RNAHeliCes – Folding space analysis based on position aware structure abstraction

Jiabin Huang and Björn Voß
Genetics & Experimental Bioinformatics
Institute of Biology 3
University of Freiburg
{jiabin.huang, bjoern.voss}@biologie.uni-freiburg.de

Abstract:

RNA has many pivotal functions especially in the regulation of gene expression by ncRNAs. Identification of their structure is an important requirement for understanding their function. Structure prediction alone is often insufficient for this task, due to algorithmic problems, parameter inaccuracies and also due to biological peculiarities. Among the latter we have base modifications, co-transcriptional folding leading to folding traps and conformational switching as in the case of riboswitches. All these require more in-depth analysis of the folding space. The major drawback, which all methods have to cope with, is the exponential growth of the folding space. Therefore, methods are often limited in the sequence length they can analyze or make use of heuristics, sampling or abstraction. Our approach adopts the abstraction strategy and remedies some problems of existing methods.

We introduce a position-specific abstraction based on helices which we term helix index shapes or *hishapes* for short. Utilizing a dynamic programming framework, we have implemented this abstraction in the program *RNAHeliCes*. Furthermore, we present results of empirical studies on the size of the search space, showing that it is a reasonable extension to existing methods. We further show the application of *RNAHeliCes* to some well-studied classes of RNAs and discuss possible use cases of our method.

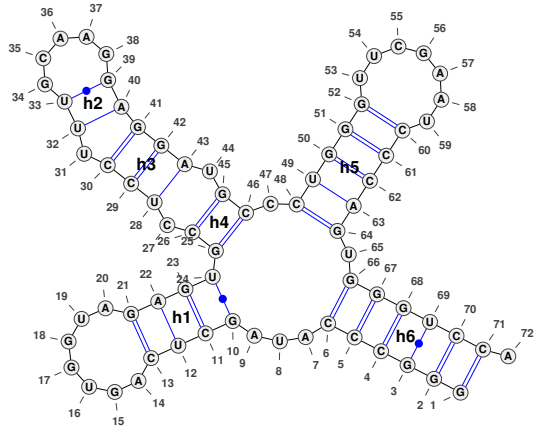
RNAHeliCes is available for download at http://www.cyanolab.de/software/RNAHeliCes.htm.

1 Introduction

Recent advances in research on RNA have led to a change in perspective regarding the role of RNA. It becomes increasingly clear that RNA has many pivotal functions, especially in the regulation of gene expression by non-coding RNAs (ncRNAs) and as *cis*-regulatory RNA elements. Generally, the correct exertion of a ncRNAs function depends on the proper formation of its structure. This is usually not a big deal for the RNA *in vivo*, but holds a lot of surprises for *in silico* analyses, which renders structure prediction an error-prone task. Beyond peculiarities of the folding process, also functional characteristics of a ncRNA may need more elaborate studies than predicting one minimum free energy (*mfe*) structure. Bistable RNAs and riboswitches, for example, can only be found when in addition to the optimal structure, suboptimal structures are considered. In general, it is often useful to analyze the folding space of a ncRNA as this gives deeper insight into structural properties. Unfortunately this does not come without a cost, which is the complexity and size of the folding space. It grows exponentially with sequence length and corresponds to a multidimensional space. Nevertheless, methods exist which can be used to carry out detailed analyses of the folding space.

Suboptimal structure prediction, with the enumeration of rigorously all possible secondary structures, is available with RNAsubopt [WFHS99]. This constitutes the most basic method for folding space analysis. Several applications exist which rely on RNAsubopt, e.g. paRNAss [VMG04] and barriers [FHSW02]. The latter is a general purpose tool which computes local minima together with the energy barriers separating them from each other. Especially the computation of energy barriers is an important feature, as this allows to draw conclusions about bi-stability, folding traps or structural well-definedness. Unfortunately, this requires a complete folding space enumeration, at least within a reasonable enrgy range above the *mfe*, which makes this approach computationally very expensive.

Approaches to overcome the exponential explosion commonly make use of heuristics for deriving the path of intermediate structures. A different approach to folding space analysis was founded with the development of shape abstraction [GVR04]. This method provides a means to partition the folding space into classes of similar structures. Together with features, such as their probabilities, shapes provide an overview of the folding space. Due to intrinsic features, shape



helix name	$h_t(i,j,k)$	$hi(h_t)$
h1	$h_{hl}(10,24,4)$	17
h2	$h_{hl}(32,40,2)$	36
h3	$h_{bl}(28,43,3)$	35.5
h4	$h_{il}(25,46,2)$	35.5
h5	$h_{hl}(48,64,5)$	56
h6	$h_{ml}(1,71,6)$	36

Figure 1: Example secondary structure and properties of its helical regions.

abstraction does not give direct access to energy barriers or estimates thereof. But, this can be achieved when combining shape abstraction with path heuristics as presented in [BMV+10]. A major drawback of shape abstraction, as it is implemented so far, is the position independence of the abstraction mappings. A single hairpin at the 5'-end has the same "[]"- shape as one at the 3'-end. As a consequence shape classes encompass structurally similar but perhaps functionally unrelated structures. This can only be overcome by a new abstraction function.

In this contribution we introduce a position specific abstraction based on helices, which we term *hishape*. We further show its application to some well-studied classes of RNAs and discuss potential use cases.

RNAHeliCes is distributed as a free software available at http://www.cyanolab.de/software/RNAHeliCes.htm.

2 Results

2.1 Defining helix index shapes

In the following, we provide formal definitions for the new abstraction based on helix indices.

Definition 1 (secondary structure of RNA) *A secondary structure of an RNA sequence* $R = r_1 r_2 r_3 \dots r_n$ $(r_i \in \{A, C, G, U\}, i = 1, 2, 3, \dots, n)$ *of length* n *is defined as a set of base pairs* $P = \{(r_i, r_j), \dots\}$*, where*

$$1 \leq i \leq j \leq n \tag{1}$$

$$j - i \geq 3 \tag{2}$$

$$(r_i, r_j) \in (A,U), (U,A), (G,C), (C,G), (G,U), (U,G) \tag{3}$$

Considering only secondary structures without pseudoknots, every base pairs (i,j) *and* (k,l) *in* P *must satisfy the following constraint:*

$$k < i < j < l \parallel k < l < i < j \tag{4}$$

Definition 2 (helix and helix index) *A helix is a series of stacking base pairs and can be denoted by* $h(i,j,k)$ *where* i *and* j *are the bases of the outermost base pair and* k *is the length of the helix (i.e. the number of base pairs). For sequence* R *and helix* $h(i,j,k)$ *it must hold* $\forall_{t=0\dots k-1}$*:* $(r_{i+t}, r_{j-t}) \in P$*. For a helix* $h(i,j,k)$*, we define the helix index of* h *as* $hi(h(i,j,k)) = (i+j)/2$ *which is the center of the helix (see Figure 1).*

Note that the same helix index may represent different helices, but each helix can only be assigned one helix index. The relationship between helix index and helix is 1-to-n.

The secondary structure of any RNA molecule can be broken down into five types of loops that are closed by helices. These are hairpin-, bulge-, internal-, stacking-, and multiloop (denoted as HL, BL, IL, SL and ML). Stacking loops

are special in our case as they elongate helices and can not introduce new ones. Thus, a helix can only be of type HL, BL, IL or ML. According to the loop they enclose, they are denoted as $h_{hl}(i,j,k)$, $h_{bl}(i,j,k)$, $h_{il}(i,j,k)$, $h_{ml}(i,j,k)$, respectively. Any RNA secondary structure is a series of helices of the four types interrupted by unpaired regions, denoted as $H = \{h_t(i,j,k), ...\}$, where $t \in \{hl, bl, il, ml\}$.

Using the function hi from Definition 2, we can map H to a list of helix indices which we term *hishape* (short for *helix index shape*). Similar to abstract shapes, we define different abstraction levels. Let π_h, π_m, π_b be mapping functions considering only HL, HL and ML, and all helices, respectively. Using these to map the same secondary structure H results in different *hishapes*.

$$\pi_h(H) = \pi_h(\{h_t(i,j,k), ...\}(t \in \{hl, bl, il, ml\})) = \{hi(h_t), ...\}(t \in \{hl\}) \tag{5}$$

For π_m and π_b, *hishapes* may be ambiguous since multiloop and symmetric internal loop helices can have helix indices equal to their enclosed hairpin loop helices. Therefore, the letter 'm' is attached to the end of helix indices derived from $h_{ml}(i,j,k)$ in π_m as well as in π_b, while the letter 'b' denotes helix indices derived from $h_{bl}(i,j,k)$ and $h_{il}(i,j,k)$ in π_b.

$$\begin{aligned} \pi_m(H) &= \pi_m(\{h_t(i,j,k), ...\}(t \in \{hl, bl, il, ml\})) = \{hi(h_t) +' x', ...\}(t \in \{hl, ml\}) \\ \text{where } x &= \begin{cases} \varnothing & \text{if } t \in \{hl\} \\ m & \text{if } t \in \{ml\} \end{cases} \end{aligned} \tag{6}$$

$$\begin{aligned} \pi_b(H) &= \pi_b(\{h_t(i,j,k), ...\}(t \in \{hl, bl, il, ml\})) = \{hi(h_t) +' x', ...\}(t \in \{hl, bl, il, ml\}) \\ \text{where } x &= \begin{cases} \varnothing & \text{if } t \in \{hl\} \\ m & \text{if } t \in \{ml\} \\ b & \text{if } t \in \{bl, il\} \end{cases} \end{aligned} \tag{7}$$

Note that these mapping functions do not preserve the nesting pattern of loops.

Definition 3 (hishape space) *Let $F(s)$ be the folding space of R, i.e. the set of all legal structures according to Definition 1. Then, the hishape space is defined as $P(s) = \{\pi(x) | x \in F(s)\}$*

Definition 4 (hishape representative) *Hishapes do not only filter similar secondary structures, but rather partition the folding space into classes of secondary structures. The member with minimum free energy is defined as the hishape representative and termed hishrep.*

2.2 Implementing *RNAHeliCes*

Dynamic programming (DP) algorithms can be implemented in any programming language, e.g. C or Java. In order to circumvent implementation specific problems, e.g. index errors, and to take advantage of already existing code the new abstraction was implemented using the algebraic dynamic programming (ADP) framework [Gie00, GM02].

The ADP framework, initially implemented in the functional programming language Haskell, was recently reimplemented in C++ and termed Bellman's GAP [SJG11, GS11]. Both implementations split a DP algorithm into a grammar and several algebras. The grammar is a set of rules describing the candidates of the search space while algebras evaluate these candidates. In the case of RNA structure analysis algebras for energy minimization, partition function [McC90] calculation, pretty printing of the structure in dot-bracket-format and others exist. Algebras can be combined using product operations which allows complex analyses to be built in a rather simple way. The thermodynamic parameters [XSB$^+$98, MSZT99, ST00] used within Bellman's GAP are taken from the Vienna RNA package [HFS$^+$94].

We use the same grammar that is used in *RNAshapes*, which handles dangling bases in an unambiguous fashion [VGR06]. In *RNAHeliCes*, for each candidate defined by the grammar, we are interested in computing the *hishape*,

algebra function	π_h	π_m	π_b
HL(i,l,j)	$(i+j)/2$	$(i+j)/2$	$(i+j)/2$
SL(i,x,j)	$\pi_h(x)$	$\pi_m(x)$	$\pi_b(x)$
BL(i,r,x,j)	$\pi_h(x)$	$\pi_m(x)$	$\pi_b(x).(i+j)/2.'b'$
BL(i,x,l,j)	$\pi_h(x)$	$\pi_m(x)$	$\pi_b(x).(i+j)/2.'b'$
IL(i,r,x,l,j)	$\pi_h(x)$	$\pi_m(x)$	$\pi_b(x).(i+j)/2.'b'$
ML(i,x,j)	$\pi_h(x)$	$\pi_m(x).(i+j)/2.'m'$	$\pi_b(x).(i+j)/2.'m'$

Table 1: Algebra functions for *hishape* analysis. '.' = string concatenation, 'x' = enclosed substructure, 'r' and 'l' = unpaired regions.

the free energy, the dot-bracket-representation and the partition function contribution of this *hishape*. We can reuse existing algebras for computation of free energy, partition function and dot-bracket-representation. The algebra for computation of *hishapes* was developed by ourselves and implemented as described in the following.

Algebra $hishape$ has three different abstraction levels according to the three mapping functions π_h, π_m and π_b. Table 1 shows, how this is reflected within the algebra functions for the different loop/helix types. The choice function h removes candidates with equal *hishape* resulting in non-redundant answer lists. Our goal is to compute the k best (in terms of energy) *hishapes* together with their free energy, partition function contribution and the *hishrep* in dot-bracket notation. Reusing the algebras mfe for free energy calculation, p_func for partition function values and $pretty$ for the dot-bracket-representation, in GAP we achieve this with the algebra product: $hishape \otimes (mfe \times p\_func) * pretty$, where $\otimes$ is the interleaved, $\times$ the cartesian and $*$ the lexicographic product operation. Details about these product types can be found in [SJG11, GS11].

The time complexity of the algorithm using the product above is $O(k^2n^3)$ where k is the number of *hishape* classes and n the sequence length. Because of the exponential growth of *hishape* classes, $k \approx \alpha^n$ where α depends on the mapping function π (see Section 2.3), the time complexity would be $O(\alpha^n)$. However, an implementation returning only the k-best (for example $k = 100$) *hishape* classes reduces the overall complexity to $O(n^3)$.

2.3 Size of the *hishape* space

While in *RNAshapes* the position of a helix is completely ignored, *RNAHeliCes* keeps track of these resulting in a more fine-grained abstraction. Thus, the *hishape* space is likely to be bigger than the shape space. The results of empirical measurments for the growth of the *hishape* space, structure space and shape space with increasing sequence length are shown in Figure 2. Two main characteristics become apparent. First, all spaces grow exponentially with sequence length, with the structure space being the fastest followed by *hishapes* and abstract shapes show the slowest growth. An exception is π_h for which the growth is slower than for the least abstract shape type 1. Secondly, comparing the three *hishape* spaces shows that the fewer the helix types considered the slower the growth.

Recently precise asymptotics for the number of abstract shapes have been derived [LPC08, NS09]. They follow the formula $\beta \cdot \alpha \cdot n^{-3/2}$ where α and β depend on the shape type, e.g. for shapes of type 5 $\beta = 1.20259^n$ and $\alpha = 5.12777$. Note that these asymptotics have been derived for structures of length n, thus, disregarding constraints imposed by the sequence, namely base pairing. Nevertheless, we used this formula to fit functions to our data, see Figure 2. For the *hishape* spaces we derive the following numbers $\beta_{\pi_h} = 1.2314249^n$, $\beta_{\pi_m} = 1.3186567^n$ and $\beta_{\pi_b} = 1.3808570^n$. Notably, we achieved R^2-values > 0.97 for all fits. Expectedly, for abstract shapes the values of α and β differ from the ones given in [NS09], e.g. for β of shape type 5 we have 1.1330641^n compared to 1.20259^n. This simply reflects the influence of the base-pairing constraint on the number of solutions. The good correlation of the data to the fit functions for *hishapes* lets us assume, that asymptotic numbers for sequences of length n according to the general formula $\beta \cdot \alpha \cdot n^{-3/2}$ can be derived straightforward.

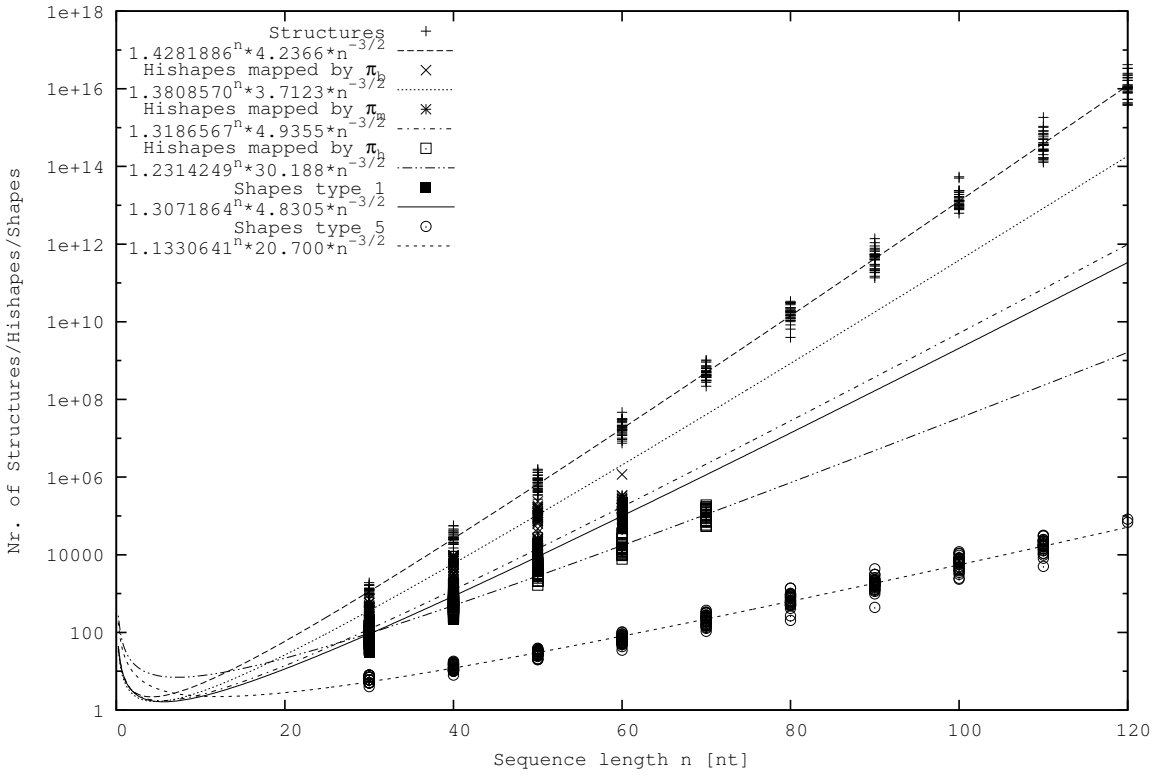


Figure 2: Comparison of secondary structure space, *hishape* spaces mapped by π_b, π_m and π_h and shape spaces (type 1 and 5).

```
AACUAAAACAAUUUUUGAAGAACAGUUUCUGUACUUCAUUGGUAUGUAGAGACUUC     E  hishape    P    shape
..((...((((((..(((((.((((...)))).)))))..))))..))))..)).....  -10.70  27   0.897904  []
.......................((((((((((((.....))))))..))))))))..   -9.00  38   0.063473  []
```

Figure 3: The alternating structures of the spliced leader RNA from *L.collosoma* with their free energy (E in $kcal/mol$), *hishapes*, *hishape* probabilities (P) and their shape.

2.4 Applications

In the following we want to show some applications of *RNAHeliCes* to well-studied RNAs.

Spliced Leader RNA from *Leptomonas collosoma*

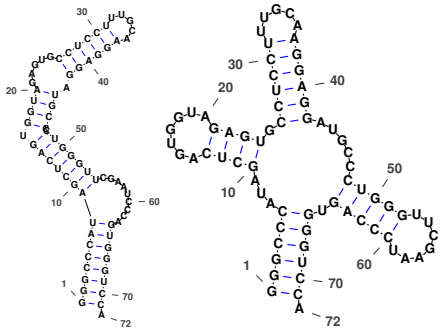
The Spliced Leader RNA from *Leptomonas collosoma* [LC93] has two alternating structures of nearly equal free energy, differing only by 1.7 kcal/mol. Figure 3 shows the results of shape and *hishape* analysis. While the two π_m *hishapes* ([27] and [38]) reflect the two conformations of the Spliced Leader RNA, *RNAshapes* yields the same abstract shape "[]" for both conformations. The probability of the "[]"-shape is 0.961912, and contributions to this come from both conformations. This example shows that for certain applications shape abstraction might be too strong and, perhaps more important, shape features, such as the shape probability, are computed over very diverse, rather than similar, structures. Conversely, *hishapes* hold position-specific structure information. In this way a more fine-grained overview of the structure space can be obtained. The probabilities of conformation 1 and conformation 2 are 0.897904 and 0.063473, respectively, and are in good agreement with the bistable character of this RNA.

tRNA, microRNA precursor and snoRNA

In order to give an impression about analyses using RNAHeliCes we will show the results for three example sequences, namely the *Natronomonas pharaonis* tRNA for alanine [KDSU$^+$97] (Database ID: tdbD00000012) from the Transfer RNA database [JMH$^+$09], the precursor of *Homo sapiens* microRNA miR-507 [BAK$^+$05] (miRBase accession no.

	hishape,π_m	E	P
1	34	-35.9	0.989670
2	34,56,36m	-32.2	0.007853
3	**17,34,56,36m**	**-31.7**	**0.001184**
4	34,56.5,36m	-31.1	0.000473
5	34,61.5,36m	-30.9	0.000408
$rank$(*hishape*)	3		
$rank$(subopt)	104		

(a)

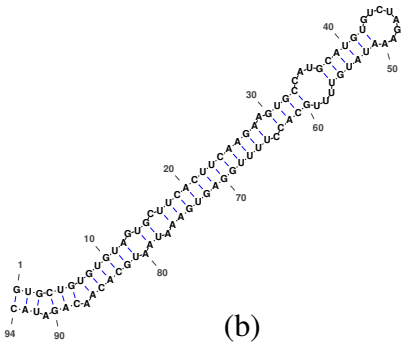


(b)

Figure 4: (a) Five best π_m *hishapes* for *Natronomonas pharaonis* tRNA-alanine. E is free energy in $kcal/mol$, P is *hishape* probability. $rank$ gives the position of the native structure in the result list. (b) Secondary structures of the 1st and 3rd *hishape*.

	hishape,π_m	E	P
1	**46.5**	**-37.9**	**0.999622**
2	47.5	-32.1	0.000293
3	49.5	-32.0	0.000070
4	7,46.5	-30.4	0.000002
5	12,46.5,43m	-30.1	0.000001
$rank$(*hishape*)	1		
$rank$(subopt)	1		

(a)



(b)

Figure 5: (a) Five best *hishapes* for *Homo sapiens* miR-507; (b) Secondary structure of the 1st *hishape*.

MI0003194) from miRBase [KGJ11] and the *Mus musculus* U73a snoRNA from [RTLP98].

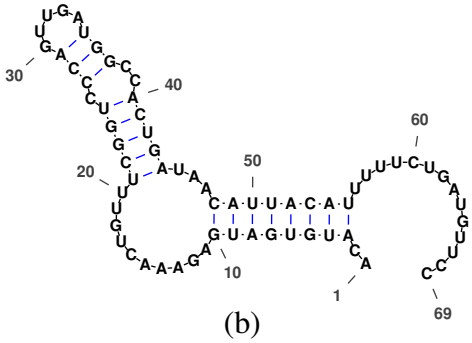
Figures 4-6 show that, in general, by using *RNAHeliCes* fewer suboptimal solutions are required to reveal the native structure compared to *RNAsubopt* [HFS$^+$94] with the most striking difference for the *Natronomonas pharaonis* tRNA (72 nt). Instead of requiring to check 104 structures with *RNAsubopt*, only three need to be checked using *RNAHeliCes*. Surprisingly, the cloverleaf *hishape* has a probability of only 0.001184, which may be due to post-transcriptional modifications [Bjö95, Agr96, HA04]. Interestingly, the anticodon-loop helix with helix index 34 is present in all *hishapes* showing that it is a well-defined structural element. Summing up the probabilities would result in a joined probability of 0.999588 for this helix.

MicroRNA genes are usually transcribed as a primary miRNA (pri-miRNA). The primary transcript is further processed to a precursor miRNA (pre-miRNA) with a characteristic stem-loop structure [LKH$^+$04]. The native structure of *Homo sapiens* precursor micoRNA miR-507 (94 nt) is a long stem with a terminal hairpin-loop, as shown in Figure 5(b). The corresponding helix is $h_{hl}(37, 56, 6)$ with a helix index of 46.5 as predicted for *hishape* 1 (see Figure 5(a)). Noteworthy, *hishapes* 2 and 3 carry helices which are close "neighbors" to that of *hishape* 1. *Hishapes* 4 and 5 have additional helices $3'$ and/or $5'$ to helix 46.5. Altogether, this shows the molding of the folding space to ensure the functional structure where miRNA and miRNA* are paired to form a proper Dicer substrate.

SnoRNAs are a class of nucleolus-located RNAs that guide chemical modifications of other types of RNAs, e.g. small nuclear RNAs (snRNAs). Two types of snoRNAs exist, box C/D and H/ACA type, each having their characteristic secondary structure. The *Mus musculus* U73a snoRNA (69 nt) is of the box C/D-type and contains a C, D and D$'$ box [RTLP98]. The first *hishrep* (see Figure 6) shows good correspondence to the native structure regarding this crucial D$'$ box. Further analyses using π_b reveals that *hishrep* 2 is equal to the native structure (*hishape*, π_b:32,32.5b,33b,35b,33.5b; $E = -10.0\ kcal/mol$; $P = 0.050530$).

	hishape,π_m	E	P
1	**32**	**-10.7**	**0.543800**
2	20,35.5,29m	-10.0	0.085077
3	32,59	-9.9	0.043850
4	17,35.5,29m	-9.8	0.026664
5	20,36,29m	-9.7	0.046650
$rank$(*hishape*)	1 (slight difference in bulge pattern)		
$rank$(subopt)	4		

(a)



(b)

Figure 6: (a) Five best *hishapes* for *Mus musculus* U73a snoRNA; (b) Secondary structure of the 1st *hishape*.

3 Discussion

In the present paper we introduce the concept of *hishapes* which is closely related to the idea of abstract shapes. Briefly, we provide a new mapping function and preserve all functionalities of shape analysis. Among these are search space reduction by (hi)shape filtering and probabilistic analysis based on (hi)shape classes. Compared to abstract shapes the major advantage of the new abstraction is its position-specificity, which provides a better resolution especially for short RNAs. The cost for this is a slightly increased search space, which is still much smaller than the structure space. Setting up relations between *hishapes* is as easy as looking for common helix indices. This, together with the reduced search space, renders *RNAHeliCes* a promising candidate for folding space analysis. For instance, *hishapes* representing the union or an intersection of two *hishapes* represent intermediates on the interconversion pathway between the corresponding *hishreps*. Thus, they could be used as anchors in the computation of interconversion pathways, which could provide a significant speed-up for tasks such as energy barrier calculation. A similar idea using abstract shapes is described in [BMV$^+$10].

Besides this general functionality in folding space analysis, the example applications show that more specific use cases might be reasonable. Straightforward is the use of *hishape* probabilities as a measure of structural well-definedness. The tRNA example within this manuscript shows that the helix-based approach might allow to expand this to structural elements, such as the anticodon-loop of tRNAs. Complementary, the difference in free energy to the next *hishape* might be useful. Such measures can be used as filters in the process of *de-novo* miRNA prediction as proposed by Xue et al. [XLH$^+$05]. Their algorithm discovers new pre-miRNAs in whole genomes without utilizing comparative genomic information. It relies on identifying hairpin structures which, unfortunately, are numerous, resulting in a high number of false positives. Here the above mentioned measure of structural well-definedness may provide a means of filtering out pseudo miRNA precursors.

Hishapes can also be used for comparative structure analysis between different RNAs. For instance a distance measure based on the sum of positional distances of the helices of two *hishapes* provides a fast method for structure comparison. It further might be used for the identification of common structures of two or more RNAs where the objective is to find the set with a minimum sum of pairwise distances. This would provide an alternative to the prediction of consensus shapes [RG05].

Altogether we feel, that *hishape* based abstraction provides a valuable tool for various applications in RNA secondary structure analysis. Furthermore, modified abstractions based on other helix features may be useful and extend the range of applications for our method.

4 Acknowledgments

We gratefully acknowledge the contributions of Georg Sauthoff to the efficient implementation of Bellman's GAP. His example GAP code for folding space analysis served as a starting point for our research. This work was supported by the Deutsche Forschungsgemeinschaft (grant Vo 1450/2-1 to B.V.).

References

[Agr96] P.F. Agris. The importance of being modified: roles of modified nucleosides and Mg2+ in RNA structure and function. *Progress in nucleic acid research and molecular biology*, 53:79–129, 1996.

[BAK+05] Isaac Bentwich, Amir Avniel, Yael Karov, Ranit Aharonov, Shlomit Gilad, Omer Barad, Adi Barzilai, Paz Einat, Uri Einav, Eti Meiri, Eilon Sharon, Yael Spector, and Zvi Bentwich. Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics*, 37(7):766–770, June 2005.

[Bjö95] G.R. Björk. Genetic Dissection of Synthesis and Function of Modified Nucleosides in Bacterial Transfer RNA. *Progress in nucleic acid research and molecular biology*, 50:263–338, 1995.

[BMV+10] Sergiy Bogomolov, Martin Mann, Björn Voß, Andreas Podelski, and Rolf Backofen. Shape-based Barrier Estimation for RNAs. In Dietmar Schomburg and Andreas Grote, editors, *GCB*, volume 173 of *LNI*, pages 41–50. GI, 2010.

[FHSW02] C. Flamm, I.L. Hofacker, P.F. Stadler, and M.T. Wolfinger. Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie*, 216(2/2002):155, 2002.

[Gie00] R. Giegerich. A systematic approach to dynamic programming in bioinformatics. *Bioinformatics*, 16(8):665, 2000.

[GM02] Robert Giegerich and Carsten Meyer. Algebraic dynamic programming. In *Algebraic Methodology And Software Technology, 9th International Conference, AMAST 2002*, pages 349–364. Springer LNCS, 2002.

[GS11] Robert Giegerich and Georg Sauthoff. Yield grammar analysis in the Bellman's GAP compiler. In *Proceedings of the Eleventh Workshop on Language Descriptions, Tools and Applications*, LDTA '11. ACM, 2011.

[GVR04] R. Giegerich, B. Voß, and M. Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Research*, 32(16):4843, 2004.

[HA04] Mark Helm and Giuseppe Attardi. Nuclear Control of Cloverleaf Structure of Human Mitochondrial tRNALys. *Journal of Molecular Biology*, 337(3):545 – 560, 2004.

[HFS+94] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian L. Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, 125:167–188, 1994.

[JMH+09] Frank Jühling, Mario Mörl, Roland K. Hartmann, Mathias Sprinzl, Peter F. Stadler, and Joern Pütz. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Research*, 37(Database issue):D159–162, January 2009.

[KDSU+97] M. Kamekura, M.L. Dyall-Smith, V. Upasani, A. Ventosa, and M. Kates. Diversity of Alkaliphilic Halobacteria: Proposals for Transfer of Natronobacterium vacuolatum, Natronobacterium magadii, and Natronobacterium pharaonis to Halorubrum, Natrialba, and Natronomonas gen. nov., respectively, as Halorubrum vacuolatum comb. nov., Natrialba magadii comb. nov., and Natronomonas pharaonis comb. nov., respectively. *International Journal of Systematic and Evolutionary Microbiology*, 47(3):853, 1997.

[KGJ11] Ana Kozomara and Sam Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39(suppl 1):D152–D157, January 2011.

[LC93] K.A. LeCuyer and D.M. Crothers. The Leptomonas collosoma spliced leader RNA can switch between two alternate structural forms. *Biochemistry*, 32(20):5301–5311, 1993.

[LKH+04] Y. Lee, M. Kim, J. Han, K.H. Yeom, S. Lee, S.H. Baek, and V.N. Kim. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20):4051–4060, 2004.

[LPC08] W.A. Lorenz, Y. Ponty, and P. Clote. Asymptotics of RNA shapes. *Journal of Computational Biology*, 15(1):31–63, 2008.

[McC90] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, May 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. *Journal of Molecular Biology*, 288(5):911–940, May 1999.

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

[RG05] Jens Reeder and Robert Giegerich. Consensus shapes: An alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, 21(17):3516–3523, 2005.

[RTLP98] A. Rebane, R. Tamme, M. Laan, and I. Pata. A novel snoRNA (U73) is encoded within the introns of the human and mouse ribosomal protein S3a genes. *Gene*, 210(2):255–263, 1998.

[SJG11] Georg Sauthoff, Stefan Janssen, and Robert Giegerich. Bellman's GAP - A Declarative Language for Dynamic Programming. In *Proceedings of 13th International ACM SIGPLAN Symposium on Principles and Practice of Declarative Programming*, PPDP '11. ACM, 2011.

[ST00] Susan J. Schroeder and Douglas H. Turner. Factors Affecting the Thermodynamic Stability of Small Asymmetric Internal Loops in RNA. *Biochemistry*, 39(31):9257–9274, 2000.

[VGR06] Björn Voss, Robert Giegerich, and Marc Rehmsmeier. Complete probabilistic analysis of RNA shapes. *BMC Biol*, 4(1), February 2006.

[VMG04] B. Voss, C. Meyer, and R. Giegerich. Evaluating the predictability of conformational switching in RNA. *Bioinformatics*, 20(10):1573–1582, July 2004.

[WFHS99] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165, February 1999.

[XLH+05] C. Xue, F. Li, T. He, G. P. Liu, Y. Li, and X. Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6, 2005.

[XSB+98] Tianbing Xia, John SantaLucia, Mark E. Burkard, Ryszard Kierzek, Susan J. Schroeder, Xiaoqi Jiao, Christopher Cox, and Douglas H. Turner. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs. *Biochemistry*, 37(42):14719–14735, October 1998.