CParty: Hierarchically Constrained Partition Function of RNA Pseudoknots

Mateo Gray, Luke Trinity, Ulrike Stege, Yann Ponty, Sebastian Will, Hosna Jabbari

CParty: Hierarchically Constrained Partition Function of RNA Pseudoknots

Mateo Gray[1, ‡] **Luke Trinity**[2, ‡] **Ulrike Stege**[2] **Yann Ponty**[3] **Sebastian Will**[3] **and Hosna Jabbari**[1,*]

[1]Department of Biomedical Engineering, University of Alberta, [2]Department of Computer Science, University of Victoria and [3]Institut Polytechnique de Paris

*Corresponding author. jabbari@ualberta.ca ‡

Abstract

Motivation

Biologically relevant RNA secondary structures are routinely predicted by efficient dynamic programming algorithms that minimize their free energy. Starting from such algorithms, one can devise partition function algorithms, which enable stochastic perspectives on RNA structure ensembles. As the most prominent example, McCaskill's partition function algorithm is derived from pseudoknot-free energy minimization. While this algorithm became hugely successful for the analysis of pseudoknot-free RNA structure ensembles, as of yet there exists only one pseudoknotted partition function implementation, which covers only simple pseudoknots and comes with a borderline-prohibitive complexity of $O(n^5)$ in the RNA length n.

Results

Here, we develop a partition function algorithm corresponding to the hierarchical pseudoknot prediction of HFold, which performs exact optimization in a realistic pseudoknot energy model. In consequence, our algorithm CParty carries over HFold's advantages over classical pseudoknot prediction to characterizing the Boltzmann ensemble at equilibrium. Given an RNA sequence S *and* a pseudoknot-free structure G, CParty computes the partition function over all possibly pseudoknotted density-2 structures $G \cup G'$ of S that extend the fixed G by a disjoint pseudoknot-free structure G'. Thus, CParty follows the common hypothesis of hierarchical pseudoknot formation, where pseudoknots form as tertiary contacts only after a first pseudoknot-free 'core' G and we call the computed partition function *hierarchically constrained (by G)*. Like HFold, the dynamic programming algorithm CParty is very efficient, achieving the low complexity of the pseudoknot-free algorithm, i.e. cubic time and quadratic space. Finally, by computing pseudoknotted ensemble energies, we unveil kinetics features of a therapeutic target in SARS-CoV-2.

Availability

CParty is available at `https://github.com/HosnaJabbari/CParty`.

Key words: RNA, pseudoknots, hierarchical folding, hierarchically constrained partition functions, bisecondary structures

1. Introduction

RNA molecules play a vital role in cellular processes; many possess functional structures (Cruz and Westhof, 2009; Kozak, 2005; Mortimer et al., 2014; Warf and Berglund, 2010; Wilson and Lilley, 2015). As experimental methods to detect RNA structure are time consuming and costly, computational methods for predicting RNA structure have become indispensable. We focus on the accurate prediction of RNA secondary structure (2D), which in turn sheds light on the 3D structure of the RNA. Various algorithms have been developed to tackle this problem, aiming to predict the most energetically favorable structure based on thermodynamic models and empirical data (Bernhart et al., 2008; Reuter and Matthews, 2010; Rivas, 2020; Rivas and Eddy, 1999; Sato et al., 2011; Zuker and Stiegler, 1981). The best-known, most widely-used thermodynamics-based approaches are the algorithms by Zuker and Stiegler (1981) for predicting RNA secondary structure and due to McCaskill (1990) for computing partition functions.

The Zuker algorithm finds the minimum free energy (MFE) structure among all possible pseudoknot-free structures for the given RNA sequence (Zuker and Stiegler, 1981). RNA secondary structure prediction is NP-hard (Akutsu, 2000; Lyngsø and Pedersen, 2000) and even inapproximable (Sheikh et al., 2012) when pseudoknots are allowed. Existing efficient algorithms for exact prediction of pseudoknotted RNA secondary structure handle only restricted classes of structures, trading off run-time and structure complexity (Chen et al., 2009b; Jabbari et al., 2018; Rivas and Eddy, 1999; Reeder and Giegerich, 2004).

Offering a stochastic perspective on the entire ensemble of possible pseudoknot-free structures of an RNA, the McCaskill algorithm computes *partition functions*. Algorithmically it has strong parallels to the pseudoknot-free MFE algorithm by Zuker, since both algorithms decompose the same

‡Authors contributed equally

structure space in their dynamic programming scheme. Generally, there is a one-to-one correspondence between the search spaces considered by partition function algorithms, such as McCaskill (1990), and MFE algorithms, provided they are unambiguous and complete (Ponty and Saule, 2011). This correspondence also extends to pseudoknotted structure spaces. Consequently, the run-time vs. structure complexity trade-offs that were discussed for pseudoknot MFE algorithms like Chen et al. (2009b); Jabbari et al. (2018); Rivas and Eddy (1999) are mirrored in (hypothetical) corresponding partition function algorithms. So far, the only pseudoknot partition function algorithm, which realizes this idea, is due to Dirks and Pierce (2003). Their algorithm (D&P) handles the restricted class of simple pseudoknots. While it is implemented in NUPACK, its practical application is limited by the algorithm's $O(n^5)$ time and $O(n^4)$ space complexity.

To address the high time and space complexity of other pseudoknot prediction algorithms, we previously developed HFold (Jabbari et al., 2007, 2008). Given an RNA sequence and a pseudoknot-free structure, HFold calculates a potentially pseudoknotted secondary structure with minimum free energy in a full RNA energy model that extends the given structure with a, in a specifically defined way, 'compatible' second pseudoknot-free structure. By following this principle, HFold becomes the first MFE algorithm that adheres to the *hierarchical folding hypothesis*. This hypothesis suggests that RNA initially folds into a pseudoknot-free structure, and then additional bases pair to further lower the MFE of the structure, possibly forming pseudoknots (Tinoco Jr and Bustamante, 1999). Hierarchical folding has been experimentally observed in the formation of pseudoknotted structures (Cho et al., 2009), including frameshift stimulating pseudoknots (Chen et al., 2009a).

HFold has only cubic time and quadratic space complexity. This means it is as efficient as pseudoknot-free prediction algorithms and, for example, is faster than CCJ (Chen et al., 2009c) or D&P (Dirks and Pierce, 2003) by a quadratic factor. HFold takes a pseudoknot-free structure G as input, and predicts a pseudoknot-free structure G' such that $G \cup G'$ has minimum free energy among all *density-2* structures (Jabbari et al., 2008) (Figure 1; Methods). The class of density-2 structures allows for arbitrary depth and length of nested pseudoknots including H-type pseudoknots and kissing hairpins. This class encompasses structures not handled by CCJ and is more comprehensive than the structure class described by the partition function algorithm by Dirks and Pierce (2003). Since the selection of the non-pseudoknotted partial structure G is crucial in hierarchical folding, previous work has identified promising techniques for selecting G (Jabbari and Condon, 2014; Trinity et al., 2023). These techniques involve computing energetically favorable pseudoknot-free structures (Jabbari and Condon, 2014) or choosing partial structures compatible with chemical modification data, such as SHAPE reactivity (Trinity et al., 2023).

The main objective of this work is to develop and study a partition function counterpart to HFold. To achieve this, we present CParty, a *constrained partition function* (CPF) algorithm that considers possibly pseudoknotted density-2 structures.

The primary challenge in constructing the CParty algorithm is that HFold decomposes density-2 structures in a non-trivially redundant manner. Removing or avoiding such ambiguities is a general and recurring issue in the construction of dynamic programming algorithms to compute RNA partition functions, e.g. McCaskill (1990); Dirks and Pierce (2003). Since partition functions sum over the weights of all considered structures, any ambiguities directly lead to over-"counting". To address this problem, we have resolved all ambiguities in the decomposition process. This careful preparation step, finally enables deriving the CParty algorithm by a systematic exchange of algebras (e.g. a core idea of ADP (Giegerich, 2000)).

Similar to HFold, CParty takes an RNA sequence and an input structure G. Then, it calculates the hierarchically constrained partition function over $G \cup G'$, where G' is pseudoknot-free and $G \cup G'$ is density-2. Here, we focus on the algorithm for calculating this hierarchically constrained partition function (i.e. partition function for the ensemble of structures $G \cup G'$, where G is fixed) and we leave the base pair probability calculation as a future direction.

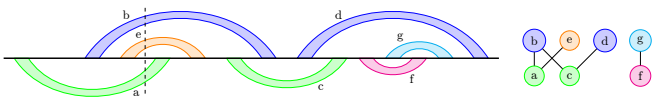

Fig. 1: Bands of a bisecondary structure (left) with the corresponding bipartite crossing graph (right). Note that the connected components of the crossing graph can be understood as groups of bands that directly or transitively cross each other. The structure has density three, since there are positions which are simultaneously covered by three transitively crossing bands. For example, the three bands a, b, and e cover the position indicated by the dashed line. When removing one of the bands a, b, or e, the remaining structure is density-2 (see Methods). The dynamic programming algorithms CParty and HFold exploit that the crossing bands of density-2 structures are arranged in chains, in the sense that bands a,b,c,d form a 4-chain or f,g form a 2-chain. This allows decomposing density-2 structures by recursively decomposing such chains.

To clarify, CParty is designed to compute hierarchically constrained partition functions of RNAs, which ensures very high efficiency. In some applications, this hierarchical approach can be advantageous, as it allows for the integration of prior knowledge. Additionally, its efficiency makes it suitable for iterated use in meta-strategies (cf. Jabbari and Condon (2014)).

1.1. Contributions

We introduce the novel hierarchical constraint partition function algorithm CParty as a counterpart to HFold. CParty decomposes the density-2 structure class completely and, in contrast to HFold, unambiguously. We implement CParty to perform realistic computations using a full-featured pseudoknot energy model (HotKnots 2.0 (Andronescu et al., 2010)), thoroughly scrutinize the implementation, and study its properties. Through empirical time complexity analysis, we demonstrate that CParty outperforms the only other existing pseudoknotted partition function algorithm in NUPACK. Applying our novel tool to the SARS-CoV-2 frameshift element, we compute constrained ensemble energies and unveil a key kinetic transition of its pseudoknot (Kelly et al., 2020).

2. Methods

2.1. RNA Secondary Structure

An RNA *sequence* of length n, known as the *primary structure* of RNA, is represented as a string in $\{A, C, G, U\}^n$. Its *secondary structure* is a set of base pairs $i.j$, where $1 \leq i < j \leq n$, and each *base* $1 \leq i \leq n$ occurs in at most one pair (no triplets). A secondary structure is called *crossing* or *pseudoknotted* if there are at least two base pairs, $i.j$ and $i'.j'$, that *cross* each other (i.e. $i < i' < j < j'$ or $i' < i < j' < j$). Otherwise, it is called *pseudoknot-free* or *non-crossing*. The base pairs $i.j$ and $i'.j'$ are *nested* if $i < i' < j' < j$ or $i' < i < j < j'$. Given a secondary structure G that pairs i, $bp_G(i)$ denotes the other end of the base pair of i in G; similarly, $bp(i)$ refers to the other end in $G \cup G'$.

Features of pseudoknotted and density-2 structures.

Due to space restrictions, we review important features of pseudoknotted structures briefly, and refer to the literature (Dirks and Pierce, 2003; Jabbari et al., 2008) for full detail; see also Fig. 1. Pseudoknotted structures can be classified by considering specific subsets of base pairs called *bands* (Dirks

and Pierce, 2003). A band of an RNA structure is a maximal subset of base pairs with the properties that 1) all of its base pairs are pairwisely nested; 2) each base pair of the remaining structure crosses either all or no base pairs of the band; and, moreover, 3) the base pairs of a band cross at least one base pair of the structure.

The literature distinguishes various classes of RNA structures such as simple pseudoknots, kissing hairpins, k-knots, and genus g that all can be characterized by specific restrictions on the crossing configurations of bands.

In this work, we focus on a specific subclass of *bisecondary* structures (Fontana et al., 1993; Hasslinger and Stadler, 1999; Witwer et al., 2004), which can be decomposed into two pseudoknot-free secondary structures. Specifically, we consider the subclass of density-2 structures defined by Jabbari et al. (2008) to precisely describe the search space of HFold. Envision the *crossing graph* of a structure that consists of one node for each band and one edge between any pair of crossing bands. In this graph, we can identify connected components of bands, that are in direct or transitive crossing relation to each other. This allows to characterize bisecondary structures graph-theoretically as the structures with bipartite crossing graphs. In density-k structures, the number of bands per connected component that cover a single position is less or equal k. For example, in Fig. 1, positions are covered by up to three bands of the connected component $\{a, b, c, d, e\}$. Thus, the example shows the bands of a density-3 structure, the density-2 property is violated by the bands a, b and e; e.g. removing the band e leaves a density-2 structure.

We require additional technical definitions from HFold's description (Jabbari et al., 2008): In density-2 structures, a *region* $[i, j]$ (denoting positions $i, i+1, \ldots, j$ is *closed*, either if i pairs with j, or if they are transitively connected due to a chain of crossing bands. In the latter case, i and j are the left and right ends of a *pseudoloop*, which is closed by base pairs of i and j as well as the outer base pairs of the other bands in the chain. For example, in Fig. 1, the outermost base pairs of bands a,b,c, and d form such a chain and close a pseudoloop. Sec. Supp1.1 (Supplementary Information) summarizes further definitions.

2.2. Energy Model

To assess the energy of RNA structures, we distinguish different types of structural elements, called *loops*, i.e. hairpin loops, stacks, bulges, interior loops, or multiloops. Loops are generally defined by their outer and potentially inner closing base pairs (Rastegari and Condon, 2007).

Nearest neighbor energy models define the free energy $E(G)$ of a secondary structure G as the sum of the energies of its loops $E(G) = \sum_{L \in G} E^{\text{loop}}(L)$. A prominent example is the Turner 2004 energy model (Turner and Matthews, 2009) for pseudoknot-free RNAs, which is used by RNAfold. For pseudoknotted RNAs, Dirks and Pierce (2003) introduced the DP03 energy model, used for pseudoknot prediction in NUPACK; it extends the Turner model by adding penalties for pseudoknots and bands, as well as parameters to score multiloops that 'span' a band.

CParty's energy model and Vienna RNA based implementation. In CParty, we utilize the DP09 energy parameters of HotKnots 2.0, which improve upon the DP03 energy model due to training on known pseudoknotted structures (Andronescu et al., 2010). Specific parameters and loop energy functions are provided in Supplementary Table 1. While HFold calculates loop energies based on SimFold (Andronescu et al., 2005), CParty uses the Vienna RNA library (Lorenz et al., 2011). For this purpose, the energy model parameters were translated to a compatible format, allowing for better interoperability and comparability with the Vienna RNA package. Additionally, our CParty implementation supports hard constraints that restrict the partition functions to structures that leave specified bases unpaired. We note that CParty is limited to the constraints of its energy model and hence, limited to $A.U$, $G.C$ and $G.U$ base pairings.

2.3. Problem statement: partition functions over density-2 structures

Given an RNA sequence S, a pseudoknot-free secondary RNA structure G, CParty computes the hierarchically constrained partition function (CPF)

$$Z_S^G = \sum_{\substack{G' : G' \text{ is secondary structure of } S \\ \text{s.t. } G \cap G' = \{\} \text{ and } G \cup G' \text{ is density-2}}} \exp(-E_S(G \cup G')/(RT)), \quad (1)$$

where T denotes the temperature (e.g, $T = 37\,^\circ\mathrm{C}$) and R denotes the universal gas constant ($R \approx 1.987\,\mathrm{cal\,K^{-1}\,mol^{-1}}$).

Analogous to the pseudoknot-free partition function developed by McCaskill (McCaskill, 1990), this partition function is defined as the sum of *Boltzmann weights* $\exp(-E_S(\hat{G})/(RT))$ of RNA structures $\hat{G}$, where E_S computes the RNA energy. Extending this result, our constrained partition function sums over all density-2 structures that are the union of a given (constrained) structure G and a secondary structure G'. The energy E_S is evaluated using a pseudoknot energy model (specifically, DP09). Note that this is a true generalization, reducing to the pseudoknot-free partition function of McCaskill when the structure G is empty.

Boltzmann weights, $B(e) := \exp(-e/(RT))$, and partition functions have several immediate applications in the description of the potential structures (called *ensemble*) of an RNA at equilibrium. For example, we obtain the *conditional equilibrium probability* of each structure $G \cup G'$: $\Pr(G \cup G'|G, S) = B(E(G \cup G'))/Z_S^G$, and the *ensemble free energy* $E_S^G = B^{-1}(Z_S^G) - RT \ln Z_S^G$ of the constrained ensemble.

2.4. The HFold algorithm and its ambiguity

HFold efficiently minimizes the free energy over all density-2 structures $G \cup G'$ that are hierarchically constrained by a given pseudoknot-free structure G. Like G, G' must be pseudoknot-free. Energies are defined by a D&P pseudoknot energy model for the given sequence S. As a dynamic programming (DP) algorithm, HFold can be fully defined in terms of its recurrences.

HFold computes the total minimum free energy (MFE) as the entry $W(1, n)$ of its DP matrix W, where $W(i, j)$ denotes the MFE of the subsequence $s_i s_{i+1} \ldots s_j$. Each $W(i, j)$ is computed using HFold's W-recurrence with the help of additional DP matrices. These matrices store MFEs under specific conditions: for example, $V(i, j)$ is the MFE over "closed" structures that pair i and j, $WMB(i, j)$ requires that i and j are the ends of a *pseudoloop*, and $VP(i, j)$ is the MFE over the loop closed by $i.j$ that spans a band.

Compared to pseudoknot-free prediction algorithms, HFold requires a large number of helper matrices to decompose density-2 structures and optimize correctly in the DP09 model. For instance, it distinguishes pseudoloops with the rightmost band in G' ($WMB'(i, j)$), bands in G (BE), parts of a multiloop (WI), and parts of a multiloop that span a band (WI').

Ambiguity. Several of HFold's recurrences are non-trivially ambiguous, preventing a direct translation of the HFold recurrences for CParty. A good example is the decomposition of multiloops spanning a band in the VP recurrence, as illustrated in Figure 2.

2.5. Dataset

For the time and space complexity analysis of CParty we obtained 2808 sequences from the RNASTRAND V2.0 database (Andronescu et al., 2008). The smallest sequence has a length of 8 nucleotides, while the largest is 1500 nucleotides long. For each sequence, we identified the 20 most stable stem-loop by calculating only hairpin and stacking base pair energies across the whole sequence. These stem-loop structures were then used as constraints.

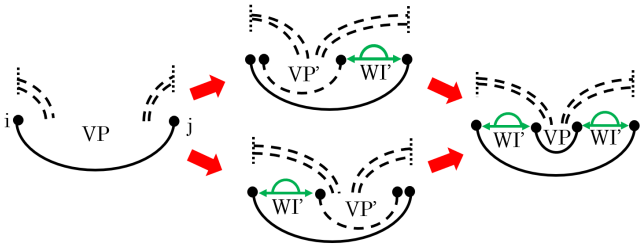

Fig. 2: The ambiguity of computing $VP(i,j)$ in HFold. Since $i.j$ is a base pair of a band, it can be crossed by other bands to the left or right (dashed arcs). To handle cases where $i.j$ closes a multiloop that spans the band, HFold utilizes two ambiguous recursion cases (middle) (Jabbari et al., 2008) to allow further multiloop branches on the left and/or right of the next base pair of the band. These different cases can converge (right) and produce the same structures in different ways, leading to ambiguity.

To assess the impact of constraint variation on CParty, we obtained 4 sequences of length 968 from the RNASTRAND V2.0 database (Andronescu et al., 2008). We generated 24 dinucleotide-shuffled versions for each sequence, resulting in a total of 100 sequences, using the MEME suite (Bailey et al., 2015). Using the same method as in the time and space complexity analysis, we generated the 20 most stable stem-loops for each sequence to be used as constraints, as well as the output of RNAfold, for a total of 21 input constraints for each sequence.

3. The CParty Algorithm

To address the partition function problem corresponding to HFold's energy minimization problem, we build on HFold's decomposition of the constrained density-2 structure space. However, the ambiguity in HFold's decomposition prevents a straightforward rewriting of the energy minimization recurrences into correct partition function recurrences by simply swapping the minimization algebra $(\min, +)$ with the 'partition function algebra' $(+, \cdot)$. Therefore, as our core contribution, we resolve all these ambiguities by carefully rewriting HFold's recurrences and introducing new structure classes and recurrences. This enhancement ensures a complete and unambiguous decomposition of the density-2 class of structures.

Here, we discuss the main recurrences of the CParty algorithm and refer readers to the Supplementary Information for a detailed explanation of the remaining recurrences.

3.1. General density-2 structures

Corresponding to the $W(i,j)$ recurrence in HFold, $Z_W(i,j)$ denotes the partition function over all density-2 structures $R_{i,j} = G_{i,j} \cup G'_{i,j}$ for the subsequence $s_i \dots s_j$ and input substructure $G_{i,j}$, taken over all choices of $G'_{i,j}$.

Call a base r *covered by* G, write $\text{isCovered}(G,r)$, iff it is *covered* by some base pair $k.\ell \in G$, i.e. $k < r < \ell$. Note that $Z_W(i,j)$ is defined only for *weakly closed* regions, where no base in the region $[i,j]$ pairs with a base outside of the region. For empty region $(i > j)$, $Z_W(i,j) = 1$—accounting for the empty structure with energy 0. Moreover, $Z_W(i,j) = 0$, if i or j is covered by G.

$$Z_W(i,j) = \sum \begin{cases} (1) & \underbrace{\sum_{i \le r < j} Z_W(i,r-1) \cdot Z_V(r,j)}_{\text{isCovered}(G,r)} \\ (2) & Z_W(i,j-1) \\ (3) & \underbrace{\sum_{i \le r < j} Z_W(i,r-1) \cdot Z_P(r,j) \cdot B(P_s)}_{\text{isCovered}(G,r)} \end{cases}$$

Figure 3 illustrates the three cases of Z_W. Case (1) decomposes the structures, where j is paired to some k in $[i,j]$; it recurses to $Z_V(i,r)$, the partition function over all structures closed by $r.j$. Case (2) handles structures where j is unpaired. Case (3) is analogous to Case (1), but r and j are left and right ends of a pseudoloop. The case recurses to $Z_P(r,j)$ (see below), and penalizes the pseudoknot initiation (P_s).

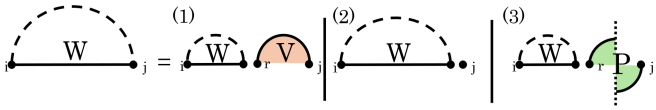

Fig. 3: $Z_W(i,j)$ recurrence in graphical notation: Dashed arcs indicate possible structure, each solid arc represents a base pair. The dotted vertical line indicates an overlapping chain of bands of arbitrary length and that the chain can begin or end via either G (above horizontal line) or G' (below horizontal line). Filled in circles show regions covered by specific structure classes, orange for Z_V, and green for Z_P.

3.2. Structures closed by a pseudoloop

The partition function over $[i,j]$ where i and j are ends of a pseudoloop is calculated as $Z_P(i,j)$. The decomposition splits of the rightmost band of the pseudoloop with ends i and j. The band can be in either G or G'. We handle the former case in the recurrence of Z_P and the latter in $Z_{PG'}$.

Figure 4 illustrates cases of the Z_P recurrence. The vertical dashed line in the figure symbolizes a series of crossing alternating bands of unspecified length.

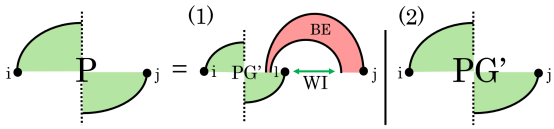

Fig. 4: Cases of Z_P. (1) j is paired in G and there must be some base, l, between $bp_G(j)$ and j that is paired in G'. (2) j is not paired in G, then move directly to $Z_{PG'}$. Filled in circles show regions covered by specific structure classes, red for Z_{BE}, and green for Z_P and $Z_{PG'}$. Detailed recurrences are provided in the Supplementary Information.

We distinguish whether j is paired in G (Case 1) or in G' (Case 2). In Case 1, each valid structure must contain a base pair in G' that crosses $bp_G(j)$, where $bp_G(j)$ denotes the base pair of j in G. This forms part of the pseudoloop. We consider all possible choices for the right end of this base pair, denoted as l. Each l determines unique inner and outer base pairs of the rightmost band (Jabbari et al., 2008).

Note that for a given G, only one case can be applicable (depending on whether j is paired in G). To maintain unambiguity, the corresponding sets of structures for different l must be disjoint, which is true for density-2

structures. Each single entry of Z_P is computed in linear time, and there is a quadratic number of entries.

3.3. Pseudoknotted structures with rightmost band in G'

The partition function of the structures closed by a pseudoloop with ends i and j and rightmost band in G' is calculated in $Z_{PG'}$ (Fig. 5).

In case (1), j pairs with l such that $l.j$ crosses a band of G. $Z_{VP}(l,j)$ accounts for the contribution of region closed by $l.j$, and Z_{BE} accounts for the contribution of the band in G. We then recurse back to $Z_{PG'}$ to consider the contribution of the rest of the structure. Case (2) is similar to case (1) with the only difference being the nested substructures allowed between the bands, which is handled by $Z_{PG'^w}$ in this case. The introduction of $Z_{PG'^w}(i,j)$ prevents multiple adjacent weakly closed subregions in the pseudoloop.

Cases (3-4) of $Z_{PG'}$ are end cases, where only one or two bands, respectively, need to be accounted for. If $i \geq j$, $Z_{PG'} = Z_{PG'^w} = 0$.

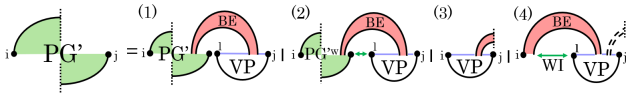

Fig. 5: Cases of $Z_{PG'}$. (1) handles two rightmost elements of the chain and continues. (2) is similar to (1) except there is a weakly closed region between the bands, this will be handled by $Z_{PG'^w}$ structure class to preserve the cubic time complexity. For the end cases we have (3) leftmost band of chain in G'; and (4) leftmost band in G. Dashed arcs indicate possible structure, each solid arc represents a base pair. Filled in circles show regions covered by specific structure classes, green for $Z_{PG'^w}$. Colored lines correspond with structure classes that may or may not have any substructures: Z_{WI} in green, and purple for Z_{VP}. Detailed recurrences are provided in the Supplementary Information.

3.4. Structures closed in G', crossing G

$Z_{VP}(i,j)$ is the partition function over all structures $R_{i,j}$ in which $i.j \in G'$ and crosses a base pair in G (Fig. 6). If $i \geq j$, i or j is paired in G, or $i.j$ does not cross any base pair of G, then $Z_{VP}(i,j) = 0$.

Cases (1-3) of $Z_{VP}(i,j)$ handle nested substructures where there are no other base pairs in $[i,j]$ that cross the same band(s) that $i.j$ crosses. These nested substructures are managed by the WI recurrence (see Supplementary Information). The three cases are disjoint: either i is covered in G (Case 1), j is covered in G (Case 2), or both are covered (Case 3). In Case (4), $i.j$ and $(i+1).(j-1)$ form a stacked pair; substructures created by $(i+1).(j-1)$ are handled recursively by Z_{VP}. In Case (5), $i.j$ and $r.r'$ close an internal loop, and we recurse back to $Z_{VP}(r,r')$ for the structures formed by $r.r'$. Cases (6-9) handle $i.j$ closing a multiloop that spans a band. In these cases, one band of the multiloop crosses the same band in G that $i.j$ crosses, and the rest of the multiloop bands and unpaired bases are handled by WI' recurrences as nested substructures. In Case (6), $r.(j-1)$ crosses the base pair in G that $i.j$ crosses, and $[i+1, r-1]$ is a non-empty weakly closed region. In Case (7), $(i+1).r$ crosses the base pair in G that $i.j$ crosses, and $[r+1, j-1]$ is a non-empty weakly closed region. In Case (8), $[i+1, r-1]$ is a non-empty weakly closed region, $r.bp(r)$ crosses the base pair in G that $i.j$ crosses, and $[bp(r)+1, j-1]$ is weakly closed. We introduce $Z^R_{VP}(i,j)$ (see the bottom left part of Fig. 6), the partition function over all structures such that $i.r \in G'$ crosses a band in G, and $r \neq j$ (distinct from Case 6). Finally, in Case (9), $[r+1, j-1]$ is a non-empty weakly closed region, $bp(r).r$ crosses the base pair in G that $i.j$ crosses, and $[i+1, bp(r)-1]$ is empty. We introduce $Z^L_{VP}(i,j)$ (see the bottom right part of Fig. 6), the partition function over all structures such that $r.j \in G'$ crosses a base pair in G, $r \neq i$ (distinct from Case 7), and $[i, r-1]$ is empty.

4. Correctness

In the following, we argue that the cases of Z_W fully decompose the density-2 structure class, and are unambiguous. The proof sketch for correctness works by structural induction, showing the correctness of each case.

Theorem 1 *The recurrence of $Z_W(i,j)$ is complete, correct, and unambiguous.*

Recall that $Z_W(i,j)$ is the partition function over the set of structures $G_{i,j} \cup G'_{i,j}$ for the subsequence $s_i \dots s_j$ taken over all choices of $G'_{i,j}$ (which is pseudoknot-free, disjoint from $G_{i,j}$, and such that $G_{i,j} \cup G'_{i,j}$ is density-2).

By definition of density-2 there are three possible cases, Case (1): j pairs with r, $i \leq r < j$, such that $r.j$ closes a pseudoknot-free loop, Case (2): j is unpaired, or Case (3): j is the rightmost end of a chain of crossing base pairs. These cases are disjoint; additionally, if j is paired and closes a pseudoknot-free loop, it cannot also be paired in the rightmost band of a pseudoloop. Therefore, the recurrence is unambiguous. Since every density-2 structure falls into one of these three cases, the $Z_W(i,j)$ recurrence is complete. Finally, it is correct, since partition functions can be correctly inferred from smaller subproblems (which are correct by induction hypothesis). ■

Similarly we have constructed each recurrence to be complete and unambiguous by construction. Of particular importance are Cases $(6-9)$ of Z_{VP} that handle a multiloop that spans a band. For a complete decomposition that preserves the $O(n^3)$ time complexity, Z^R_{VP} and Z^L_{VP} are introduced asymmetrically such that there is only one possible path to reach each structure. For example, Z_{VP} Case (8) enforces a structure somewhere in the region between i and r, and moving to Z^R_{VP} Case (1) enforces an additional structure in the subregion adjacent to j. To compare with Z_{VP} Case (9), similarly we enforce a structure somewhere in the region between r and j, but moving to Z^L_{VP} there is no possible case to introduce an additional structure adjacent to i. Thus, we avoid any ambiguity in Z_{VP} decomposition.

5. Complexity

Starting with the Z_W recurrence, we observe that its time and space complexity depend on those of Z_V and Z_P. Since Z_V handles pseudoknot-free loops, its time complexity is $O(n^3)$, and its space complexity is $O(n^2)$, where n is the length of the input sequence.

Z_P deals with pseudoloops. As Z_P matches the WMB recurrence of HFold, and HFold has been proven to have time and space complexities of $O(n^3)$ and $O(n^2)$ respectively, the same applies to Z_P. We further empirically verify Z_P's time and space complexity (see Empirical Results). Therefore, Z_W's time and space complexity remain $O(n^3)$ and $O(n^2)$, respectively.

Similarly, all other cases remain within the $O(n^3)$ time and $O(n^2)$ space complexity. For example, the time complexity for both $Z_{PG'}$ and $Z_{PG'^w}$ is $O(n^3)$, as both cases involve searching over all values of l for a given region $[i,j]$. The time complexity of Z_{VP} is dominated by the search over the region $[i,j]$ to find the value of r, which is also $O(n^3)$.

6. Empirical Results

Since CParty solves the conditional partition function for density-2 structures for the first time, it cannot be directly *benchmarked* against existing

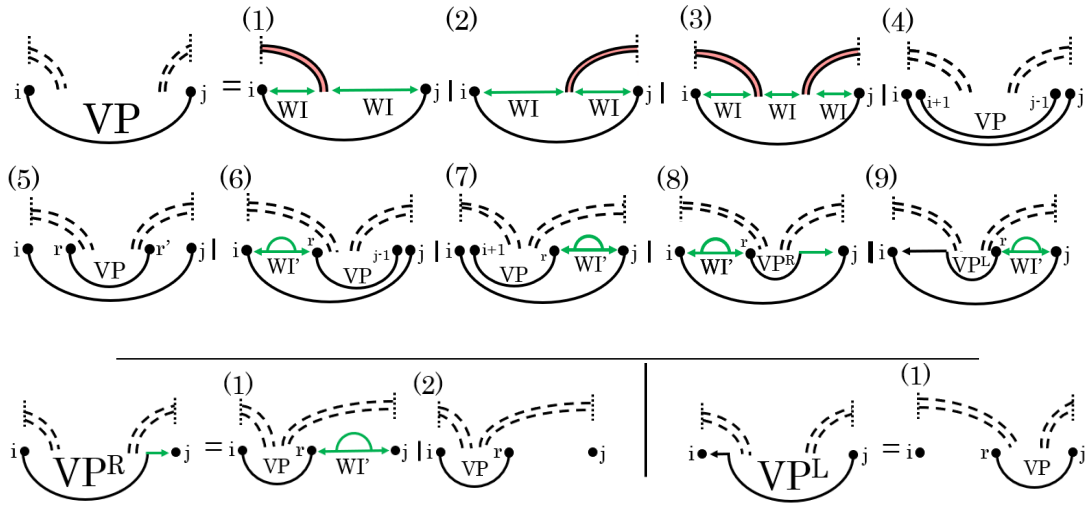

Fig. 6: Cases of VP, VP^R, and VP^L. Top: VP $(1-3)$ either two or three WI subregions (green) between i and j, band regions excluded. $(4-5)$, stacked pair and internal loop, respectively. $(6-9)$, $i.j$ closes a multiloop spanning a band. Bottom-left: VP^R, i.e., $i.bp(i)$ in G' crosses base pair in G, $bp(i) \neq j$. VP^R (1) weakly closed non-empty region $[r+1, j]$, (2) empty region $[r+1, j]$. Bottom-right: VP^L, i.e., $bp(j).j$ in G' crosses base pair in G, $bp(j) \neq i$. VP^L (1) empty region $[i, r-1]$. Dashed arcs indicate possible structure, each solid arc represents a base pair. Colored lines correspond with structure classes: Z_{WI} in green may or may not have any substructure, but for $Z_{WI'}$ which also has a green arc, there must be some substructure. Detailed recurrences are provided in the Supplementary Information.

algorithms. Nevertheless, some comparisons to RNAfold and NUPACK remain meaningful and can provide insights.

CParty and RNAfold compute identical partition functions on non-crossing structures.

Recall that in the special case of an empty input structure G, CParty computes a *pseudoknot-free partition function* Z_{pkfree}. As plausibility check, we first compared the ensemble free energy computed by CParty for Z_{pkfree} to the ensemble free energy for pseudoknot-free structures computed by RNAfold (Lorenz et al., 2011). Here, CParty perfectly reproduces the results of RNAfold (Fig. 7a), using Turner 2004 parameters (Mathews et al., 2004) without dangle energies.

The MFE variant of CParty resembles HFold.

To further validate CParty's results, we compared CParty-MFE (the MFE variant of CParty) with HFold using 945 density-2 pseudoknotted structures from the RNASTRAND database (Andronescu et al., 2008). For each sequence, we extracted a partial structure to use as input. We found that HFold and CParty predicted the same output energy for every sequence and produced identical structures in 913 out of 945 cases. In the remaining cases, the predicted structures differed but resulted in the same energy, representing alternate structures. Both programs were run using the DP09 parameters from HotKnots V2 (Andronescu et al., 2010).

These alternate structures are generated due to significant rewrites in both the codebase and the recursions used by CParty. One of the most significant changes is in how CParty traverses the matrix, facilitating the generation of alternative structures. Another major modification is the rewrite of the multiloop recurrence to permit unpaired bases on both sides of a pseudoknot, a feature not allowed in HFold.

CParty does not 'invent' pseudoknots in pseudoknot-free RNAs.

To assess the robustness of CParty against potential mispredictions of pseudoknots, we study the 24 pseudoknot-free tRNA structures from the RNASTRAND database having completely determined sequence and hairpins of at least size 3. For these RNAs, we do not expect energetically strong pseudoknotted extensions of the RNASTRAND reference, which would manifest as differences in the results from CParty and RNAfold. Demonstrating the desirable behavior of CParty, we compare the ensemble energies predicted by CParty and RNAfold, each time constrained by the reference structure, in Fig. 8.

CParty's empirical time and space outperform NUPACK's and closely match RNAfold's.

We then sought to assess the empirical time and space of computing the CParty partition function, Z, against RNAfold and NUPACK. We chose RNAfold as a benchmark for our lower bound and NUPACK as it is the only pseudoknotted partition function calculation algorithm. Since CParty requires an input structure in addition to the RNA sequence, for each sequence we identified the stem-loop structure with the lowest free energy, as detailed in Section Dataset, and used it as input to CParty and RNAfold (NUPACK's algorithm does not accept a partial structure as input), for a fair runtimes comparison. All experiments were performed on the Digital Research Alliance of Canada's Cedar cluster. We measured runtime using user time (see Fig. 9b), and memory using maximum resident set size (see Fig. 9a). The maximum time and memory used by CParty was 103.42 seconds and 117212 KB. In comparison, RNAfold had a maximum time of 36.39 seconds and 52208 KB. The expected increase in time and space usage when transitioning from pseudoknot-free to pseudoknotted structures in CParty is due to the need for new data structures and additional recurrence relations. As NUPACK requires a large amount of memory, its results were limited to sequences of max length 100. The maximum time and space for NUPACK on this subset of our dataset were 23.05 seconds and 460908 KB (see Fig. 9b and 9a in blue). As seen in Figure 9, CParty's time and space complexities closely match those of RNAfold.

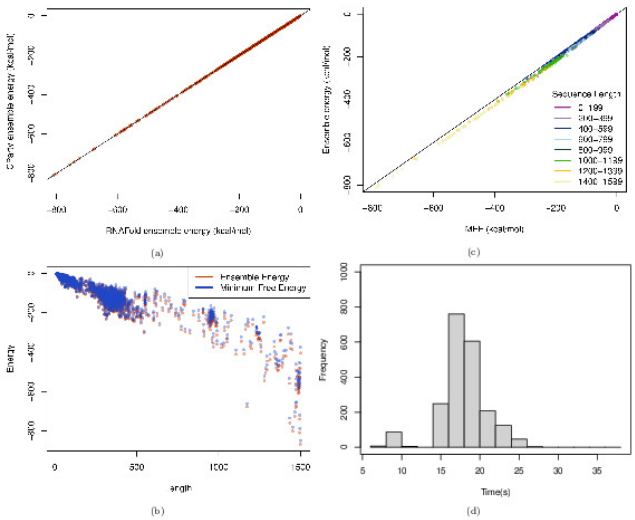

Fig. 7: We considered 2733 sequences with length up to 1500 analyzed from the RNASTRAND V2.0 database (Andronescu et al., 2008). (a) Ensemble free energies without constraints via CParty (as the y-axis) and RNAfold (as the x-axis). Agreement is observed between the two. (b) Ensemble free energies vs minimum free energy of CParty with constraints (pseudoknotted). (c) Each sequence is plotted as its ensemble energy from CParty vs the minimum free energy from HFold. Colors represent the lengths of the sequences. A diagonal line represents a 1 to 1 for ensemble energy to minimum free energy. (d) We plot the results of CParty given a sequence and an input structure. We took 4 sequences of equal length from the RNASTRAND database (Andronescu et al., 2008) and created 24 dinucleotide shuffled versions for each of them. Each sequence had 21 varying input structures; time was placed in a histogram to show the distribution given different inputs.

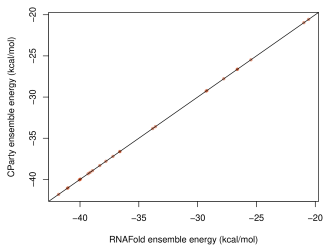

Fig. 8: Constrained ensemble free energies by CParty compared to the ones of RNAfold for 24 selected pseudoknot-free tRNAs from RNASTRAND. The virtually identical ensemble energies show that CParty does not predict strong pseudoknotted extensions in its ensemble.

Input structure has minimum effect on CParty's performance.
To assess the potential impact of the input structure on the performance of CParty, we calculated the constrained partition function on 100 sequences of length 968 with a total of 2100 various input structures, as detailed in Section Dataset. As shown in Figure 7d, little variation is observed in memory given different input structures with the 25th and 75th percentiles showing a difference of 10 KB. Figure 7d also provides a median time of 18 seconds with the 25th and 75th percentiles showing a difference of only 2 seconds. While variations in CParty's time and space usage are expected, those were not deemed significant.

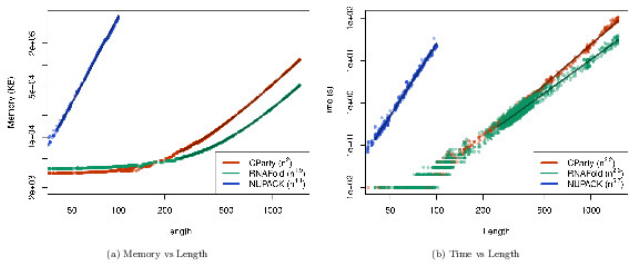

Fig. 9: Time and space consumption of CParty vs. RNAfold and NUPACK on our dataset, when given an RNA sequence and a pseudoknot-free structure as input. (a) Memory Usage (maximum resident set size in KB) versus length (log-log plot) over all benchmark instances. The solid line shows an asymptotic fit $(c_1+c_2n^x)$ for sequence length n, constants c_1,c_2, and exponent x for the fit. We ignored all values < 250 for CParty and RNAfold and all values < 40 for NUPACK. (b) Run-time (s) versus length (log-log plot) over all benchmark instances. For each tool in both plots, we report (in parenthesis) the exponent x that we estimated from the benchmark results; it describes the observed complexity as $\Theta(n^x)$.

6.1. Analysis of SARS-CoV-2 frameshift structure

There has been extensive research into predicting structure of the SARS-CoV-2 frameshift sequence, which includes both computational efforts (Schlick et al., 2021b; Trinity et al., 2023, 2024) and experimental probing experiments (Huston et al., 2021; Manfredonia et al., 2020; Yang et al., 2021; Zhang et al., 2021). The frameshift sequence is believed to form a density-2 pseudoknotted structure (Kelly et al., 2020; Schlick et al., 2021a; Jones and Ferré-D'Amaré, 2022).

Employing CParty with different fixed input structures, here we provide a view of suboptimal structures for the SARS-CoV-2 frameshift stimulating structure ensemble. Combining the available SHAPE reactivity probing datasets and various thermodynamic-based algorithms, we previously identified the top-most energetically favourable initial stems for the SARS-CoV-2 77 nucleotide frameshift pseudoknot sequence (Schlick et al., 2021b; Trinity et al., 2023, 2024). Here, we utilize the top two stems (referred to as initial stem 1 and 2) to explore the structural ensemble for the frameshift sequence. These two stems were identified as pivotal for formation of two of the main structural motifs, referred to as 3_3 and 3_6 (Schlick et al., 2021a)(see Fig. 11b).

Following the pipeline of Fig. 10, with each of the two initial stems as constraint, we employ CParty to compute ensemble free energy for sequences of decreasing length (taking 7 bases one at a time from the 5' end), to simulate the effects of the translocating ribosome (Atkins et al., 2016; Dinman, 2012).

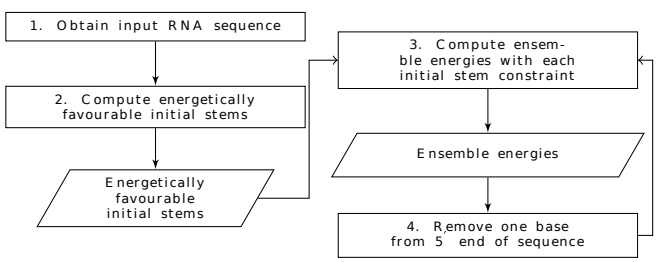

Fig. 10: CParty constrained ensemble energy pipeline. Rectangles dictate actions, parallelograms denote outputs.

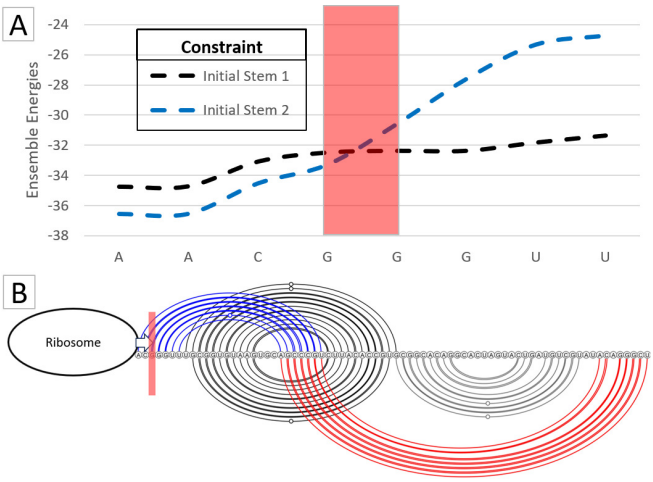

Fig. 11: SARS-CoV-2 secondary structure motif transition. (a): Constrained ensemble energies for decreasing SARS-CoV-2 sequence lengths (decreasing from 77 to 70 nt, left to right, sequence labeled on x-axis corresponding with 5' end of RNA strand shown in (b)). (b) Arcs represent base pairs. Initial stem 1 in black (included in both top and bottom pseudoknot motifs), initial stem 2 in blue, psuedoknot-free stem in green. Top arc diagram: 3_3 motif (Schlick et al., 2021a), Bottom arc diagram: 3_6 motif (also referred to as the native structure). (a & b): Red rectangles highlight the location of a transition from the 3_3 motif to the 3_6 motif. When the ribosome destabilizes the 3_3 motif base pairs (blue arcs) to the left of the red rectangle, refolding of the native-type pseudoknot (red arcs) is expected.

As seen in Fig. 11, ensemble free energies constrained by initial stems 1 and 2 are close to one another for the 77 length frameshift sequence. However, at the 5th base removal from the 5' end (see the red rectangle in Fig. 11), the ensemble free energy for initial stem 2 increases, suggesting a significant change of structural ensemble at this point. We further investigated this possible structural change using Iterative HFold (Jabbari and Condon, 2014); with frameshift sequence and initial stems 1 and 2 as input and decreasing the length by 7 bases (one base at a time) from the 5' end. We noticed that both initial stems 1 and 2 can form the 3_3 structural motif at original sequence length (77). However, at the marked transition form (red rectangle), the 3_3 motif is destabilized while the 3_6 motif maintains its stability. Therefore, the ensemble constrained by initial stem 1 is not affected. This transition observed through both ensemble free energy change as well as structurally supports the hypothesis that destabilization of initial stem 2 *facilitates* subsequent refolding of the native-type pseudoknot (Trinity et al., 2024).

7. Discussion

In this work, we introduce CParty, a novel biologically motivated algorithm that follows the hierarchical folding hypothesis to efficiently compute the constrained partition function (CPF) for density-2 RNA secondary structures. CParty takes an RNA sequence and a pseudoknot-free structure G as input and computes the CPF over all density-2 structures $G \cup G'$, where G' is pseudoknot-free and disjoint from G.

CParty was developed by addressing the ambiguities in the HFold algorithm (Jabbari et al., 2008). While HFold relies on SimFold (Andronescu et al., 2005) for pseudoknot-free energy calculations, CParty utilizes the efficient and well-maintained ViennaRNA library (Lorenz et al., 2011) and supports various energy models.

CParty handles the class of density-2 structures, which includes a wide range of pseudoknots such as kissing hairpins and interleaved bands of infinite length with arbitrarily nested substructures of the same class. By employing a hierarchical folding approach, CParty achieves a runtime complexity of $O(n^3)$ and a space complexity of $O(n^2)$. We evaluated the empirical time and space usage of CParty against RNAfold and NUPACK on a large dataset of RNA sequences of varying lengths, demonstrating CParty's efficiency in handling large RNA sequences.

Correctly identifying the input structure G is an important factor when using our algorithm. As noted in previous studies (Jabbari and Condon, 2014; Trinity et al., 2023), utilizing the most stable pseudoknot-free stem-loops is effective in identifying both the minimum free energy (MFE) structure and low-energy suboptimal structures—those energetically close to the MFE structure. Repeatedly sampling the hierarchical distribution with multiple fixed structure choices for G can help identify possible folding paths to different secondary structure motifs. Although the input structure influences our algorithm's runtime and memory usage, we found this impact to be minimal.

In this work, we demonstrated CParty's application in characterizing structural motifs in the SARS-CoV-2 frameshift element. We believe our algorithm can be similarly used in other structure-function characterizations and aid in the development of novel therapeutics.

Under hierarchical folding assumptions, CParty enables us to calculate the probability of observing a density-2 structure $G \cup G'$ at equilibrium for an RNA S as the product of the pseudoknot-free probability of G (following McCaskill's method) and the conditional probability $\Pr(G \cup G' | G, S)$.

Building on this concept, CParty supports sampling from the corresponding hierarchical structure probability distribution. While we plan to study hierarchical sampling explicitly in future work, it can be achieved through direct stochastic traceback from CParty's dynamic programming matrices. This process involves a two-step approach: first, sampling pseudoknot-free structures G (Ding and Lawrence, 2003), and then drawing from the hierarchical distribution constrained by G.

By leveraging these capabilities, CParty offers a powerful and efficient method for exploring RNA secondary structures, paving the way for further advancements in RNA research.

8. Competing interests

No competing interest is declared.

9. Author contributions statement

L.T., S.W., Y.P., U.S. and H.J. conceptualized the work. M.G. implemented the algorithm, and M.G. and L.T. conducted the experiments and analysis. All authors wrote and reviewed the manuscript.

10. Acknowledgments

We thank and acknowledge the Computational Biology Research and Analytics Lab for invaluable feedback. This work was partially supported by funding from NSERC Discovery Grant (HJ), NRC Digital Health Cluster (HJ), and ANR grant ANR-21-CE45-0034-01 "INSSANE" (SW).

References

T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.*, 104(1-3):45–62, 2000.

M. Andronescu, Z. C. Zhang, and A. Condon. Secondary structure prediction of interacting RNA molecules. *Bioinformatics*, 345:987–1001, 2005. doi: 10.1016/j.jmb.2004.10.082.

M. Andronescu, V. Bereg, H. H. Hoos, and A. Condon. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinf.*, 9(1):1–10, 2008.

M. S. Andronescu, C. Pop, and A. E. Condon. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA*, 16 (1):26–42, 2010.

J. F. Atkins, G. Loughran, P. R. Bhatt, A. E. Firth, and P. V. Baranov. Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.*, 44 (15):7007–7078, 2016.

T. Bailey, J. Johnson, C. E. Grant, and W. S. Noble. The meme suite. *Nucleic Acids Research*, 43(W1):W39–W49, November 2015. doi: 10.1093/nar/gkv416.

S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. Rnaalifold: improved consensus structure prediction for rna alignments. *BMC Bioinf.*, 9(1):1–13, 2008.

G. Chen, K.-Y. Chang, M.-Y. Chou, C. Bustamante, and I. Tinoco. Triplex structures in an RNA pseudoknot enhance mechanical stability and increase efficiency of–1 ribosomal frameshifting. *PNAS*, 106(31): 12706–12711, 2009a.

H.-l. Chen, A. Condon, and H. Jabbari. An o(n5) algorithm for mfe prediction of kissing hairpins and 4-chains in nucleic acids. *Journal of Computational Biology*, 16:803–815, 2009b. doi: 10.1089/cmb.2008.0219.

H.-L. Chen, A. Condon, and H. Jabbari. An o (n^5) algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids. *J. Comput. Biol.*, 16(6):803–815, 2009c.

S. S. Cho, D. L. Pincus, and D. Thirumalai. Assembly mechanisms of RNA pseudoknots are determined by the stabilities of constituent secondary structures. *PNAS*, 106(41):17349–17354, 2009.

J. A. Cruz and E. Westhof. The dynamic landscapes of RNA architecture. *Cell*, 136:604–609, Feb 2009. doi: 10.1016/j.cell.2009.02.003.

Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, 31:7280–7301, Dec. 2003. ISSN 1362-4962.

J. D. Dinman. Mechanisms and implications of programmed translational frameshifting. *Wiley Interdiscip. Rev. RNA*, 3(5):661–673, 2012.

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

W. Fontana, D. A. Konings, P. F. Stadler, and P. Schuster. Statistics of rna secondary structures. *Biopolymers*, 33(9):1389–1404, 1993.

R. Giegerich. A systematic approach to dynamic programming in bioinformatics. *Bioinformatics*, 16(8):665–677, Aug. 2000. ISSN 1367-4803. doi: 10.1093/bioinformatics/16.8.665.

C. Hasslinger and P. F. Stadler. RNA structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties. *Bull. Math. Biol.*, 61: 437—467, 1999. doi: 10.1006/bulm.1998.0085.

N. C. Huston, H. Wan, M. S. Strine, R. d. C. A. Tavares, C. B. Wilen, and A. M. Pyle. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Mol. Cell*, 81(3):584–598, 2021.

H. Jabbari and A. Condon. A fast and robust iterative algorithm for prediction of RNA pseudoknotted secondary structures. *BMC Bioinf.*, 15(1): 147, 2014.

H. Jabbari, A. Condon, A. Pop, C. Pop, and Y. Zhao. Hfold: Rna pseudoknotted secondary structure prediction using hierarchical folding. In R. Giancarlo and S. Hannenhalli, editors, *Algorithms in Bioinformatics*, pages 323–334, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-74126-8. doi: 10.1007/978-3-540-74126-8_30.

H. Jabbari, A. Condon, and S. Zhao. Novel and Efficient RNA Secondary Structure Prediction Using Hierarchical Folding. *J. Comput. Biol.*, 15(2): 139–163, Mar. 2008. doi: 10.1089/cmb.2007.0198.

H. Jabbari, I. Wark, C. Montemagno, and S. Will. Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. *Bioinformatics*, 34:3849–3856, Nov 2018. doi: 10.1093/bioinformatics/bty420. URL `https://doi.org/10.1093/bioinformatics/bty420`.

C. Jones and A. R. Ferré-D'Amaré. Crystal structure of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) frameshifting pseudoknot. *RNA*, 28:437—467, 2022. doi: 10.1261/rna.078825.121.

J. A. Kelly, A. N. Olson, K. Neupane, S. Munshi, J. San Emeterio, L. Pollack, M. T. Woodside, and J. D. Dinman. Structural and functional conservation of the programmed- 1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *J. Biol. Chem.*, 295(31):10741–10748, 2020.

M. Kozak. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, 361:13–37, Nov 2005. doi: 10.1016/j.gene.2005.06.037.

R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. Viennarna package 2.0. *Algorithms Mol. Biol*, 6:1–14, 2011.

R. B. Lyngsø and C. N. Pedersen. Pseudoknots in RNA secondary structures. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 201–209, New York, NY, United States, 2000. Association for Computing Machinery.

I. Manfredonia, C. Nithin, A. Ponce-Salvatierra, P. Ghosh, T. K. Wirecki, T. Marinus, N. S. Ogando, E. J. Snijder, M. J. van Hemert, J. M. Bujnicki, et al. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res.*, 48(22): 12436–12452, 2020.

D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *PNAS*, 101(19):7287–7292, 2004.

J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers: Original Research on Biomolecules*, 29(6-7):1105–1119, 1990.

S. A. Mortimer, K. M. Anne, and J. A. Doudna. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics*, 15: 469—479, May 2014. doi: 10.1038/nrg3681.

Y. Ponty and C. Saule. A combinatorial framework for designing (pseudoknotted) RNA algorithms. In M.-F. S. T. Przytycka, editor, *Algorithms in Bioinformatics*, number 6833 in LNBI, pages 250–269, Saarbrücken, Germany, Jan. 2011. Springer. ISBN 978-3-642-23037-0.

B. Rastegari and A. Condon. Parsing nucleic acid pseudoknotted secondary structure: algorithm and applications. *J. Comput. Biol.*, 14(1):16–32, 2007.

J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinf.*, 5(104):1–12, 2004.

J. S. Reuter and D. H. Matthews. RNAstructure: software for RNA secondary structure prediction and analysis. *Bioinformatics*, 11, Mar 2010. doi: 10.1186/1471-2105-11-129.

E. Rivas. RNA structure prediction using positive and negative evolutionary information. *PLoS Comput Biol.*, 16(10):1–25, Oct 2020. doi: 10.1371/journal.pcbi.1008387.

E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285(5):2053–2068, 1999.

K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinf.*, 27(13):i85–i93, 2011.

T. Schlick, Q. Zhu, A. Dey, S. Jain, S. Yan, and A. Laederach. To knot or not to knot: multiple conformations of the SARS-CoV-2 frameshifting RNA element. *J. Am. Chem. Soc.*, 143(30):11404–11422, 2021a.

T. Schlick, Q. Zhu, S. Jain, and S. Yan. Structure-altering mutations of the SARS-CoV-2 frameshifting RNA element. *Biophys. J.*, 120(6):1040–1053, 2021b.

S. Sheikh, R. Backofen, and Y. Ponty. Impact of the energy model on the complexity of rna folding with pseudoknots. In *Annual Symposium on Combinatorial Pattern Matching*, pages 321–333. Springer, 2012.

I. Tinoco Jr and C. Bustamante. How RNA folds. *J. Mol. Biol.*, 293(2):271–281, 1999.

L. Trinity, I. Wark, L. Lansing, H. Jabbari, and U. Stege. Shapify: Paths to SARS-CoV-2 frameshifting pseudoknot. *PLoS Comput. Biol.*, 19(2):e1010922, 2023.

L. Trinity, U. Stege, and H. Jabbari. Tying the knot: Unraveling the intricacies of the coronavirus frameshift pseudoknot. *PLOS Computational Biology*, 20(5):e1011787, 2024.

D. H. Turner and D. H. Matthews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38:D280—-D282, 2009. doi: 10.1093/nar/gkp892.

M. B. Warf and J. A. Berglund. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci.*, 35:169–178, Mar 2010. doi: 10.1016/j.tibs.2009.10.004.

T. J. Wilson and D. M. Lilley. RNA catalysis—is that it? *RNA*, 21:534–537, Apr 2015. doi: 10.1261/rna.049874.115.

C. Witwer, I. L. Hofacker, and P. F. Stadler. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 1(2):66–77, 2004.

S. L. Yang, L. DeFalco, D. E. Anderson, Y. Zhang, J. G. A. Aw, S. Y. Lim, X. N. Lim, K. Y. Tan, T. Zhang, T. Chawla, et al. Comprehensive mapping of SARS-CoV-2 interactions in vivo reveals functional virus-host interactions. *Nat. Commun.*, 12(1):1–15, 2021.

K. Zhang, I. N. Zheludev, R. J. Hagey, R. Haslecker, Y. J. Hou, R. Kretsch, G. D. Pintilie, R. Rangan, W. Kladwang, S. Li, et al. Cryo-EM and antisense targeting of the 28-kDa frameshift stimulation element from the SARS-CoV-2 RNA genome. *Nat. Struct. Mol. Biol.*, 28(9):747–754, 2021.

M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9 (1):133–148, 1981.