS}_{\mathrm{num}}$:

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	21.073	58.200	136.67	140.63	205.54	328.56
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	16.202	98.357	327.26	327.27	327.51	418.80
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	25.205	142.50	453.03	453.03	453.10	527.04
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	24.883	130.04	392.78	392.79	393.05	494.79
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	34.898	173.73	513.05	513.06	513.08	595.26
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	101.69	326.26	708.52	708.94	709.42	805.87
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	101.77	294.14	717.92	717.92	718.37	811.29
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	102.65	331.18	717.08	717.54	718.10	818.76
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	104.09	300.04	730.09	730.09	730.54	826.21

$\mathrm{DS}_{\mathrm{num}}$:

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	355.32	130.22	81.796	33.125	22.585	4.8848
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	802.27	244.52	60.504	60.030	59.916	28.764
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	652.75	125.69	24.687	24.687	24.687	16.019
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	752.71	208.65	48.257	47.797	47.691	21.838
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	592.84	103.04	18.921	18.921	18.921	12.053
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	238.30	15.045	5.6854	5.4122	5.1806	3.2274
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	125.37	8.2070	2.6736	2.6736	2.6736	2.4123
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	244.62	16.121	6.1883	5.8268	5.6244	3.1974
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	126.84	8.2815	2.7296	2.7296	2.7296	2.3869

Table 15: Results related to the shapes of sampled structures, obtained from our tRNA database (by 10-fold cross-validation procedures, using sample size 1000).

$\mathrm{CSP}_{\mathrm{freq}}$ (selection principle MF struct.):

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0000	0.0009	0.0078	0.0513	0.0261	0.6353
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.0000	0.0026	0.0052	0.0131	0.0357	0.7128
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0000	0.0052	0.0139	0.0331	0.0522	0.7502
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.0000	0.0044	0.0113	0.0314	0.0766	0.7781
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0009	0.0096	0.0244	0.0609	0.1027	0.8207
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.2002	0.4239	0.4700	0.4857	0.9426	0.9861
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0000	0.0087	0.0087	0.0522	0.9321	0.9948
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.1984	0.4221	0.4710	0.4857	0.9391	0.9835
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0000	0.0087	0.0087	0.0522	0.9313	0.9948

$\mathrm{CSP}_{\mathrm{freq}}$ (selection principle MEA struct.):

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0000	0.0052	0.0139	0.0835	0.0696	0.6640
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.0000	0.0000	0.0000	0.0000	0.0261	0.3820
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0000	0.0009	0.0009	0.0035	0.0566	0.4769
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.0000	0.0000	0.0000	0.0009	0.0261	0.3977
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0000	0.0009	0.0009	0.0035	0.0557	0.5387
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.1062	0.3891	0.4290	0.4378	0.9051	0.9835
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0000	0.0078	0.0078	0.0514	0.9078	0.9957
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.1062	0.4065	0.4456	0.4535	0.8990	0.9835
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0000	0.0078	0.0078	0.0540	0.9078	0.9948

$\mathrm{CSP}_{\mathrm{freq}}$ (selection principle Centroid):

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0000	0.0026	0.0104	0.0775	0.0731	0.7214
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.0000	0.0000	0.0000	0.0000	0.0104	0.1097
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0000	0.0000	0.0000	0.0000	0.0148	0.1279
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.0000	0.0000	0.0000	0.0000	0.0078	0.1236
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0000	0.0000	0.0000	0.0009	0.0139	0.1549
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.0966	0.2916	0.3238	0.3316	0.8703	0.9686
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0000	0.0061	0.0061	0.0426	0.8982	0.9887
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.0949	0.2951	0.3281	0.3386	0.8660	0.9712
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0000	0.0070	0.0070	0.0453	0.8982	0.9861

Table 16: Results related to the shapes of selected predictions, obtained from our 5S rRNA database (by 10-fold cross-validation procedures, using sample size 1000).

$\mathrm{CSO}_{\mathrm{freq}}$:

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0009	0.1662	0.3063	0.7580	0.6883	0.9817
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.0000	0.2855	0.4526	0.9852	0.9974	1.0000
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0017	0.4135	0.5754	0.9861	0.9983	0.9991
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.0000	0.3308	0.4883	0.9904	0.9974	1.0000
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0026	0.4509	0.6372	0.9904	0.9974	0.9991
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.6258	0.8912	0.9295	0.9504	0.9948	1.0000
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0000	0.0374	0.0392	0.5588	0.9957	1.0000
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.6197	0.8938	0.9286	0.9547	0.9948	1.0000
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0000	0.0435	0.0453	0.5822	0.9948	1.0000

$\mathrm{CS}_{\mathrm{num}}$:

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0009	0.7571	3.4207	36.641	30.288	600.35
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	0.0000	0.5432	1.1811	20.640	51.834	573.72
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0017	1.1428	2.6615	32.051	64.332	608.06
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	0.0000	0.6651	1.4309	22.983	54.635	569.80
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0026	1.3795	3.1949	36.673	71.080	609.58
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	42.599	347.33	421.29	455.78	881.11	983.88
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	0.0000	8.2238	8.3039	51.288	890.71	993.23
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	42.962	347.97	422.19	457.71	875.13	983.67
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	0.0000	8.2082	8.2918	51.573	884.49	993.06

$\mathrm{DS}_{\mathrm{num}}$:

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	710.75	333.72	237.71	93.335	63.661	7.0951
SCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	999.67	941.77	866.98	336.69	167.10	16.476
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	999.18	884.49	764.79	249.02	129.35	14.198
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	999.93	947.19	874.03	331.75	163.09	15.620
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	999.68	885.81	762.67	239.28	123.91	13.558
LSCFG	$\min_{HL} = 1, \min_{\mathrm{hel}} = 1$	318.99	24.878	19.283	8.2879	4.4246	1.2088
	$\min_{HL} = 1, \min_{\mathrm{hel}} = 2$	148.01	10.076	8.5355	4.4627	3.5160	1.1297
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 1$	325.18	26.023	20.266	8.4599	4.4933	1.2114
	$\min_{HL} = 3, \min_{\mathrm{hel}} = 2$	150.56	10.411	8.8139	4.5323	3.5690	1.1279

Table 17: Results related to the shapes of sampled structures, obtained from our 5S rRNA database (by 10-fold cross-validation procedures, using sample size 1000).

CSP_{freq} (selection principle MF struct.):

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0661	0.1255	0.1586	0.2050	0.2183	0.4834
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.0530	0.0993	0.1191	0.1324	0.1589	0.3776
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.0398	0.1193	0.1457	0.1656	0.1856	0.4106
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.0530	0.1259	0.1390	0.1590	0.1789	0.4107
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.0530	0.1258	0.1522	0.1788	0.1985	0.4240
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.0199	0.0465	0.0597	0.0663	0.0995	0.3245
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.0132	0.0465	0.0532	0.0664	0.0797	0.2982
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.0199	0.0532	0.0664	0.0730	0.0995	0.3179
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.0132	0.0532	0.0598	0.0664	0.0731	0.2916

CSP_{freq} (selection principle MEA struct.):

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0660	0.1123	0.1453	0.1984	0.2051	0.4902
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.0264	0.0927	0.0993	0.1125	0.1325	0.3778
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.0264	0.1193	0.1391	0.1523	0.1789	0.4239
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.0264	0.0927	0.0993	0.1125	0.1325	0.3777
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.0197	0.1127	0.1391	0.1656	0.2055	0.4109
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.0132	0.0397	0.0530	0.0530	0.0927	0.2118
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.0000	0.0332	0.0332	0.0398	0.0663	0.2254
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.0132	0.0397	0.0530	0.0596	0.0794	0.2118
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.0000	0.0398	0.0465	0.0465	0.0663	0.1854

CSP_{freq} (selection principle Centroid):

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.0793	0.1321	0.1653	0.1917	0.2449	0.5100
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.0197	0.0861	0.1059	0.1190	0.1258	0.3181
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.0197	0.0795	0.0926	0.1191	0.1192	0.3578
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.0197	0.0795	0.0926	0.1125	0.1125	0.3181
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.0197	0.0927	0.1125	0.1390	0.1391	0.3577
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.0132	0.0397	0.0530	0.0530	0.0729	0.1656
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.0000	0.0265	0.0332	0.0332	0.0663	0.1590
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.0066	0.0397	0.0530	0.0596	0.0728	0.1722
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.0000	0.0332	0.0398	0.0398	0.0596	0.1590

Table 18: Results related to the shapes of selected predictions, obtained from the S-151Rfam database (by 2-fold cross-validation procedures, using sample size 1000).

$\mathrm{CSO}_{\mathrm{freq}}$:

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	0.3638	0.4433	0.4766	0.5231	0.6488	0.7947
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.2520	0.5497	0.6095	0.6888	0.7683	0.9604
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.2717	0.5630	0.6158	0.7284	0.8079	0.9605
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.2518	0.5429	0.6093	0.7218	0.7815	0.9472
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.2715	0.5564	0.6027	0.7087	0.7484	0.9604
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	0.0463	0.2518	0.4041	0.5496	0.5960	0.8408
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	0.0397	0.2320	0.3381	0.4635	0.5033	0.7282
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	0.0463	0.2582	0.3908	0.5295	0.5825	0.8075
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	0.0331	0.1922	0.2982	0.4305	0.4635	0.6818

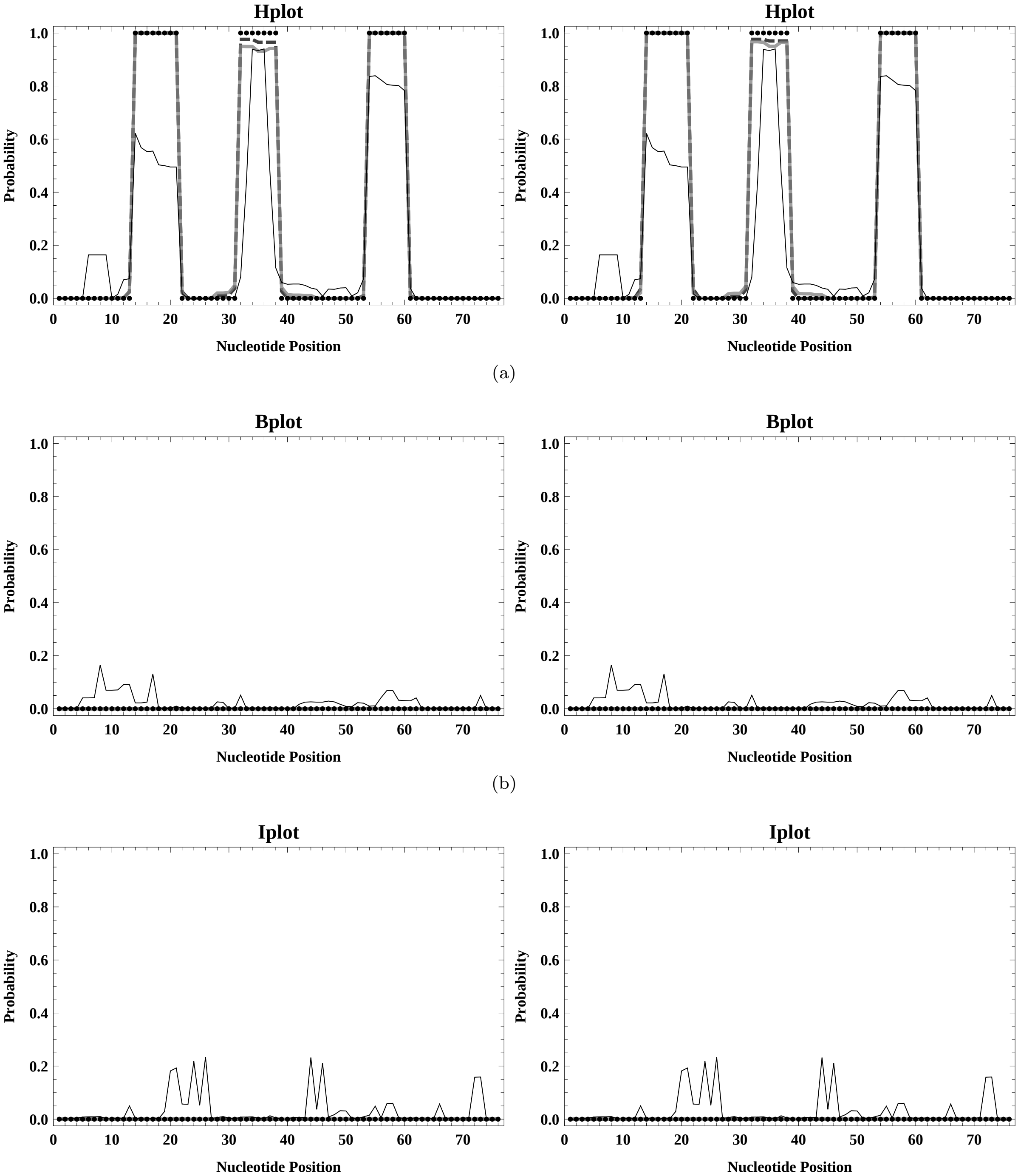
$\mathrm{CS}_{\mathrm{num}}$:

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	40.390	88.886	121.55	158.32	195.83	453.58
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	10.743	47.281	63.587	97.088	121.64	362.44
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	12.968	58.796	78.776	115.96	139.09	387.16
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	12.468	51.569	67.603	104.67	125.50	365.84
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	15.059	63.707	83.965	125.82	142.99	391.39
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	4.6818	30.691	44.362	62.552	92.031	305.66
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	3.2041	36.090	48.338	62.027	98.212	293.97
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	4.0326	28.718	41.792	59.675	86.897	300.72
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	3.3858	35.005	46.601	57.815	92.288	286.40

$\mathrm{DS}_{\mathrm{num}}$:

Approach	Parameters	Shape Level					
		0	1	2	3	4	5
PF	$\max_{BL} = 30$	540.74	304.36	255.40	150.89	117.24	18.795
SCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	892.14	600.39	526.36	368.49	322.88	99.601
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	849.32	538.56	466.17	322.99	286.12	84.480
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	888.89	588.97	516.66	358.72	315.25	94.603
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	840.03	522.53	452.04	307.61	273.92	77.536
LSCFG	$\min_{HL} = 1, \min_{\text{hel}} = 1$	729.44	249.69	201.75	102.87	78.918	13.381
	$\min_{HL} = 1, \min_{\text{hel}} = 2$	568.66	172.46	143.60	72.662	57.327	9.5317
	$\min_{HL} = 3, \min_{\text{hel}} = 1$	725.66	264.33	217.20	110.27	85.455	13.484
	$\min_{HL} = 3, \min_{\text{hel}} = 2$	563.23	180.29	151.89	74.803	59.977	8.9805

Table 19: Results related to the shapes of sampled structures, obtained from the S-151Rfam database (by 2-fold cross-validation procedures, using sample size 1000).

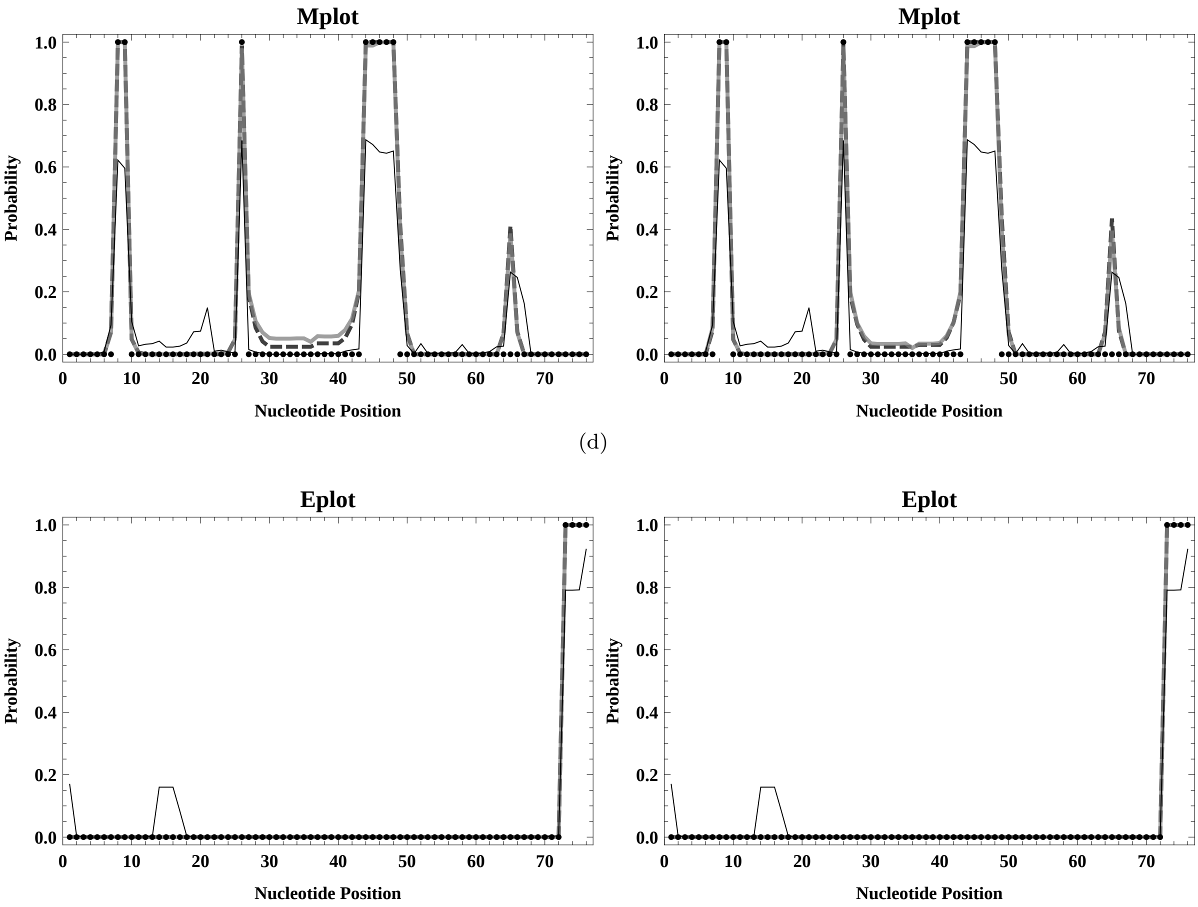

Hplot
Hplot
Bplot
Bplot
Iplot
Iplot
Probability
1.0
0.8
0.6
0.4
0.2
0.0
Nucleotide Position
0
10
20
30
40
50
60
70

(a)

(b)

(c)

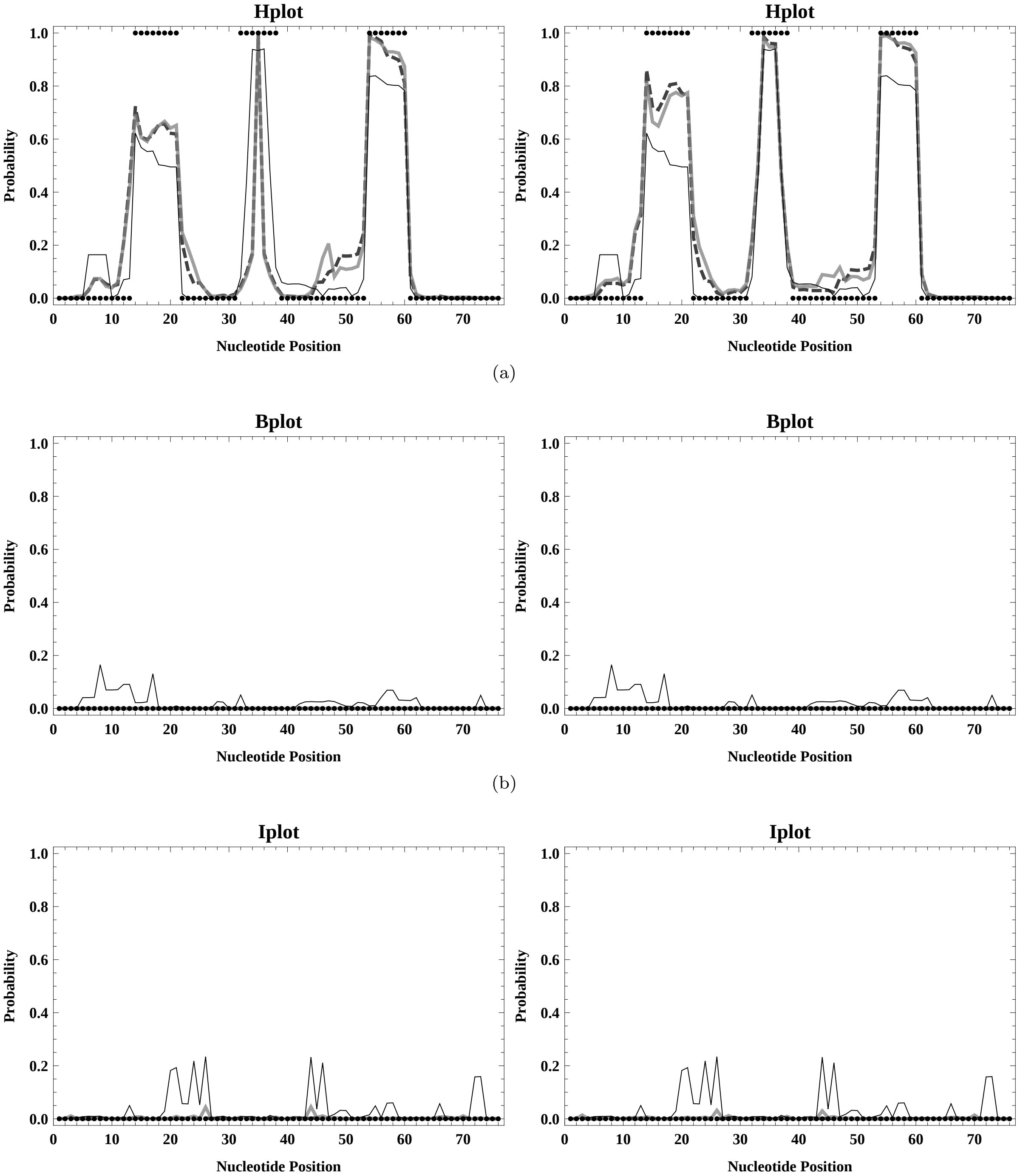
Figure 3



(d)

(e)

Figure 3: Comparsion of loop profiles for *E.coli* tRNA^{Ala}, obtained with the PF approach and the length-dependent SCFG variant. Hplot, Bplot, Iplot, Mplot and Extplot display the probability that an unpaired base lies in a hairpin, bulge, interior, multibranched and exterior loop, respectively. For each considered variant, these five probabilities are computed by a sample of 1000 structures generated by using $\max_{BL} = 30$. Results for the PF approach are displayed by the thin black lines. For the SCFG approach, we chose $\min_{\text{hel}} = 1$ (thick gray lines) and $\min_{\text{hel}} = 2$ (thick dashed darker gray lines), combined with $\min_{HL} = 1$ (figures shown on the left) and $\min_{HL} = 3$ (figures on the right), respectively. The corresponding probabilities for the correct structure of *E.coli* tRNA^{Ala} are also displayed (by black points).

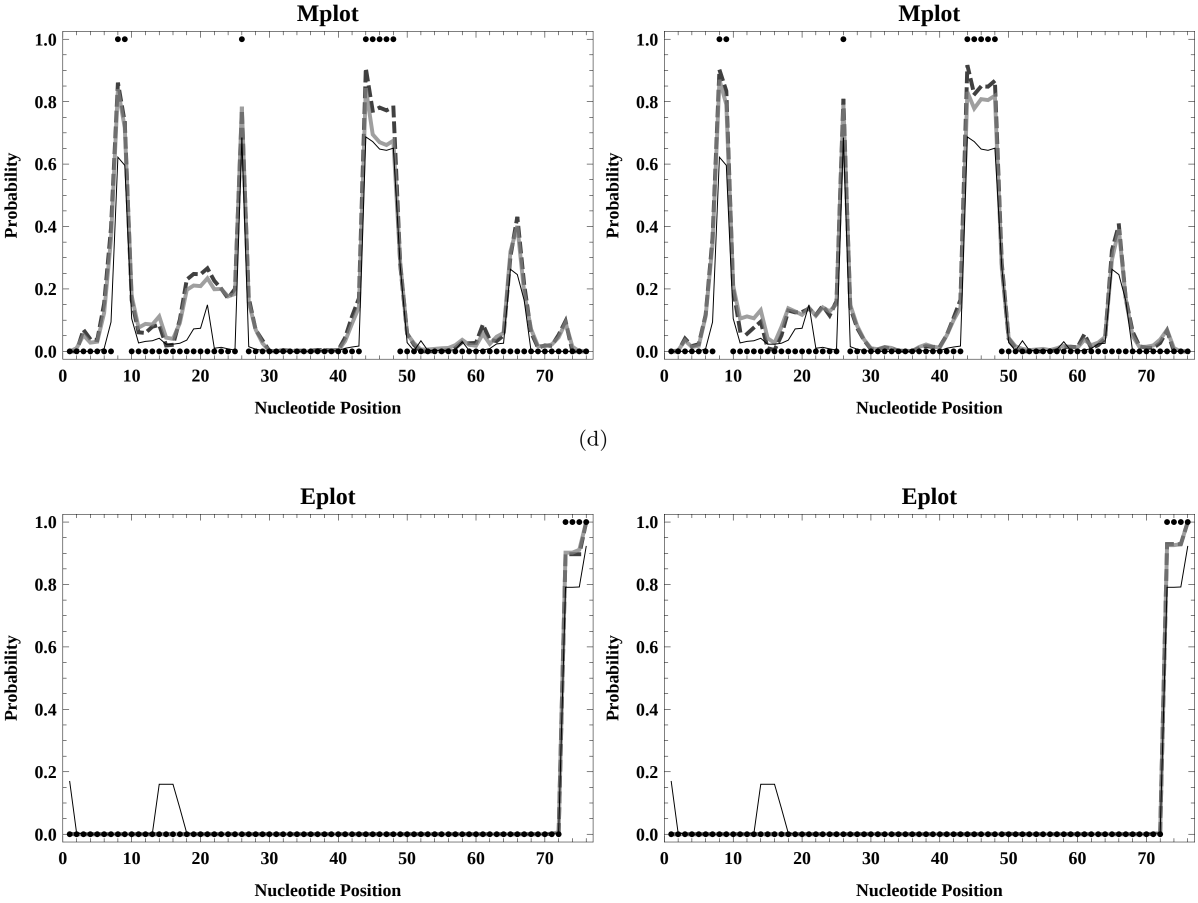

Hplot
Hplot
Bplot
Bplot
Iplot
Iplot
Probability
0.0
0.2
0.4
0.6
0.8
1.0
Nucleotide Position
0
10
20
30
40
50
60
70

(a)

(b)

(c)

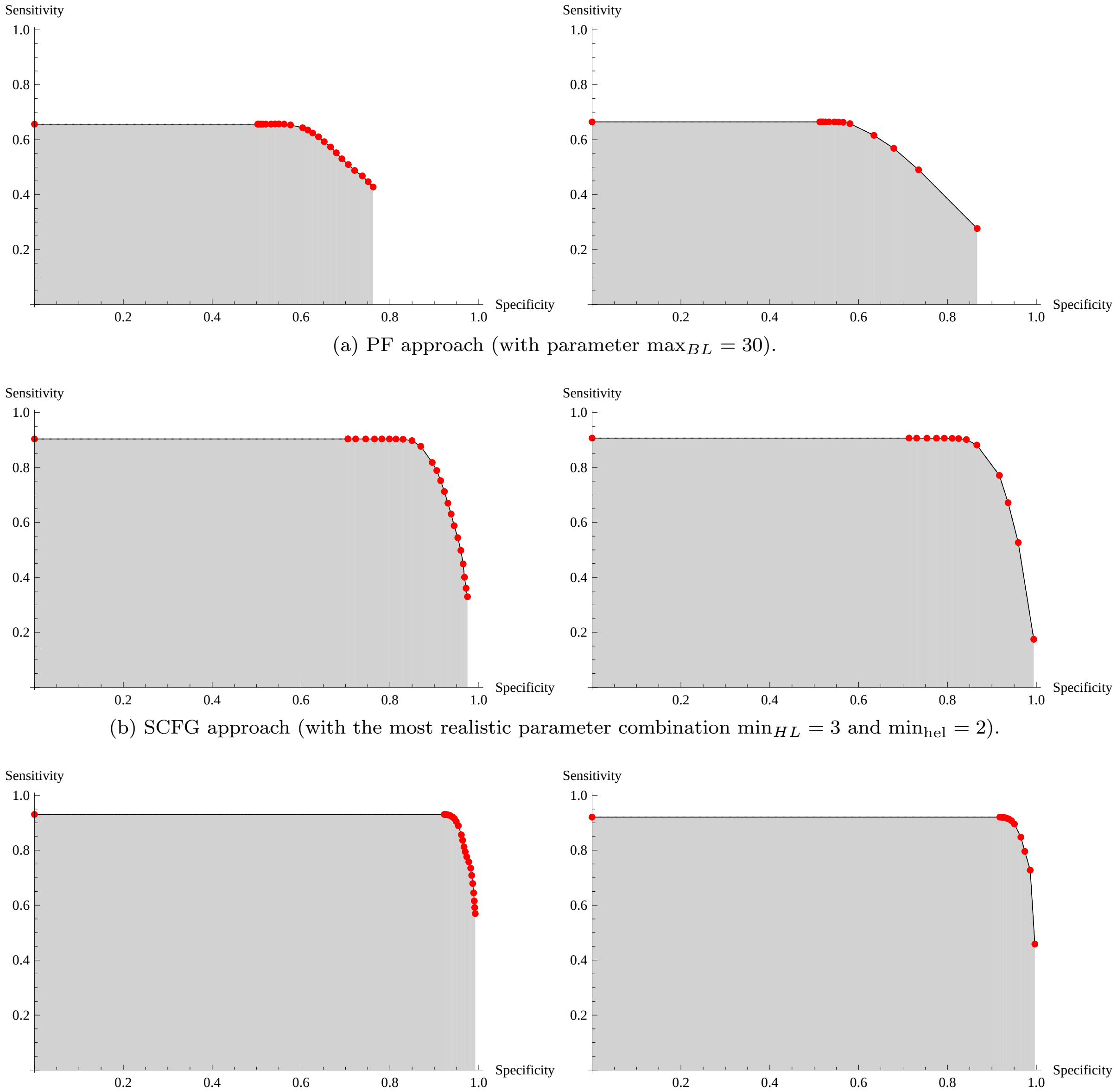
Figure 4



(d)

(e)

Figure 4: Loop profiles for *E.coli* tRNA^{Ala} corresponding to those presented in Figure 3, obtained with the PF approach and the traditional SCFG variant that does not incorporate length-dependencies.

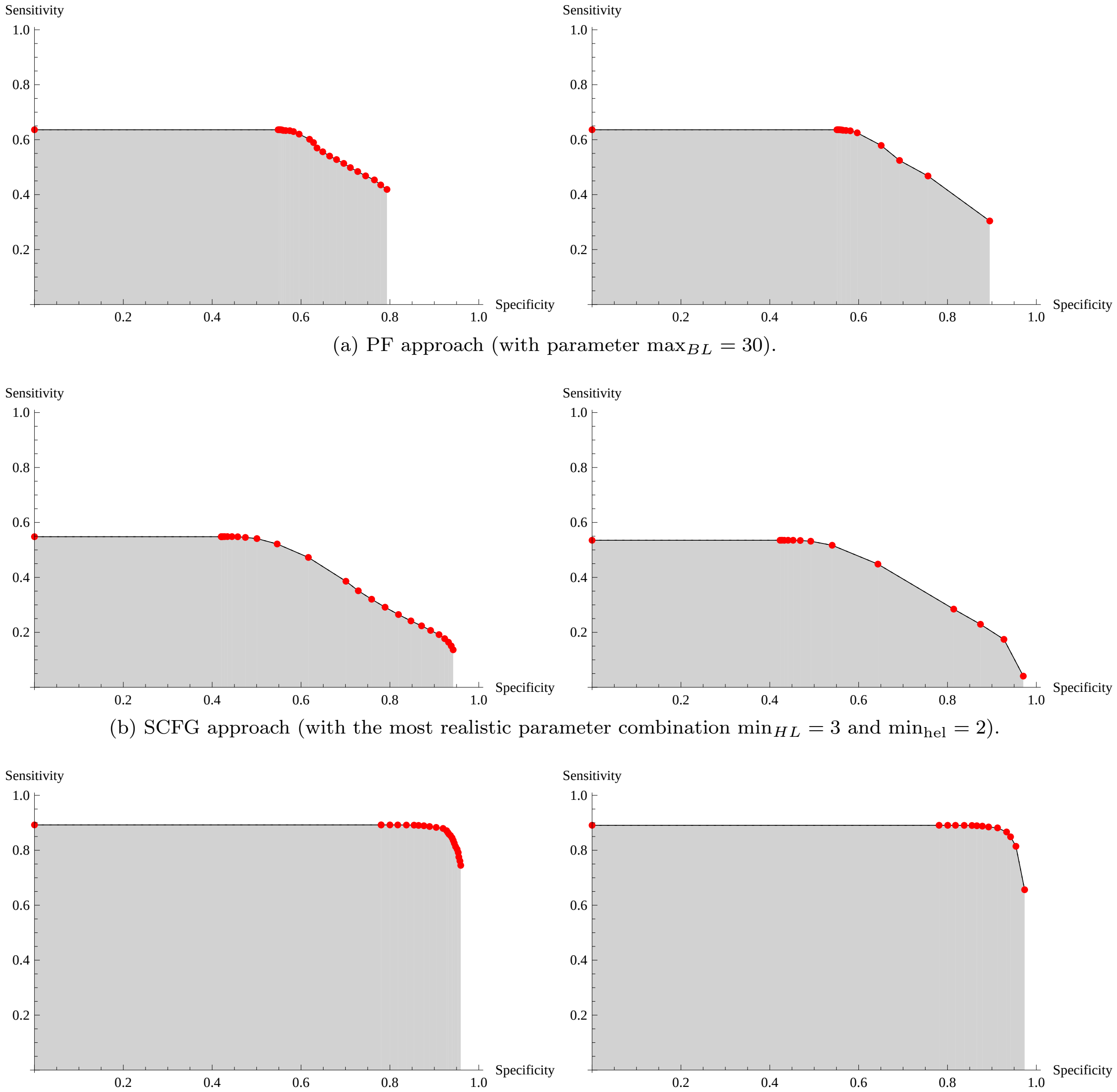


(a) PF approach (with parameter $\max_{BL} = 30$).

(b) SCFG approach (with the most realistic parameter combination $\min_{HL} = 3$ and $\min_{\mathrm{hel}} = 2$).

(c) Length-dependent SCFG approach (also with $\min_{HL} = 3$ and $\min_{\mathrm{hel}} = 2$).

Figure 5: Comparison of the (areas under) ROC curves obtained for our tRNA database (computed by 10-fold cross-validation procedures, using sample size 1000). For each considered sampling variant, the corresponding ROC curves are shown for prediction principle MEA structure (figure on the left) and centroid (figure on the right), respectively.

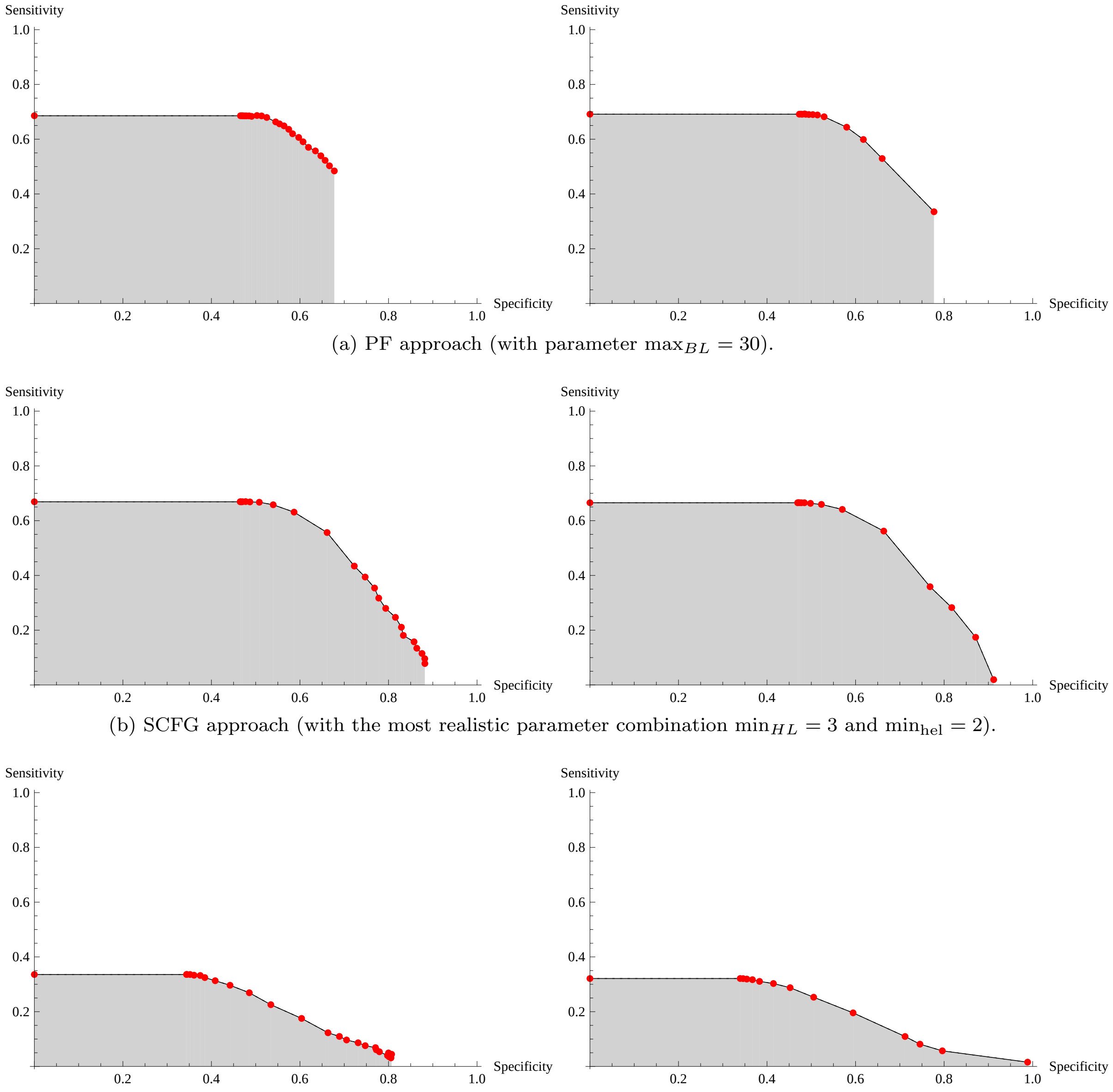


(a) PF approach (with parameter $\max_{BL} = 30$).

(b) SCFG approach (with the most realistic parameter combination $\min_{HL} = 3$ and $\min_{\text{hel}} = 2$).

(c) Length-dependent SCFG approach (also with $\min_{HL} = 3$ and $\min_{\text{hel}} = 2$).

Figure 6: Comparison of the (areas under) ROC curves obtained for our 5SrRNA database (computed by 10-fold cross-validation procedures, using sample size 1000). For each considered sampling variant, the corresponding ROC curves are shown for prediction principle MEA structure (figure on the left) and centroid (figure on the right), respectively.



(a) PF approach (with parameter $\max_{BL} = 30$).

(b) SCFG approach (with the most realistic parameter combination $\min_{HL} = 3$ and $\min_{\text{hel}} = 2$).

(c) Length-dependent SCFG approach (also with $\min_{HL} = 3$ and $\min_{\text{hel}} = 2$).

Figure 7: Comparison of the (areas under) ROC curves obtained for the mixed S-151Rfam database (computed by two-fold cross-validation procedures, using the same folds as in [DWB06] and sample size 1000). For each considered sampling variant, the corresponding ROC curves are shown for prediction principle MEA structure (figure on the left) and centroid (figure on the right), respectively.