Random k-noncrossing RNA Structures

William Y.C. Chen , Hillary S. W. Han , and Christian M. Reidys *

*Center for Combinatorics, LPMC-TJKLC
Nankai University
Tianjin 300071
P.R. China
Phone: *86-22-2350-6800
Fax: *86-22-2350-9272

duck@santafe.edu

Submitted to Proceedings of the National Academy of Sciences of the United States of America

In this paper we introduce a combinatorial framework which provides an interpretation of RNA pseudoknot structures as sampling paths of a Markov process. Our results give insight into the cross-serial interactions of RNA pseudoknot structures, i.e. their crossings. They facilitate a variety of applications ranging from the energy based sampling of pseudoknot structures as well as the *ab initio* folding via hidden Markov models. Our main result is an algorithm which generates RNA pseudoknot structures with uniform probability in linear time. This algorithm serves as a stepping stone to sequence specific as well as energy based transition probabilities. The approach employs a correspondence between pseudoknot structures, parametrized in terms of the maximal number of mutually crossing arcs and certain tableau sequences. The latter can be viewed as lattice paths, whose generating functions are shown to be D-finite. The main idea of this paper is to view each such lattice path as a sampling path of a stochastic process and to make use of D-finiteness for the efficient computation of the corresponding transition probabilities.

RNA pseudoknot structure | k-noncrossing structure | uniform generation | tableau | lattice path

Abbreviations:

Pseudoknots have long been known as important structural elements [34], see Fig. 1. These cross-serial interactions between RNA nucleotides are functionally important in tRNAs, RNaseP [20], telomerase RNA [28], and ribosomal RNAs [18]. Pseudoknots in plant virus RNAs mimic tRNA structures, and *in vitro* selection experiments have produced pseudoknotted RNA families that bind to the HIV-1 reverse transcriptase [31]. Import general mechanisms, such as ribosomal frame shifting, are dependent upon pseudoknots [2].

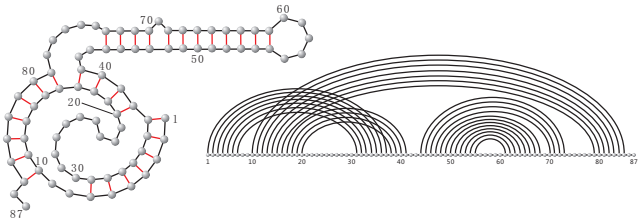

Fig. 1. The Hepatitis Delta Virus (HDV)-pseudoknot structure represented as a planar graph and as a diagram: we display the structure as folded by the *ab initio* folding algorithm cross [10] (left) and the diagram representation (right).

Despite their biological importance, pseudoknots are typically excluded from large-scale computational studies. Although the problem has attracted considerable attention in the last decade, and several software tools [26] have become available, the required resources have remained prohibitive for applications beyond individual molecules. Lyngso *et al.* [22] have shown that the prediction of general RNA pseudoknot structures is NP-complete. In the literature, oftentimes some variant of the dynamic programming (DP) paradigm is employed [26]. This DP-method generates certain subclasses of pseudoknots. We will discuss below that the DP-paradigm is ideally suited for an inductive, or context-free, structure-class. However, due to the cross-serial bonds RNA pseudoknot structures cannot be recursively generated. Accordingly, we *a priori* know that the DP-paradigm is only of limited applicability for RNA pseudoknot structures. In addition, DP-based approaches are not even particularly time efficient, point in case is [26] exhibiting a time complexity of $O(n^6)$. The algorithmic difficulties are confounded by the fact that the thermodynamics of pseudoknots is poorly understood; we suspect that this is at least in part the case because of the well-known difficulties in making use of such information even if it were available.

Within the DP-paradigm, it is unlikely that substantial improvements can still be made. Here, we introduce the mathematical framework for a completely different view on pseudoknotted structures that is not based on recursive decomposition, i.e., parsing w.r.t. to (some extension of) context-free grammars (CFG). The approach that we take here is based on the observation that pseudoknotted RNA structures are in a natural way related to well-understood combinatorial objects. The key algorithmic innovation is a Markov process that efficiently generates pseudoknotted structures with a uniform measure. Biophysical realism can be added by modifying the transition rates of this fundamental Markov process.

In order to put our approach into context, let us give a retrospective overview. Three decades ago Waterman *et al.* [33, 24, 13] analyzed RNA secondary structures. Secondary structures are coarse grained RNA contact structures, see Fig. 2.

Reserved for Publication Footnotes

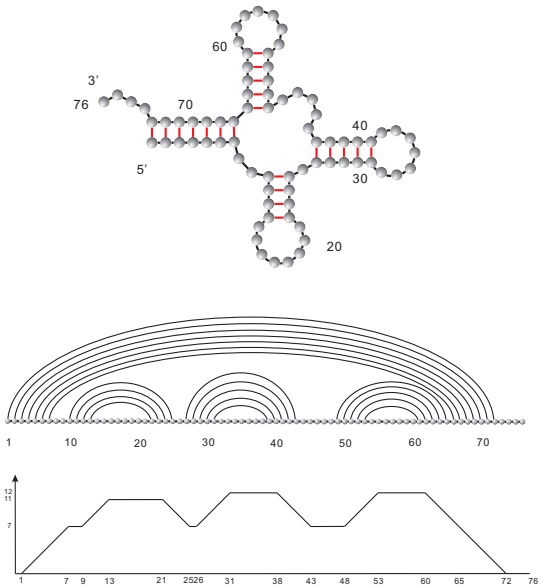

Fig. 2. The phenylalanine tRNA secondary structure represented as a planar graph (top), 2-noncrossing diagram (middle) and Motzkin-path (bottom), where up/down/horizontal-steps correspond to start/end/unpaired vertices, respectively.

They can be represented as diagrams, i.e. labeled graphs over the vertex set $[n] = \{1, \ldots, n\}$ with vertex degrees ≤ 1, represented by drawing its vertices on a horizontal line and its arcs (i, j) $(i < j)$, in the upper half-plane, see Fig. 1 and Fig. 3. Here, vertices and arcs correspond to the nucleotides **A**, **G**, **U** and **C** and Watson-Crick (**A-U**, **G-C**) and (**U-G**) base pairs, respectively.

In a diagram two arcs (i_1, j_1) and (i_2, j_2) are called crossing if $i_1 < i_2 < j_1 < j_2$ holds. Accordingly, a k-crossing is a sequence of arcs $(i_1, j_1), \ldots, (i_k, j_k)$ such that $i_1 < i_2 < \cdots < i_k < j_1 < j_2 < \cdots < j_k$, Fig. 3.

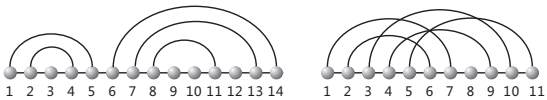

Fig. 3. k-noncrossing diagrams: a noncrossing (left) and a 4-noncrossing diagram (right) containing the three mutually crossing arcs $(1, 7), (4, 9), (5, 11)$.

We call diagrams containing at most $(k-1)$-crossings, k-noncrossing diagrams (k-noncrossing partial matchings). RNA secondary structures have no crossings in their diagram representation, see Fig. 3 (l.h.s.) and Fig. 2, and are therefore 2-noncrossing diagrams.

The efficient minimum free energy (mfe) folding of secondary structures is a consequence of the following relation of the numbers of RNA secondary structures over n nucleotides, $S_2(n)$, [33]

$$S_2(n) = S_2(n-1) + \sum_{j=0}^{n-3} S_2(n-2-j)S_2(j), \qquad \textbf{[1]}$$

where $S_2(n) = 1$ for $0 \leq n \leq 2$. Accordingly, RNA secondary structures satisfy a constructive recursion. As mentioned above, this relation suggests the DP-recursions used for the polynomial time folding of secondary structures [24] and has therefore profound algorithmic implications. The uniform generation of RNA secondary structures is wellknown [32] and can be derived in linear time, using the framework of Flajolet *et al.* [4].

k-noncrossing RNA structures [15, 16], are k-noncrossing diagrams without arcs of the form $(i, i+1)$ and represent a natural generalization. The notion k-noncrossing stipulates that the complexity of a pseudoknot is related to the maximal number of mutually crossing bonds. Indeed, most natural RNA pseudoknots are 3-noncrossing [9]. Due to the cross-serial interactions, the numbers of pseudoknot structures do not satisfy a recursion of the type of eq. (1), rendering the *ab initio* folding into minimum free energy configurations [10, 22] as well as the derivation of detailed statistical properties, a nontrivial task. Indeed, in order to derive statistical properties, the entire space of structures has to be exhaustively generated, which is only possible for small sequence lengths. Only a few statistical results, derived using singularity analysis of the bivariate generating functions are known [17].

There exists no general framework for the uniform generation of elements of a non-inductive combinatorial class. However, in the context of graphs the subject of uniform generation via Markov-processes has been studied. Work on the uniform generation of specific graphs in the context of parallel random access machine (PRAM) can be found in [39] and Jerrum *et al.* [11, 12] studied approximation algorithms in the context of rapidly mixing Markov-chains [1]. We also refer to the paper of Wilf [35] as well as the book [36].

Our approach is as follows: we translate k-noncrossing diagrams into specific lattice walks, see Fig. 4

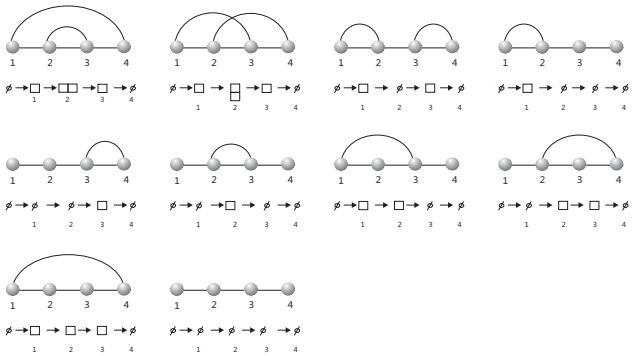

Fig. 4. Translating diagrams into sequences of "shapes". We display all 3-noncrossing diagrams over four vertices and draw their corresponding sequences of shapes underneath.

and view the latter as sampling paths of a stochastic process, see Fig. 5:

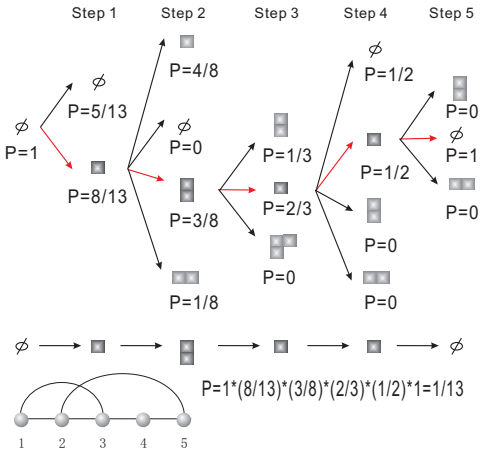

Fig. 5. Uniform generation: the stochastic process over shapes (top), a sampling path (middle) and its pseudoknot structure (bottom). The transition probabilities are computed in Theorem 2 as a pre-processing step.

The key observation is that the generating function of these walks is D-finite or equivalently, P-recursive [29]. This means, there exists a finite recurrence relation with polynomial coefficients, see Corollary 1. Consequently, the *numbers* of these walks can be derived in linear time and these allow us to compute the transition probabilities of the process displayed in Fig. 5. The implication is profound: the transition probabilities can be derived as a pre-processing step in polynomial time after which a pseudoknot structure can be generated uniformly in linear time. Indeed, each structure is generated by the stochastic process having exactly n steps each of which requiring constant time.

From structures to lattice paths and back

In this section we translate RNA pseudoknot structures into lattice paths. For this purpose we introduce shapes, $*$-tableaux and the Robinson-Shensted-Knuth (RSK) algorithm [30].

A shape is a collection of squares arranged in left-justified rows with weakly decreasing number of boxes in each row. A Young tableau is a filling of these squares by numbers which is weakly decreasing in each row and strictly decreasing in each column. A $*$-tableaux of shape λ^n is a sequence of shapes $\varnothing = \lambda^0, \lambda^1, \ldots, \lambda^n$ such that for $1 \leq i \leq n$, λ^i is obtained from λ^{i-1} by either adding/removing one square or doing nothing (hesitating step), see Fig. 6.

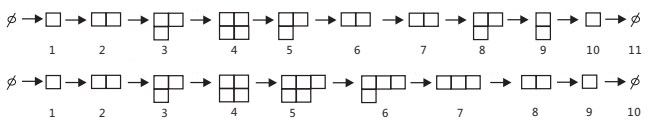

Fig. 6. $*$-tablaux with (top) and without (bottom) hesitating steps. The hesitation step in the top is at $(6, 7)$, In Fig. 9 we show how the top $*$-tableaux induces a unique k-noncrossing structure.

The RSK-algorithm is a procedure which row-inserts elements into a Young tableau, T. Suppose we want to insert k into T. Let $T_{i,j}$ denote the element in the ith row and jth column. Let j be the largest integer such that $T_{1,j-1} \leq k$. (If $T_{1,1} > k$, then $j = 1$.) If $T_{1,j}$ does not exist, then simply add k at the end of the first row. Otherwise, if $T_{1,j}$ exists, then replace $T_{1,j}$ by k. Next insert $T_{1,j}$ into the second row following the above procedure and continue until an element is inserted at the end of a row, see Fig. 7.

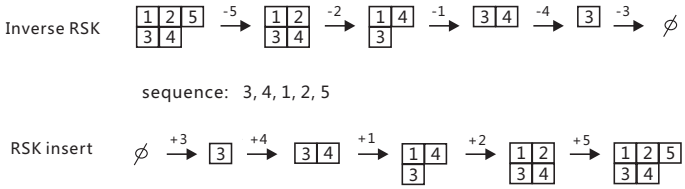

Fig. 7. The RSK-algorithm and its inverse. First we extract via the inverse RSK and then reinsert using RKS, recovering the original Young tableau. Below it is the origins of arcs that are RSK inserted and extractions appear when squares are removed.

The RSK-algorithm has also an inverse. Suppose we are given two shapes $\lambda^i \subsetneq \lambda^{i-1}$, which differ by exactly one square. Let T_{i-1} and T_i be Young tableaux of shape λ^{i-1} and λ^i, respectively. Then there exists a unique j contained in T_{i-1} and a unique tableau T_i such that T_{i-1} is obtained from T_i by inserting j using the RSK-algorithm, see Fig. 7.

We are now ready to describe the correspondence between diagrams and $*$-tableaux due to [3].

From k-noncrossing structures to $*$-tableaux: starting with the empty shape, consider the sequence $(n, n-1, \ldots, 1)$ and do the following:

- if j is the endpoint of an arc (i, j), then RSK-insert i
- if j is the startpoint of an arc (j, s), then remove the square containing j.
- if j is an isolated point, then do nothing, see Fig. 8.

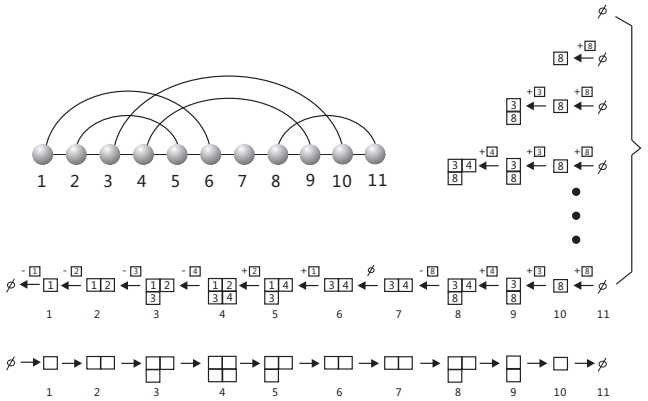

Fig. 8. From k-noncrosssing diagrams to $*$-tableaux using RSK insertion of the origins of arcs and removal of squares at the termini. Here we generate the $*$-tableaux in the top of Fig. 6.

From $*$-tableaux to k-noncrossing structures: Given a $*$-tableaux of empty shape, $(\varnothing, \lambda^1, \ldots, \lambda^{n-1}, \varnothing)$, reading $\lambda^i \setminus \lambda^{i-1}$ from left to right, at step i, we do the following:

- for a $+\square$-step we insert i into the new square
- for a $\varnothing$-step we do nothing
- for a $-\square$-step we extract the unique entry, $j(i)$, of the tableaux T^{i-1}, which via RSK-insertion into T^i recovers it (Fig. 7). The latter extractions, see Fig. 7, generate the arc-set $\{(i, j(i)) \mid i \text{ is a } -\square\text{-step}\}$ of a k-noncrossing diagram, see Fig. 9.

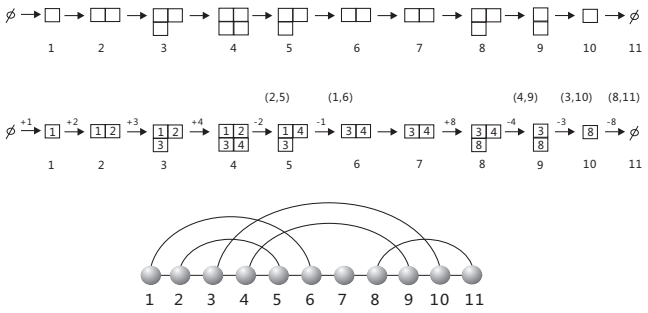

Fig. 9. From $*$-tableaux to partial matchings. If $\lambda^i \setminus \lambda^{i-1} = -\square$, then the unique number is extracted, which, if RSK-inserted into λ^i, recovers λ^{i-1}. This yields the arc-set of a k-noncrossing, partial matching.

Therefore, each $*$-tableaux of length n, containing shapes with at most $(k-1)$-rows, corresponds uniquely to a k-noncrossing partial matching on $[n]$ [3]. We denote the numbers of $*$-tableaux and those without hesitating steps (oscillating tableaux) of shape λ^i and length $(n-i)$, by $\mathsf{O}_k^*(\lambda^i, n-i)$ and $\mathsf{O}_k^0(\lambda^i, n-i)$, respectively.

Reflection and D-finiteness

The reflection principle [25, 5, 19] is a powerful technique in combinatorial enumeration. However, it is not directly applicable to RNA pseudoknot structures. Additional arguments [15] are needed for dealing with the non-reflectable, minimum arc-length condition, see Lemma 1.

Given a $*$-tableaux of shape λ, $(\lambda^i)_{i=0}^n$, we consider the number of squares in the sth row of shape λ^i, denoted by $x_s(i)$. It is evident that a $*$-tableaux of shape λ with at most $(k-1)$ rows uniquely corresponds to a walk of length n which starts at $a=(k-1,k-2,\ldots,1)$ and ends at $b=(k-1+x_1(n),\ldots,1+x_{k-1}(n))$ having steps $0,\pm e_i$, $1\leq i\leq k-1$ such that $0<x_{k-1}<\ldots<x_1$ at any step, see Fig. 10. That is, a $*$-tableaux of shape λ with at most $(k-1)$ rows corresponds to a lattice path in $\mathbb{Z}^{k-1}$ that remains in the interior of the dominant Weyl chamber [5].

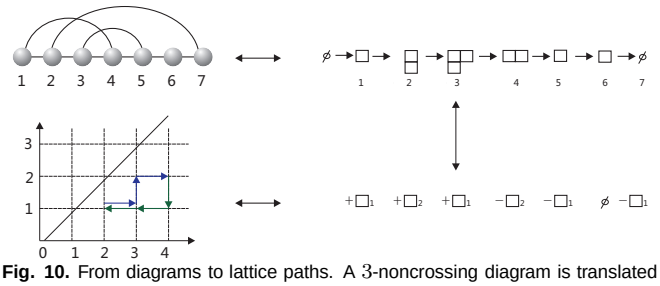

Fig. 10. From diagrams to lattice paths. A 3-noncrossing diagram is translated into a sequence of shapes ($*$-tableaux) which in turn induces a walk that stays in the dominant Weyl-chamber of $\mathbb{Z}^2$ starting and ending at $(2,1)$.

For $a,b\in\mathbb{Z}^{k-1}$, let $\Gamma(a,b)$ denote the set of walks without 0-steps of length n. Clearly, $\Gamma_n^0(a,b)=\mathsf{O}_k^0(\lambda,n)$, where λ represents the unique shape with at most $(k-1)$ rows that corresponds to the lattice point $b\in\mathbb{Z}^{k-1}$. Let $I_r(2x)$ denote the hyperbolic Bessel function of the first kind of order r.

Using the reflection principle, Grabiner [8] derived the following relation between the generating function and a determinant of Bessel functions

$$\sum_{n\geq 0}\Gamma_n^0(a,b)\frac{x^n}{n!}=\det[I_{a_i-b_j}(2x)-I_{a_i+b_j}(2x)]|_{i,j=1}^{k-1}. \qquad [2]$$

In [15] it is shown, using eq. (2), that for $k\geq 2$, the numbers of k-noncrossing RNA pseudoknot structures with minimum arc-length 2, $S_k(n)$, are P-recursive and given by

$$S_k(n)=\sum_{b\leq\lfloor\frac{n}{2}\rfloor}(-1)^b\binom{n-b}{b}\mathsf{O}_k^*(\varnothing^0,n-2b), \qquad [3]$$

where $\mathsf{O}_k^*(\lambda^i,n-i)$ satisfies

$$\mathsf{O}_k^*(\lambda^i,n-i)=\begin{cases}\sum_{l=0}^{\frac{n}{2}}\binom{n-i}{2l}\mathsf{O}_k^0(\lambda^i,n-i-2l), & \text{for } (n-i) \text{ even}\\ \sum_{l=0}^{\frac{n}{2}}\binom{n-i}{2l+1}\mathsf{O}_k^0(\lambda^i,n-i-2l-1), & \text{for } (n-i) \text{ odd.}\end{cases} \qquad [4]$$

As a result, the number of k-noncrossing RNA pseudoknot structures can be derived from the quantities $\mathsf{O}_k^0(\lambda^i,n)$, given by eq. (2).

Uniform generation

Eq. (2), combined with the fact that D-finite functions form an algebra [29] implies, that the ordinary generating function $\sum_{n\geq 0}\Gamma_n^0(a,b)x^n$ is D-finite. Since D-finiteness is equivalent to the P-recursiveness [30] of its coefficients, we derive

Corollary 1. *For fixed shape λ with at most $(k-1)$ rows and $n\in\mathbb{N}$, there exists some $m\in\mathbb{N}$ and polynomials $p_0(n),\ldots,p_m(n)$ such that*

$$p_m(n)\mathsf{O}_k^0(\lambda,n+m)+\cdots+p_0(n)\mathsf{O}_k^0(\lambda,n)=0. \qquad [5]$$

In particular, the numbers $\mathsf{O}_k^0(\lambda,n)$ can be computed in $O(n)$ time.

We remark, that for fixed n and λ, the derivation of eq. (5) is a pre-processing step. It has to be derived only once, for instance employing Zeilberger's algorithm [38, 27]. The recursions of Corollary 1 can be found empirically with the MAPLE package `gfun` using the command `listtorec`.

Theorem 2. *A random k-noncrossing structure can be generated, after polynomial pre-processing time, with uniform probability in linear time. The algorithmic implementation, see Algorithm 1, has $O(n^{k+1})$ pre-processing time and $O(n^k)$ space complexity. Each k-noncrossing structure is generated with $O(n)$ space and time complexity.*

Let $\mathsf{W}_k^*(\lambda^i,n-i)$ denote the number of $*$-tableaux of shape λ^i with at most $(k-1)$ rows of length $(n-i)$ that do not contain any $(+\square_1,-\square_1)$-steps, then we have

Algorithm 1.

PShape ← ArrayP*(n,k) (computation of* $\mathsf{O}_k^*(\lambda^i,n-i)$, $i=0,1,\ldots,n-1$, λ^i*)*
SShape ← ArrayS*(n,k) (computation of* $\mathsf{W}_k^*(\lambda_j^i,n-i)$, $j=0,1^+,1^-,\ldots,(k-1)^+,(k-1)^-$; $i=0,1,\ldots,n-1$, *stored in the* $k\times n$ *array SShape)*
while $i<n$ **do**
  flag ← 1
  X[0] ← $\mathsf{W}_k^*(\lambda_0^{i+1},n-(i+1))$
  X[1]← $\mathsf{W}_k^*(\lambda_{1^+}^{i+1},n-(i+1))-\mathsf{W}_k^*(\lambda_{1^-}^{i+2},n-(i+2))$
  if *flag=0 and j=2* **then**
    X[2]← 0
  else
    X[2]← $\mathsf{W}_k^*(\lambda_{1^-}^{i+1},n-(i+1))$
  end if
  sum ← *X[0]+X[1]+X[2]*
  for j *from* 2 *to* $k-1$ **do**
    X[2j-1] ← $\mathsf{W}_k^*(\lambda_{j^+}^{i+1},n-(i+1))$
    X[2j] ← $\mathsf{W}_k^*(\lambda_{j^-}^{i+1},n-(i+1))$
    sum←*sum+X[2j-1]+X[2j]*
  end for
  Shape ← Random*(sum) (*Random *generates the random shape* λ_j^{i+1} *with probability X[j]/sum)*
  $i\leftarrow i+1$
  if *Shape* =$\lambda_{1^+}^i$ **then**
    flag ← 0
  end if
  Insert λ_j^{i+1} *into Tableaux*
end while
Map*(Tableaux)*

We remark that for $k=3$, explicit formulas based on the work of [19, 6, 7] allow us to derive the transition probabilities directly.

Conclusion

This paper provides an interpretation of RNA pseudoknot structures as sampling paths of a Markov process. This point of view has the potential to offer radically new ways of dealing with the complex cross-serial interactions in RNA molecules.

It is not obvious that cross-serial interactions can be expressed in terms of a Markov process, since the latter are by construction local, having no memory of the sampling path, except of the last step. Our construction shows *why* this is the case: $(k-1)$-crossings in RNA molecules can be expressed *locally* by a $(k-1)$-row of squares in the associated shape-sequence. This locality holds for arbitrary k. The

conclusion that cross serial interactions are indeed local is good news for designing *ab initio* folding algorithms.

The framework presented here is a stepping stone towards the sampling with non-uniform transition probabilities. The uniform sampling of pseudoknot structures, compatible to a given sequence, is displayed in Fig. 11. Here we insert the nucleotides into the squares and in analogy to Theorem 2 consider compatible paths, thereby sampling uniformly. For instance, Theorem 2 immediately allows to sample sequence-specific *locally uniformly* in linear time, by setting all incompatible uniform transition probabilities to zero and rescale.

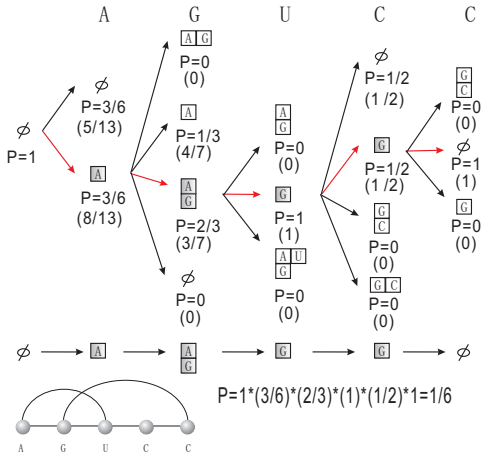

Fig. 11. Sequence specific, uniform and locally uniform sampling of RNA pseudoknot structures. Here we display the two sampling variants for the sequence **AGUCC**.

It is also possible to assign transitions that induce base pairs (i.e. extractions) particular weights. This leads to the energy based sampling of pseudoknot structures, which can be made context dependent along the lines of [37]. Stacking bonds could be included as well in our framework. Higher-order Markov processes naturally model the dependencies of adjacent arcs.

Since our approach is path-based, it offers the possibility to formalize the kinetics of the folding. In addition, we can generalize a class of stochastic CFG-foldings for RNA secondary structures [14]. Examining a set of known molecular foldings, it is now possible to derive the maximum likelihood estimators of the model parameters [23] and to fold pseudoknot structures using hidden Markov models [21]. One important advantage of this approach is to avoid explicit knowledge of the energy parameters of pseudoknot-loops.

Our algorithm is available[1] in C and in MAPLE.

Proof of the main result

A 1-arc corresponds to a subsequence of shapes $(\lambda^i, \lambda^{i+1}, \lambda^{i+2} = \lambda^i)$, obtained by first adding and then removing a square in the first row. This sequence corresponds to a pair of steps $(+\square_1, -\square_1)$, where $+\square_1$ and $-\square_1$ indicate that a square is added and subtracted in the first row, respectively. In terms of $*$-tableaux having at most $(k-1)$ rows, eq. (3) can be rewritten as follows

$$\mathsf{W}_k^*(\varnothing^0, n) = \sum_{b=0}^{\frac{n}{2}} (-1)^b \binom{n-b}{b} \mathsf{O}_k^*(\varnothing^0, n-2b).$$

In order to prove our main result we have to generalize this relation from the empty shape, $\varnothing$ to arbitrary shapes, λ.

Lemma 1. *Let λ^i be an arbitrary shape with at most $(k-1)$ rows, then*

$$\mathsf{W}_k^*(\lambda^i, n-i) = \sum_{b=0}^{\frac{n-i}{2}} (-1)^b \binom{(n-i)-b}{b} \mathsf{O}_k^*(\lambda^i, n-i-2b). \quad \textbf{[6]}$$

Let $\mathcal{Q}_k^*(\lambda^i, n-i, j)$ denote the set of $*$-tableaux of shape λ^i of length $(n-i)$ having at most $(k-1)$ rows containing exactly j pairs $(+\square_1, -\square_1)$ and set $\mathsf{Q}_k^*(\lambda^i, n-i, j) = |\mathcal{Q}_k^*(\lambda^i, n-i, j)|$.

Proof: Let $(\lambda^s)_{s=0}^{(n-2b)-i}$ be a $*$-tableaux of shape λ^i. We select from the set $\{0, \ldots, (n-2b)-i-1\}$ an increasing sequence of labels $(r_1, \ldots, r_b)$. For each r_s we insert a pair $(+\square_1, -\square_1)$ after the corresponding shape λ^{r_s}, see Fig. 12. This insertion generates a $*$-tableaux of length $(n-i)$ of shape λ^i.

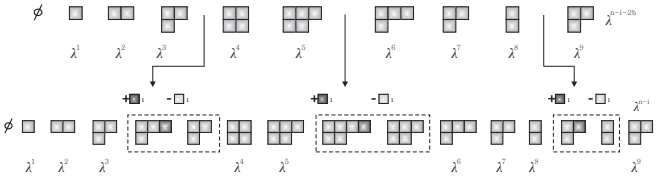

Fig. 12. Illustration of the proof idea: pairs $(+\square_1, -\square_1)$ are inserted at positions $3, 5$ and 8, respectively.

Considering the above insertion for all sequences $(r_1, \ldots, r_b)$, we arrive at a family $\mathcal{F}_b$ of $*$-tableaux of length $(n-i)$ containing at least b pairs, $(+\square_1, -\square_1)$. Since we can insert at any position $0 \le h \le ((n-i)-2b-1)$, $\mathcal{F}_b$ has cardinality $\binom{(n-i)-b}{b} \mathsf{O}_k^*(\lambda^i, n-i-2b)$. By construction, each $*$-tableaux $(\lambda^s)_{s=0}^{n-i} \in \mathcal{F}_b$, that exhibits exactly j pairs $(+\square_1, -\square_1)$ appears with multiplicity $\binom{j}{b}$, whence

$$\sum_{j \ge b} \binom{j}{b} \mathsf{Q}_k^*(\lambda^i, n-i, j) = \binom{(n-i)-b}{b} \mathsf{O}_k^*(\lambda^i, n-i-2b). \quad \textbf{[7]}$$

We consider $F_k(x) = \sum_{j \ge 0} \mathsf{Q}_k^*(\lambda^i, n-i, j) x^j$. Taking the bth derivative and setting $x = 1$ we obtain $\frac{1}{b!} F_k^b(1) = \sum_{j \ge b} \binom{j}{b} \mathsf{Q}_k^*(\lambda^i, n-i, j) 1^{j-b}$ and computing the Taylor expansion of $F_k(x)$ at $x = 1$

$$\begin{aligned} F_k(x) &= \sum_{b \ge 0} \frac{1}{b!} F_k^b(1) (x-1)^b \\ &= \sum_{b=0}^{\frac{n-i}{2}} \binom{(n-i)-b}{b} \mathsf{O}_k^*(\lambda^i, n-i-2b) (x-1)^b. \end{aligned}$$

Since $\mathsf{W}_k^*(\lambda^i, n-i) = \mathsf{Q}_k^*(\lambda^i, n-i, 0)$ is the constant term of $F_k(x)$, the lemma follows.

Proof of Theorem 2: The idea is to interpret $*$-tableaux without pairs of steps, $(+\square_1, -\square_1)$, (good $*$-tableaux) as paths of a stochastic process. To this end, we index the shapes λ^{i+1} according to their predecessors: let $i = 0, 1, \ldots, n-1$ and $j \in \{0, 1^+, 1^-, \ldots, (k-1)^+, (k-1)^-\}$. Setting $\lambda_j^0 = \varnothing$, we write λ_j^{i+1}, if λ^{i+1} is obtained via

- doing nothing (λ_0^{i+1})
- adding a square in the jth row $(\lambda_{j^+}^{i+1})$
- deleting a square in the jth row $(\lambda_{j^-}^{i+1})$.

[1] http://www.combinatorics.cn/cbpc/unif.html

With this notation, the number of good $*$-tableaux of shape $\lambda_{1^+}^{i+1}$ of length $(n-(i+1))$ is given as follows:

$$\mathsf{V}_k^*(\lambda_{1^+}^{i+1}, n-(i+1)) = \mathsf{W}_k^*(\lambda_{1^+}^{i+1}, n-(i+1)) - \mathsf{W}_k^*(\lambda_{1^-}^{i+2}, n-(i+2)).$$

In order to derive transition probabilities, we establish two equations: first, for any λ_j^i, where $j \neq 1^+$, we have $\mathsf{W}_k^*(\lambda_j^i, n-i) =$

$$\mathsf{V}_k^*(\lambda_{1^+}^{i+1}, n-(i+1)) + \mathsf{W}_k^*(\lambda_{1^-}^{i+1}, n-(i+1)) + \sum_{h=2}^{k-1} \left(\mathsf{W}_k^*(\lambda_{h^+}^{i+1}, n-(i+1)) + \mathsf{W}_k^*(\lambda_{h^-}^{i+1}, n-(i+1)) \right) + \mathsf{W}_k^*(\lambda_0^{i+1}, n-(i+1))$$

and second, in case of $j = 1^+$, we have $\mathsf{V}_k^*(\lambda_{1^+}^i, n-i) =$

$$\mathsf{V}_k^*(\lambda_{1^+}^{i+1}, n-(i+1)) + \mathsf{W}_k^*(\lambda_0^{i+1}, n-(i+1)) \sum_{h=2}^{k-1} \left(\mathsf{W}_k^*(\lambda_{h^+}^{i+1}, n-(i+1)) + \mathsf{W}_k^*(\lambda_{h^-}^{i+1}, n-(i+1)) \right).$$

We are now in a position to specify the process $(X^i)_{i=0}^n$:

- $X^0 = X^n = \varnothing$ and X^i is a shape having at most $(k-1)$ rows
- for $0 \leq i \leq n-1$, X^i and X^{i+1} differ by at most one square
- there exists no subsequence $X^i, X^{i+1}, X^{i+2} = X^i$ obtained by first adding and second removing a square in the first row
- for $j \neq 1^+$

$$\mathbb{P}(X^{i+1} = \lambda_l^{i+1} \mid X^i = \lambda_j^i) = \begin{cases} \frac{\mathsf{W}_k^*(\lambda_l^{i+1}, n-(i+1))}{\mathsf{W}_k^*(\lambda_j^i, n-i)}, & \text{for } l \neq 1^+ \\ \frac{\mathsf{V}_k^*(\lambda_{1^+}^{i+1}, n-(i+1))}{\mathsf{W}_k^*(\lambda_j^i, n-i)}, & \text{for } l = 1^+ \end{cases} \quad \textbf{[8]}$$

- for $j = 1^+$

$$\mathbb{P}(X^{i+1} = \lambda_l^{i+1} \mid X^i = \lambda_{1^+}^i) = \begin{cases} \frac{\mathsf{W}_k^*(\lambda_l^{i+1}, n-(i+1))}{\mathsf{V}_k^*(\lambda_{1^+}^i, n-i)}, & \text{for } l \neq 1^+, 1^- \\ \frac{\mathsf{V}_k^*(\lambda_{1^+}^{i+1}, n-(i+1))}{\mathsf{V}_k^*(\lambda_{1^+}^i, n-i)}, & \text{for } l = 1^+. \end{cases} \quad \textbf{[9]}$$

We observe that eq. (8) and eq. (9) imply

$$\prod_{i=0}^{n-1} \mathbb{P}(X^{i+1} = \lambda^{i+1} \mid X^i = \lambda^i) = \frac{\mathsf{W}_k^*(\lambda^n = \varnothing, 0)}{\mathsf{W}_k^*(\lambda^0 = \varnothing, n)} = \frac{1}{\mathsf{W}_k^*(\varnothing, n)}. \quad \textbf{[10]}$$

Consequently, the process $(X^i)_{i=0}^n$ generates random k-noncrossing structures with uniform probability in $O(n)$ time and space. According to Corollary 1, we can for any λ^i, having at most $(k-1)$ rows, compute $\mathsf{O}_k^0(\lambda^i, n-i)$ in $O(n)$ time. Consequently, we can generate the arrays $(\mathsf{O}_k^*(\lambda^i, n-i))_{\lambda^i, n-i}$ and $(\mathsf{W}_k^*(\lambda^i, n-i))_{\lambda^i, n-i}$ in $O(n^2) + O(n^2)\, O(n^{k-1})$ time and $O(n^k)$ space.

A random k-noncrossing structure is then generated as a $*$-tableaux with at most $(k-1)$ rows using the array $(\mathsf{W}_k^*(\lambda^i, n-i))_{\lambda^i, n-i}$ with $O(n)$ time and space complexity. □

ACKNOWLEDGMENTS. We want to thank M. Marz, M. Nebel and P.F. Stadler for their help and suggestions. This work was supported by the 973 Project, the PCSIRT Project of the Ministry of Education, and the National Science Foundation of China.

1. **D. Aldous (1982), *Random walks on finite groups and rapidly mixing Markov chains* Séminaire de Probabilités XVII, Lecture Notes in Mathematics 986 (Springer, Berlin 1982) 243-297.**
2. **M. Chamorro, N. Parkin and H.E. Varmus, (1991) *An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA*, J. Proc. Natl. Acad. Sci. USA, 89, 713-717.**
3. **W.Y.C. Chen, E.Y.P. Deng, R.R.X. Du, R.P. Stanley and C.H. Yan, (2007) *Crossings and nestings of matchings and partitions*, Trans. Amer. Math. Soc., 359, 1555-1575.**
4. **P. Duchon, P. Flajolet, G. Louchard and G. Schaeffer, (2004) *Boltzmann Samplers for the random generation of combinatorial structures*, Combin. Probab. Comput., 13, 577-625.**
5. **I.M. Gessel, D. Zeilberger, (1992) *Random walk in a Weyl Chamber*, Proc. Amer. Math. Soc., 115, 27-31.**
6. **I.M. Gessel, X.G. Viennot, (1989) *Determinants, paths, and plane partitions*, preprint.**
7. **D. Gouyou-Beauschamps, (1989) *Standard Young tableaux of height 4 and 5*, Europ. J. Combin., 10, 69-82.**
8. **D.J. Grabiner, P. Magyar, (1993) *Random walks in Weyl chambers and the decomposition of Tensor Powers*, J. Algebr. Comb., 2, 239-260.**
9. **C. Haslinger and P.F. Stadler, (1999) *RNA structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties*, Bull. Math. Bio., 61(3), 437-467.**
10. **F.W.D. Huang, W.W.J. Peng and C.M. Reidys, (2009) *Folding 3-Noncrossing RNA Pseudoknot Structures*, J. Comput. Biol., to appear.**
11. **M.R. Jerrum, L.G. Valiant, and V.V. Vazirant, (1986) *Random generation of combinatorial structures from a uniform distribution*, Theor. Comput. Sci., 43, 169-188.**
12. **M. Jerrum and A. Sinclair, (1990) *Fast uniform generation of regular graphs*, Theor. Comput. Sci., 73, 91-100.**
13. **D. Kleitman, (1970) *Proportions of Irreducible Diagrams*, Studies in Appl. Math. 49, 297-299.**
14. **B. Knudsen and J. Hein, (1999) *RNA secondary structure prediction using stochastic context-free grammars and evolutionary history*, Bioinformatics, 15, 446-454.**
15. **E.Y. Jin, J. Qin and C.M. Reidys, (2008) *Combinatorics of RNA structures with pseudoknots*, Bull. Math. Biol., 70(1), 45-67.**
16. **E.Y. Jin and C.M. Reidys, (2008) *Combinatorial design of pseudoknot RNA*, Adv. Appl. Math., (42), 135-151.**
17. **E.Y. Jin, C.M. Reidys, (2007) *Central and local limit theorems for RNA structures*, J. Theor. Biol., 250(3), 547-559.**
18. **D.A.M. Konings and R.R. Gutell, (1995) *A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs*, RNA 1, 559-574.**
19. **B. Lindström, (1973) *On the vector representation of induced matroids*, Bull. London Math. Soc., 5, 85-90.**
20. **A. Loria and T. Pan, (1996). *Domain structure of the ribozyme from eubacterial ribonuclease P*, RNA 2, 551-563.**
21. **U. Laube and M.E. Nebel, (2009) *Maximum Likelihood Analysis of Algorithms and Data Structures*, Theo. Comp. Sc., to appear.**
22. **R. B. Lyngso and C. N. S. Pedersen, *RNA pseudoknot prediction in energy-based models*, J. Comput. Biol. 7 (2000), 409-427.**
23. **D. Metzler and M.E. Nebel (2008) *Predicting RNA Secondary Structures with Pseudoknots by MCMC Sampling*, J. of Math. Biol. 56, 161-181.**
24. **R. Nussinov, G. Pieczenik, J.R. Griggs and D.J. Kleitman, (1978) *Algorithms for loop matchings* SIAM J. of Appl. Math. 35, 68-82.**
25. **M. Renault, (2008) *Lost (and found) in translation: André's actual method and its application to the generalized Ballot problem*, Amer. Math. Monthly., 115, 358-363.**
26. **E. Rivas and S.R. Eddy, (1999) *A dynamic programming algorithm for RNA structure prediction including pseudoknots*, J. of Mol. Biol., 285(5), 2053-2068.**
27. **B. Salvy, P. Zimmerman, (1994) *GFUN: a Maple package for the manipulation of generating and holonomic functions in one variable*, ACM Transactions on Mathematical Software (TOMS), 20, 163-177.**
28. **D.W. Staple and S.E. Butcher, (2005) *Pseudoknots: RNA Structures with Diverse functions* PLoS Biol. 3(6), 956-959.**
29. **R.P. Stanley, (1980) *Differentiably finite power series*, Europ. J. Combin., 1, 175-188.**
30. **R.P. Stanley, (1996) *Enumerative Combinatorics(Volume 1)*, Cambridge University Press, Cambridge.**
31. **C. Tuerk, S. MacDougal and L. Gold, (1992) *RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase*, Proc. Natl. Acad. Sci. USA, 89:6988-6992.**
32. **I. L. Hofacker, (2003) *Vienna RNA secondary structure server*, Nucl. Acids. Res. 31(13), 3429-3431.**
33. **M.S. Waterman, (1979) *Combinatorics of RNA hairpins and cloverleafs*, Stud. Appl. Math., 60, 91-96.**
34. **E. Westhof and L. Jaeger, (1992) *RNA pseudoknots*, Current Opinion Struct. Biol. 2, 327-333.**
35. **H.S. Wilf (1977) *A unified setting for sequencing, ranking, and selection algorithms for combinatorial objects*, Advances in Mathematics, 24, 281-291.**
36. **A. Nijenhuis and H.S. Wilf, (1978) *Combinatorial Algorithms*, Academic Press.**
37. **C. Witwer, I.L. Hofacker and P.F. Stadler, (2004) *Prediction of Consensus RNA Secondary Structures Including Pseudoknots*, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 1, 66-77.**
38. **D. Zeilberger, (1990) *A holonomic systems approach to special functions identities*, J. Comput. Appl. Math, 32, 321-368.**
39. **M. Zito, I. Pu, M. Alnos and A. Gibbons, (1996) *RCN algorithms for the uniform generation of combinatorial structures*, Society for Industrial and Applied Mathematics, 429-437.**