genes

MDPI

*Article*

# RNA Secondary Structures with Limited Base Pair Span: Exact Backtracking and an Application

**Ronny Lorenz [1,*] and Peter F. Stadler [1,2,3,4,5,*]**

1 Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Vienna, Austria
2 Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics; German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Competence Center for Scalable Data Services and Solutions, and Leipzig Research Center for Civilization Diseases, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
3 Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
4 Facultad de Ciencias, Universidad National de Colombia, Sede Bogotá 111321, Colombia
5 Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
* Correspondence: ronny@tbi.univie.ac.at (R.L.); studla@bioinf.uni-leipzig.de (P.F.S.); Tel.: +43-142-77-52734 (R.L.); +49-341-97-16690 (P.F.S.)

**Abstract:** The accuracy of RNA secondary structure prediction decreases with the span of a base pair, i.e., the number of nucleotides that it encloses. The dynamic programming algorithms for RNA folding can be easily specialized in order to consider only base pairs with a limited span $L$, reducing the memory requirements to $O(nL)$, and further to $O(n)$ by interleaving backtracking. However, the latter is an approximation that precludes the retrieval of the globally optimal structure. So far, the ViennaRNA package therefore does not provide a tool for computing optimal, span-restricted minimum energy structure. Here, we report on an efficient backtracking algorithm that reconstructs the globally optimal structure from the locally optimal fragments that are produced by the interleaved backtracking implemented in RNALfold. An implementation is integrated into the ViennaRNA package. The forward and the backtracking recursions of RNALfold are both easily constrained to structural components with a sufficiently negative $z$-scores. This provides a convenient method in order to identify hyper-stable structural elements. A screen of the *C. elegans* genome shows that such features are more abundant in real genomic sequences when compared to a di-nucleotide shuffled background model.

**Keywords:** RNA secondary structure prediction; scanning algorithm; hyper-stable RNA elements



**Citation:** Lorenz, R.; Stadler, P.F. RNA Secondary Structures with Limited Base Pair Span: Exact Backtracking and an Application. *Genes* **2021**, *12*, 14. https://dx.doi.org/10.3390/genes1010000

Received: 24 November 2020
Accepted: 21 December 2020
Published: 24 December 2020



## 1. Introduction

Long-range base pairs are notoriously difficult to predict in RNA structures. The main reasons are that parts of the folding process in very long RNAs, say beyond a few hundred nucleotides, are likely to be influenced by co-transcriptional folding, and that RNAs are rarely, if ever, isolated in the cell. Consequently, long-range base pairs often do not fold as predicted by thermodynamic folding rules alone [1–3]. Performance limitations are also a consideration for very long sequences, since the effort grows cubicly with the sequence length $n$. Several tools have become available, which restrict the span of base pairs $(k, l)$ to $l - k + 1 \leq L$, including RNALfold [4], Rfold [5], and LocalFold [6]. An alternative approach penalizes long-range base pairs by reducing their energy contribution [2]. The two ideas were combined in [3]. Here, a sigmoidal function is used in order to interpolate between the full energy parameters and an upper bound on the base pair span. Restrictions on the base pair span are easily incorporated into the dynamic programming recursions [4–6]. This has the added benefit of resulting in an asymptotically linear resource consumption, namely $O(L^2 n)$ time and $O(Ln)$ memory in terms of the sequence length $n$ and the maximum base pair span $L$. The main memory requirement can be reduced to

$O(n + L^2)$ by writing intermediate results to the disk. This makes it possible to scan an entire genome for local secondary structure elements.

In this contributionm we first close a gap in the implementation of the ViennaRNA package [7,8]. The RNALfold program provides a tool for computing minimum free energy structure with limited span in $O(n + L^2)$ memory, but it only produces local candidate structures. Here, we show that these local structures can be assembled efficiently, in $O(nL)$ time, in order to yield the global minimum energy structure. We then discuss an additional restriction to unexpectedly stable local secondary structures, and, finally, sketch some application scenarios.

## 2. Theory

### *2.1. Backtracking from External Memory*

Consider the problem of folding RNA structure with a maximal base pair span $L$, i.e., for every base pair $(i, j)$ holds $j - i + 1 \leq L$. In the following, we write $D_{kl}$ for the optimal sub-structure on the sequence interval $[k, l]$ subject to the additional condition that the interval contains a single component, i.e., a substructure that is enclosed by base pair that itself is not contained inside any other base pair. Furthermore, we write $C_{kl}$ for the minimal free energy of a structure that is enclosed by the base pair $(k, l)$. In order to accomodate the so-called dangling ends appearing in the standard (Turner) energy model [9], we set

$$D_{kl} = \min\{C_{kl}, C_{k+1,l} + d^{5'}(k), C_{k,l-1} + d^{3'}(l), C_{k+1,l-1} + d^{*}(k,l)\} \tag{1}$$

Here, $d^{5'}(\,.\,)$, $d^{3'}(\,.\,)$, and $d^{*}(\,.\,)$ denote the 5′- and 3′- dangle parameters, and the dangling mismatch energy contributions, resp. The recursions that involve $D_{kl}$ correspond to an ambiguous decomposition of the secondary structure. Thus they can be used for energy minimization, but they cannot be translated directly for probabilistic models and partition function calculations. In the absence of dangling end contributions, we may use $D_{kl} = C_{kl}$.

The basic idea of RNALfold [4] can be summarized, as follows: denote, by $f_k$, the optimal free energy of a secondary structure with maximal base pair span $L$ on the interval $[k, n]$. This quantity satisfies the recursion

$$f_k = \min\begin{cases} f_{k+1} \\ \min_{k<l\leq k+L-1}[D_{kl} + f_{l+1}] \end{cases} \tag{2}$$

The first alternative shown in Equation (2) corresponds to $k$ being unpaired (and not subject to a dangling end contribution). The second alternative corresponds to a structure beginning with a single component structure of energy $D_{kl}$. As noted in [4], a structure realizing $D_{kl}$ needs to be considered to be a possible part of the minimum free energy structure only if $f_k < f_{k+1}$. Otherwise, the extension of a substructure on $[k + 1, l']$ by an unpaired base at position $k$ can be chosen instead. The index position $l'$ is determined by an evaluation of Equation (2) in the next step of the recursion.

The idea of span-restricted structures is also of interest in the context of maximum expected accuracy (MEA) methods [10,11]. The expected accuracy of a given secondary structure $\Upsilon$ is the sum of its base pairing probabilities $\hat{p}_{ij}$, $(i, j) \in \Upsilon$, plus the sum of probabilities $\hat{p}_k := 1 - \sum_{j<k} \hat{p}_{jk} - \sum_{j>k} \hat{p}_{kj}$ for the unpaired positions. Instead of treating this as a maximization problem, we mininimize $S_{ij}$, the negative of the accuracy. This highlights the similiarity of the MEA recursions with the thermodynamic folding models. MEA requires the base pairing probabolites $\hat{p}_{kl}$ as input. Therefore, these are computed with alternate methods, e.g. the partition function version of RNALfold [4], Rfold [5], or as the average over sequence windows that enclose the base pair of interest, as in RNAplfold [12] or LocalFold [6]. MEA models also use the weight $\hat{p}_k := 1 - \sum_{j<k} \hat{p}_{jk} - \sum_{j>k} \hat{p}_{kj}$ for base $k$ to be unpaired, which leads to a slight generalization of Equation (2):

$$f_k = \min\begin{cases} D_k + f_{k+1} \\ \min\limits_{k<l\leq k+L-1}[D_{kl} + f_{l+1}] \end{cases} \tag{3}$$

Here, an unpaired position $k$ contributes $D_k = -\hat{p}_k$ instead of 0, and the contributions of a single-component structure that is enclosed by the pair $(k, l)$ becomes $D_{kl} = (-\hat{p}_{kl}) + S_{k+1,l-1}$. The negative expected accuracy $S_{kl}$ follows the Nussinov-like [13] recursion

$$S_{kl} = \min \begin{cases} (-\hat{p}_k) + S_{k+1,l} \\ \min\limits_{k<j\leq l} \left[ (-\hat{p}_{kj}) + S_{k+1,j-1} + S_{j+1,l} \right] \end{cases} \tag{4}$$

with $S_{kk} := -\hat{p}_k$ and the convention that $S_{l+1,l} = 0$ for the empty interval. The condition for local structure candidates also needs to be modified in order to account for the contribution of unpaired bases and it becomes

$$f_k < D_k + f_{k+1}\,, \tag{5}$$

in the general case. Note that it is not necessary to store all of the values of the matrix $D_{kl}$ in the forward recursion. Instead, we backtrack at position $k$ the optimal structure whenever $f_k$ satisfies the second alternative shown in Equation (3) and only record that structure and the values of $-\hat{p}_{kl}$ for the base pairs in that structure. Optimal structures contained as substructures in a larger one are omitted, as in the MFE case.

The same scheme applies to generalization of the MEA structures. In order to obtain the centroid structure [14], it suffices to consider only the base pairs with $\hat{p}_{kl} \geq 1/2$. The $\gamma$-centroid that was proposed in [15] is obtained by $\hat{p}_k \leftarrow 0$ and $\hat{p}_{kl} \leftarrow (\gamma + 1)\hat{p}_{kl} - 1$. Similar expressions pertain to the estimators discussed in [16].

### *2.2. Increased Memory Efficiency by Reduced Redundancy*

RNALfold reduces the memory requirements of the algorithm by only keeping a small part of the $D$ (or $C$) matrices in memory, namely the range $[k, k + L - 1]$ that is necessary to evaluate Equation (2). Instead of storing the dynamic programming tables to enable backtracking, RNALfold immediately backtracks a single-component structure $\Psi_{kl_k}$ for $k$ and then stores it on disk. Here, $l_k$ is the end position of the first component of an optimal structure on $[k, n]$. Because of the restriction on the span $L$, we know *a priori* that this structure cannot reach beyond the interval $[k, k + L]$, i.e., $l_k \leq k + L$. More precisely, the end position $l_k$ is the value of $l$, for which the minimum in Equation (2) is attained. In the most straighforward version, the triple $(k, \Psi_{kl_k}, D_{kl_k})$ is written to disk, where triples of the form $(k, '.', D_{kk})$ can be omitted in minimum energy directed folding, since $D_{kk} = 0$, by definition, for all exterior bases.

In the simplest case, the substructure $\Psi_{kl_k}$ are stored in dot-parenthesis notation. Ref. [4] already noted that (together with the sequence information) these are sufficient for constructing the globally optimal secondary structure. In the case of MEA structures, $D_{kl}$ also needs to be stored explicitly. The required disk space is $O(nL)$. It can be reduced by a considerable constant factor; however, since the energy of a given structure can be evaluated in linear time. Therefore, it suffices to store structures $(k, \Psi_{kl_k})$ that are maximal in the sense that there is no $(k', \Psi_{k'l_{k'}})$, such that $\Psi_{kl_k}$ is proper substructure of $\Psi_{k'l_{k'}}$. In addition, inthe case of MEA structures, we need the proabilities $\hat{p}_j$ and $\hat{p}_{ij}$ for the unpaired bases and the base pairs in the candidate structure in order to be able to compute the $D_{kl}$ value for substructures that have not been explicitly stored.

In the case of MFE structures, the energy of a stored structure can be directly evaluated from the energy model. In the case of MEA structures; however, the base pairing probabilities, or more generally the derived scores, $\hat{p}_{kl}$ of all pairs $(k, l)$ as well as the scores of the unpaired $\hat{p}_k$ of all unpaired positions must be available in the input. An implementation of the MEA option is forthcoming.

So far, backtracking of the global MFE or MEA structure in not available in RNALfold. Here, we close this gap. Backtracking starts from the 5′ end, i.e., from the end of the file storing the candidate fragments. For a given $k$, the task is to find $l_k$, such that $f_k = D_{k,l_k} + f_{l+1}$, unless $f_k = D_k + f_{k+1}$. In the latter case position $k$ is unpaired and the recursion

continues with $k \leftarrow k+1$. The difficulty in the first case is that there is not necessarily an entry for $k$ in the output of RNALfold. However, it suffices to consider the set of candidate structure that contain position $k$, i.e.,

$$\mathcal{L}(k) := \{\Psi_{k',l_{k'}} | k \in [k', l_{k'}]\} \tag{6}$$

For each $\Psi$ in $\mathcal{L}(k)$, one can determine in linear time the corresponding candidate structures, as follows: (1) determine the base pair $(p,q)$ in $\Psi$ with the smallest value of $p \geq k$. If the base following $q$ is paired, the only candidate is the restriction $\Psi[k,q]$. In the case of an model with dangling ends, both $\Psi[k,q]$ and $\Psi[k,q+1]$ must be considered. The free energies $\varepsilon(\Psi[k,q])$ and $\varepsilon(\Psi[k,q+1])$ for these (explicitly given) sub-structures can be evaluated in linear time. Then one has to check, for $\Psi \in \mathcal{L}(k)$ and $q' = q$ and, in the case of dangling ends, also $q' = q+1$, where the structure satisfies

$$f_k = \varepsilon(\Psi[k,q']) + f_{q'+1}\,. \tag{7}$$

The evaluation continues at position $k \leftarrow q'+1$, where $q'$ is the first alternative for which equality is found in Equation (7). The backtracking method is also applicable without change to computations with constrained structures [17], since it only relies on the fact that a structure fragment is available in the output of the forward recursion that satisfies Equation (7).

### *2.3. Performance Analysis and Implementation*

A naïve estimate of the CPU requirement that is required for backtracking yields an upper bound of $O(nL^2)$, since $|\mathcal{L}(k)|$ contains, at most, $O(L)$ entries of size at most $O(|L|)$, each of which certainly can be evaluated in linear time. However, it is not necessary to construct the list $\mathcal{L}(k)$ "from scratch" in each step. Instead, for each position $k$, at most one additional entry $\Psi_{k,l_k}$ is added to the list, and every other entry can be "edited" after it has been processed for position $k-1$ by removing from the 5′ end a leading unpaired position or the base pair $(k-1, l_{k-1})$, respectively, as well as any structures trailing $(k, l_{k+1})$. Assuming that the dot-parenthesis structure has been converted into an ordered list of base pairs, the effort to adjust a list entry requires only constant time to obtain $\Psi[k,l_k]$ from $\Psi[k-1,l_{k-1}]$. Thus, the evaluation of the energies or MEA scores requires only constant effort for each position that is removed. Because the total size of the stored structures is bounded by $O(nL)$, the total effort for backtracking is also bounded by $O(nL)$.

The backtracking algorithm is integrated into the ViennaRNA package and it si available for RNALfold with the command line option -b/--backtrack-global. Empirical tests, see Figure 1, show that the implementation conforms to the theoretical $O(nL)$ performance bound.

### *2.4. New Options in RNALfold*

In order to better support genome-wide screens for structured RNA elements, we added options to filter structural components, which is, maximal substructures that are enclosed by a base pair. This is motivated, in particular, by the observation that the secondary structures of many structured small ncRNAs are more stable than expected from their sequence composition [18–20]. This effect is particularly pronounced for miRNAs [21]. This relative stabilization for a candidate sequence $x$ is conveniently quantified as a $z$-score, $z(x) := (f(x) - \bar{f})/\sigma_f$, where $f(x)$ is the folding energy of $x$, and $\bar{f}$ and $\sigma_f$ denote the mean and standard deviation of the folding energies of an ensemble of sequences with the same sequence composition. Computing $z(x)$ can be viewed as a regression problem in terms of parameters that specify the composition of $x$ [22]. Here, we use the SVM model of RNAz 2.0 [23], which explicitly depends on dinucleotide frequencies.

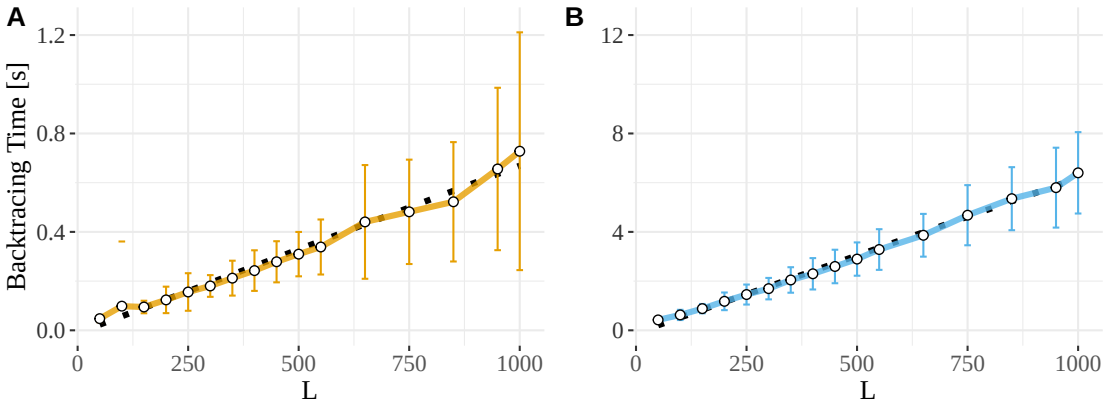


**Figure 1.** CPU requirements for backtracking the MFE structure in RNALfold. The performance of the implementation ViennaRNA conforms to a linear dependence of backtracking time per nucleotide with base pair span $L$, i.e., $t_{\mathrm{CPU}} \sim \sum nL$. Shown is the computational overhead for different window sizes $L$ that were obtained from averaging over 100 random sequences of length (**A**) 10,000 nt and (**B**) 100,000 nt. Error bars denote the standard deviation within the sets of 100 runs, and a linear fit is depicted by a dashed black line.

The regression approach, in contrast to shuffling, allows for a very fast (and deterministic) computation of the $z$-score $z^C_{kl}$ of the sub-structure that is enclosed by the base pair $(k, l)$. This can be used in two ways to restrict the predicted structure to components with $z^C_{kl}$ below a user-defined threshold $z^*$: (1) one can already restrict the forward recursion Equation (2) to base pairs enclosing components with a sufficiently negative $z$-score (pre-filter), and (2) one can suppress components with an insufficient $z$-score in the backtracking step (post-filter). The two methods are *not* equivalent. For example, in case (2), it is possible that larger component structure with better MFE but below-threshold $z$-score is computed at the expense of a smaller structure with better $z$-score.

Both of the methods have been implemented in RNALfold. The restriction of the forward recursion is accessed with the new option RNALfold --zscore-pre-filter. Backtracking is then unaffected by the restriction to components that surpass the $z$-score threshold. As an alternative, filtered backtracking of the unmodified RNALfold output (post-filter) is performed in default mode, i.e., whenever this option is omitted. For combinations of $z$-score filtering and backtracking of global MFE structures, as described in Section 2.1; however, RNALfold automatically activates the newly implemented restriction of the forward recursion. This is due to the fact that all of the structural alternatives that constitute the global solution are required for successful backtracking. Furthermore, RNALfold defaults to omit locally optimal structures if they are constituents of another, larger structure with less free energy. This might be undesirable for predictions with z-score filtering, as the substructure may exhibit a lower z-score than the larger, enclosing structure. The novel option RNALfold --zscore-report-subsumed can be used in order to alleviate this effect.

## 3. Application: Scanning Genomes for "Hyper-stable" RNA Structures

Some of the early surveys for ncRNAs used GC content and folding energy as indicators of structured RNAs. This approach was successful in particular in A/T-rich genomes of hyperthermophiles, such as *Methanococcus jannaschii* or *Pyrococcus furiosus* [24]. The extended version of RNALfold now makes it particularly easy to scan genomes for unexpectedly stable local structure.

As a show-case application, we screened the genomes of nematode *Caenorhabditis elegans* (Assembly WBcel235, Genome Assembly ID GCA_000002985.3) for highly stable component structures. For a given cut-off value $-z^*$, we recorded all of the components with a $z$-score $z \leq -z^*$ and compared the results to the ncRNA annotation available at Ensembl Release-101. In terms of the annotated elements, we found that there are only marginal differences between the two alternative strategies for $z \leq -2$, see Figure 2. As expected, recall decreases

in a class-specific manner as the $z$-scores become more negative. In particular, microRNAs persist longer than other classes of ncRNAs.

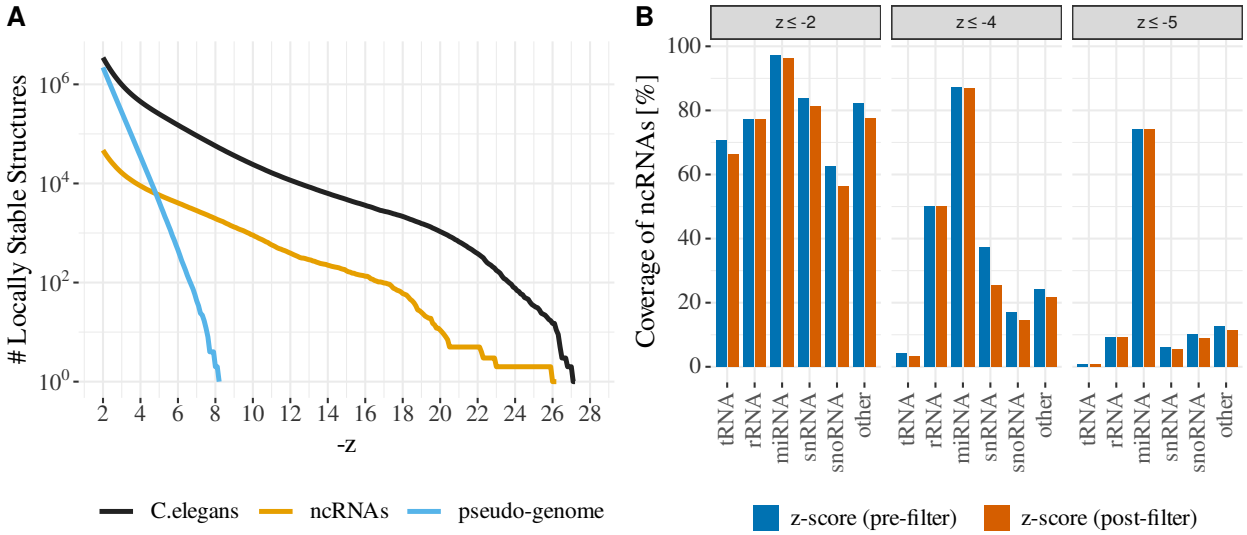


**Figure 2.** Predictions on the *C. elegans* genome. (**A**) Cumulative $z$-score distribution of predicted blocks with pre-filtering and default $z$-score threshold of $z \leq -2.0$ and window length $L = 150$. Shown are all predicted blocks for the entire *C.elegans* genome (black line) and those that sufficiently overlap with annotated *ncRNAs* (yellow line). The *pseudo-genome* (lightblue line) denotes blocks that were predicted on a di-nucleotide shuffled *C. elegans* genome. (**B**) Comparison of prediction coverage ($L = 150$) of the two $z$-score filter methods. Shown is the percentage of annotated ncRNAs that are sufficiently overlapped by the predicted locally stable structures at different $z$-score filter thresholds.

A comparison of the real data with a pseudo-genome that are generated by dinucleotide-shuffling [25] shows that the number of local structures that are more stable than a given $z$-score threshold $z^*$ decreases exponentially with $-z^*$, as in Figure 2A. The real data only follow the same distribution for small $-z^*$, but they show a tail with a smaller slope for large values of $-z$. This indicates that the genome is enriched in "hyperstable" RNA structures. A comparison of the distribution with annotated ncRNAs (including long non-coding RNAs, which are not expected to be particularly heavily structured over their whole length) suggests that this tail, indeed, corresponds to structured RNAs. The data also indicate that the vast majority of the approximately 57,000 "hyper-stable" elements with $z$-scores below $-8$ have remained unannotated. Approximately 41,000 (>70%) of those elements are predicted within low-complexity and repeat regions, as detected by RepeatMasker, while only 222 partially overlap with annotated coding sequences. For the former, we find that a large number of the predicted hyper-stable elements overlap with repeat classes/families that are annotated as DNA transposon (11,561), Simple Repeats (7977), and Satellite (7832). Long terminal repeats (LTR) are overlapped by 417 hyper-stable elements, while general low complexity regions, LINEs, and SINEs are overlapped by 153, 55, and 14, respectively. Low complexity repeats have previously been described to form highly stable structures and they have been studied, in particular, in the context of triplet repeat expansion diseases [26].

We further compared our predictions for the A/T-rich genomes of hyperthermophiles against the 32 candidate ncRNAs of length 49–238 nt listed in Klein et al. [24]. Here, we find that using default settings ($L = 150$, $-z^* = -2.0$), both approaches, pre- and post-filter, predict locally stable elements that overlap with at least 10% for virtually all of the candidate ncRNAs. The only exception is Mj8, which is not detected by the post-filter method. When requiring 50% overlap, the pre- and post-filter approach detects 21 and 22 out of a total of 22 candidate ncRNAs for *Pyrococcus furiosus*, respectively. For *Methanococcus jannaschii*, pre- and post-filter yields nine and eight of a total of 10 candidate ncRNA loci. With 12 and 10 out of 22 candidate ncRNA loci in *P. furiosus* for pre-

and post-filter, respectively, approximately one half is fully overlapped (100%) by our predictions. This is similar to the predictions for *M. jannaschii*, where we find seven and four out of 10 candidate ncRNA loci, respectively. When we increase the window length to $L = 250$ to accomodate the lengths of the queries, all 10 candidate ncRNAs in *M. jannaschii*, and the majority (20 for pre-filter, 18 for post-filter) of the 22 candidate ncRNAs in *P. furiosus* are detected with an overlap of 100%. For the elements that fully overlap with the candidates, the majority of $z$-scores is larger than $-3.0$, where the lowest $z$-score found is about $-4.8$. Still, among our pre-filtered predictions are a further 927 (*P. furiosus*) and 3007 (*M. jannaschii*) locally stable structures with $z < -4.8$, which account for 2.5% and 4.5% of the genomic DNA, respectively. A closer investigation reveals that approximately 29% (*P. furiosus*) and 22% (*M. jannaschii*) of these predicted elements overlap, at least in part, with other annotated ncRNAs, including rRNAs amd tRNAs. On the other hand, about 35% and 27% partially overlap with protein coding regions in the two genomes, respectively. This leaves 338 elements at 136 distinct non-coding loci in *P. furiosus* and 1609 elements at 287 loci in *M. jannaschii* as novel ncRNA candidates.

## 4. Conclusions

In this contribution, we have described an algorithm to reconstitute the MFE RNA secondary structure with limited base pair span from locally optimal structures. The method is applicable to effectively arbitrarily long RNA sequences and it closes the gap in the current toolkit that is provided by the ViennaRNA package. Arguably, the exact computation of such span-restricted MFE is of limited interest, since most RNA molecules of practical interest do not exceed the length range tractable without span restrictions. Furthermore, RNALfold and RNAplfold are primarily intended to scan entire genomic regions and provide local information, i.e., tasks for which local predictions that were provided by RNALfold could be used. However, the overlapping nature of these predictions is inconvenient, in particular, in the context of annotation, where one would like a partition of the input sequence into disjoint local structures. In order to become interpretable, the output of RNALfold therefore requires some form of postprocessing to reconcile overlapping local structures. The span-restricted MFE structure by construction consists of a partioning into pairwise disjoint components, i.e., base pair enclosed domains. Because the backtracking procedure that is described here has a running time of $O(nL)$, it is asymptotically optimal in the sense that postprocessing tools cannot be much faster, since the total amount of output of the forward recursion is also of size $O(nL)$.

The new backtracking functionality makes it easy to scan genome-scale data set for unusually stable structure. First and foremost, this provides a potentially useful prefilter for other, more computationally demanding methods that search for specific types of non-coding RNAs, in particular microRNAs. However, Figure 2 also indicated that there is a large number of "hyper-stable" secondary structure elements that deserve more attention in their own right. The bulk of the hyperstable structures in *C. elegans* falls into repetitive elements. Earlier studies of structured ncRNAs explicitly excluded repetitive DNA. Our results suggest that these deserve more detailed attention in future research.

**Author Contributions:** P.F.S. and R.L. jointly designed the study, P.F.S. focussed on the theory, R.L. was responsible for the implementation and conducted the showcase application. Both authors collaborated in writing the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the German Federal Ministry for Education and Research (BMBF 031A538B as part of de.NBI, to P.F.S.).

**Data Availability Statement:** The software described in this contribution is available as part of the ViennaRNA package https://www.tbi.univie.ac.at/RNA/.

**Acknowledgments:** This research originated over beer during a conference in Valetta, Malta in March 2020.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| MFE | Minimum free energy (secondary structure) |
| miRNA | microRNA |
| ncRNA | Non-coding RNA |
| MEA | Maximum expected accuracy (secondary structure) |
| TLA | Three letter acronym |

## References


1. Doshi, K.; Cannone, J.; Cobaugh, C.; Gutell, R. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinform.* **2004**, *5*, 105, doi:10.1186/1471-2105-5-105.
2. Proctor, J.R.P.; Meyer, I.M. CoFold: An RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Res.* **2013**, *41*, e102, doi:10.1093/nar/gkt174.
3. Amman, F.; Bernhart, S.H.; Doose, G.; Hofacker, I.L.; Qin, J.; Stadler, P.F.; The Students of the Bioinformatics II Lab Class 2013; Will, S. The Trouble with Long-Range Base Pairs in RNA Folding. In *Advances in Bioinformatics and Computational Biology, 8th BSB*; Setubal, J.C., Almeida, N.F., Eds.; Lect. Notes Comp. Sci.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8213, pp. 1–11, doi:10.1007/978-3-319-02624-4_1.
4. Hofacker, I.L.; Priwitzer, B.; Stadler, P.F. Prediction of Locally Stable RNA Secondary Structures for Genome-Wide Surveys. *Bioinformatics* **2004**, *20*, 191–198, doi:10.1093/bioinformatics/btg388.
5. Kiryu, H.; Kin, T.; Asai, K. Rfold: An exact algorithm for computing local base pairing probabilities. *Bioinformatics* **2008**, *24*, 367–373, doi:10.1093/bioinformatics/btm591.
6. Lange, S.J.; Daniel, M.; Möhl, M.; Joshua, J.N.; Brown, C.M.; Backofen, R. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.* **2012**, *40*, 5215–5226, doi:10.1093/nar/gks181.
7. Hofacker, I.L.; Fontana, W.; Stadler, P.F.; Bonhoeffer, L.S.; Tacker, M.; Schuster, P. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.* **1994**, *125*, 167–188, doi:10.1007/BF00818163.
8. Lorenz, R.; Bernhart, S.H.; Höner zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P.F.; Hofacker, I.L. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **2011**, *6*, 26, doi:10.1186/1748-7188-6-26.
9. Turner, D.H.; Mathews, D.H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucl. Acids Res.* **2010**, *38*, D280–D282, doi:10.1093/nar/gkp892.
10. Do, C.; Woods, D.; Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **2006**, *22*, e90–e98, doi:10.1093/bioinformatics/btl246.
11. Lu, Z.J.; Gloor, J.W.; Mathews, D.H. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA* **2009**, *15*, 1805–1813, doi:10.1261/rna.1643609.
12. Bernhart, S.; Hofacker, I.L.; Stadler, P.F. Local RNA Base Pairing Probabilities in Large Sequences. *Bioinformatics* **2006**, *22*, 614–615, doi:10.1093/bioinformatics/btk014.
13. Nussinov, R.; Piecznik, G.; Griggs, J.R.; Kleitman, D.J. Algorithms for Loop Matching. *SIAM J. Appl. Math.* **1978**, *35*, 68–82, doi:10.1137/0135006.
14. Ding, Y.; Chan, C.Y.; Lawrence, C.E. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* **2005**, *11*, 1157–1166, doi:10.1261/rna.2500605.
15. Hamada, M.; Kiryu, H.; Sato, K.; Mituyama, T.; Asai, K. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics* **2009**, *25*, 465–473, doi:10.1093/bioinformatics/btn601.
16. Hamada, M.; Sato, K.; Asai, K. Prediction of RNA secondary structure by maximizing pseudo-expected accuracy. *BMC Bioinform.* **2010**, *11*, 586, doi:10.1186/1471-2105-11-586.
17. Lorenz, R.; Hofacker, I.L.; Stadler, P.F. RNA Folding with Hard and Soft Constraints. *Algorithms Mol. Biol.* **2016**, *11*, 8, doi:10.1186/s13015-016-0070-z.
18. Le, S.Y.; Maizel, J.V., Jr. A method for assessing the statistical significance of RNA folding. *J. Theor. Biol.* **1989**, *138*, 495–510, doi:10.1016/s0022-5193(89)80047-5.
19. Rivas, E.; Eddy, S.R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **2000**, *16*, 583–605, doi:10.1093/bioinformatics/16.7.583.
20. Clote, P.; Ferré, F.; Kranakis, E.; Krizanc, D. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA* **2005**, *11*, 578–591, doi:10.1261/rna.7220505.
21. Freyhult, E.; Gardner, P.P.; Moulton, V. A comparison of RNA folding measures. *BMC Bioinform.* **2005**, *6*, 241, doi:10.1186/1471-2105-6-241.
22. Washietl, S.; Hofacker, I.L.; Stadler, P.F. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 2454–2459, doi:10.1073/pnas.0409169102.

23. Gruber, A.R.; Findeiß, S.; Washietl, S.; Hofacker, I.L.; Stadler, P.F. RNAz 2.0: improved noncoding RNA detection. *Pac. Symp. Biocomput.* **2010**, *15*, 69–79, doi:10.1142/9789814295291_0009.
24. Klein, R.J.; Misulovin, Z.; Eddy, S.R. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7542–7547, doi:10.1073/pnas.112063799.
25. Jiang, M.; Anderson, J.; Gillespie, J.; Mayne, M. uShuffle: A useful tool for shuffling biological sequences while preserving the *k*-let counts. *BMC Bioinform.* **2008**, *9*, 192, doi:10.1186/1471-2105-9-192.
26. Ciesiolka, A.; Jazurek, M.; Drazkowska, K.; Krzyzosiak, W.J. Structural Characteristics of Simple RNA Repeats Associated with Disease and their Deleterious Protein Interactions. *Front. Cell. Neurosci.* **2017**, *11*, 97, doi:10.3389/fncel.2017.00097.