CParty: Conditional Partition Function for Density-2 RNA Pseudoknots

Luke Trinity[1][0000−0002−1578−7964], Sebastian Will[2][0000−0002−2376−9205], Yann Ponty[2][0000−0002−7615−3930], Ulrike Stege[1][0000−0001−9466−7196], and Hosna Jabbari[3][0000−0002−7155−2297]

[1] University of Victoria, Victoria, British Columbia, Canada ltrinity@uvic.ca
[2] Institut Polytechnique de Paris, Paris, France
[3] University of Alberta, Edmonton, Alberta, Canada
jabbari@ualberta.ca

Abstract. RNA molecules fold into biologically important functional structures. Efficient dynamic programming RNA (secondary) structure prediction algorithms restrict the search space to evade NP-hardness of general pseudoknot prediction. While such prediction algorithms can be extended to provide a stochastic view on RNA ensembles, they are either limited to pseudoknot-free structures or extremely complex. To overcome this dilemma, we provide the theoretical framework and implementation for our algorithm, *CParty*, that follows the hierarchical folding hypothesis, i.e., the bio-physically well-motivated assumption that non-crossing structures fold relatively fast prior to the formation of pseudoknot interactions. Thus, we efficiently compute the conditional partition function (CPF) given a non-crossing structure G for a subset of pseudoknotted structures, i.e., density-2 structures $G \cup G'$ for non-crossing disjoint structure G'. Notably, this can enable sampling from the hierarchical distribution $P(G'|G)$. With CParty we develop for the first time an *unambiguous* scheme based on *HFold*, i.e., the minimum free energy hierarchical folding algorithm based on a realistic pseudoknot energy model. Thus, we develop the first partition function variant for density-2 structures. Compared to the only other available pseudoknot partition function algorithm, which covers simple pseudoknots (and follows a different strategy, mapped from a pure minimum free energy structure prediction), our method covers a much larger structure class; at the same time, it is significantly more efficient—reducing the time as well as the space complexity by a quadratic factor. Summarizing, we provide an efficient, cubic time, algorithm for the stochastic analysis of pseudoknotted RNAs, which enables novel applications. We discuss one such application, i.e., how the CPF for a pseudoknotted therapeutic target in SARS-CoV-2 can provide insight into RNA structure formation.

Keywords: RNA structure prediction, free energy, partition function, pseudoknots

1 Introduction

RNA molecules play a vital role in cellular processes; many possess functional structures. Due to the costly and time-consuming nature of experimental methods to detect RNA structure, computational tools to predict the formation of RNA structure conserve valuable resources. Our focus is on the prediction of RNA structure via a 2D model, known as secondary structure, whose accurate prediction can also inform state-of-the-art 3D prediction tasks.

RNA secondary structure prediction involves finding a set of base pairs for a given input RNA sequence, such that each base appears in at most one pair [19]. When an upstream RNA segment loops back and forms base pairs with a downstream segment, disrupting their linear alignment, we refer to these as crossing base pairs. We call structures with crossing base pairs, 'pseudoknotted'. Two widely used thermodynamic approaches for predicting non-crossing (pseudoknot-free) RNA secondary structure are the ones by Zuker [40] and McCaskill [22]. The Zuker algorithm uses dynamic programming (DP) to recursively compute the free energy of all possible RNA substructures and find the most energetically stable, the minimum free energy (MFE) structure [40]. To further improve their prediction accuracy, MFE-based prediction methods can be extended to integrate experimental data (e.g., SHAPE profiles) as soft-constraints [16].

Adopting a stochastic view of RNA, the McCaskill algorithm computes the *partition function* that describes the pseudoknot-free secondary structure of an arbitrary length RNA. In order to efficiently calculate the partition function, the algorithm must carefully consider the equilibrium ensemble of structures: all possible base pairs and associated energetic contributions from each type of *loop* (hairpin, internal/bulge, stacked pair, and multiloop). The partition function is useful for predicting the relative probability of any structure as well as base pair probabilities [21,22]. The likelihood of any particular RNA structure occurring can be determined based on the energy of the structure itself relative to the total energy in the system.

The hierarchical folding hypothesis postulates that RNA first folds into a simple non-crossing structure, then additional (possibly crossing) bases pair to lower the MFE of the structure [31]. In particular, hierarchical folding has been observed in formation of crossing structures [7] including frameshift stimulating pseudoknots [5]. Hierarchical folding can also be thought of as a special kind of kinetics, where the initial non-crossing base pairs form more quickly, followed by crossing or non-canonical base pairs. Assuming fast convergence to the equilibrium over non-crossing structures, and an overly large energy barrier, the probability to observe a given crossing structure can be expressed as the product of the classic Boltzmann probability $P(G)$ of the non-crossing subset G, and the conditional probability $P(G'|G)$. If a Boltzmann-distributed non-crossing structure is irrevocable, then the distribution to consider is not the usual Boltzmann distribution:

$$\mathbb{P}(G, G') = \mathbb{P}(G \cup G') = \frac{e^{-\beta(E_{2D}(G)+E_{PK}(G'))}}{\sum_{H \cup H'} e^{-\beta(E_{2D}(H)+E_{PK}(H'))}}$$

but instead the *hierarchical distribution:*

$$\mathbb{P}(G) \times \mathbb{P}(G' \mid G) = \frac{e^{-\beta E_{2D}(G)}}{\sum_{H} e^{-\beta E_{2D}(H)}} \times \frac{e^{-\beta E_{PK}(G')}}{\sum_{H \cup H'} e^{-\beta E_{PK}(H')}}$$

$E_{2D}(G)$ and $E_{PK}(G')$ are the energies associated with the non-crossing and crossing structures of the RNA, respectively. β is the inverse temperature factor given the Boltzmann constant, and the usual summation is over all possible structures with and without crossing base pairs, respectively. Here, the hierarchical distribution probability coincides with the emission probability of structures generated by a process that first samples non-crossing structures in the Boltzmann distribution $\mathbb{P}(G)$ using stochastic backtrack [8], and then samples from the conditional distribution $\mathbb{P}(G' \mid G)$. However, sampling from the conditional distribution requires computing the *conditional partition function* (CPF), followed by an adaptation into a conditional version of the stochastic backtrack. In this work, we develop and implement the algorithmic framework for calculating the CPF: *CParty*.

There is a one-to-one correspondence between the search spaces considered by MFE algorithms and McCaskill partition function methods [22], provided the underlying DP scheme is unambiguous and complete [23]. This relationship only holds if each substructure is unambiguously computed a single time during the DP MFE calculations [22], i.e., redundant recurrence relation schemes are incompatible with the partition function approach.

Examples of pseudoknotted MFE predictions methods that can be extended to the partition function approach include the most general algorithm PKnots [26], with $O(n^6)$ time complexity, also NUPACK [37] and CCJ [6] both with $O(n^5)$ time complexity, the latter allows for more complex structures like four interleaved *stems* (consecutive adjacent base pairs). Above-mentioned algorithms that can predict complex RNA structures are slow, motivating novel partition function methods that handle a pseudoknotted RNA structure class efficiently. Calculating a partition function for the general pseudoknotted case is NP-hard [1,17], therefore pseudoknotted partition function methods require assumptions that limit the complexity of predicted pseudoknots [10]. Heuristic methods or methods that rely on approximations, such as HotKnots V2.0 [2] and IPKnots [28], cannot be extended to compute partition function, although both methods handle a broad class of pseudoknotted structures efficiently.

Here, we develop an unambiguous recurrence relation scheme based on the hierarchical-folding algorithm *HFold* [13], which for a given non-crossing structure G, identifies a non-crossing structure G' that is disjoint from G. Specifically, HFold finds the *density-2* structure $G \cup G'$ that is the MFE structure given G (see [13] for formal definitions, for intuition cf. Appendix Fig.8-9). The density-2 structure class allows for arbitrary depth and length of nested crossing base pairs including H-type pseudoknots and kissing hairpins. This class of structures includes some structures not handled by CCJ, and is larger and more comprehensive than the structure class described by the partition function algorithm from Dirks and Pierce [10]. HFold has a time complexity of $O(n^3)$, significantly improving over PKnots, CCJ, and NUPACK.

The HFold recurrence relation scheme is redundant in decomposing density-2 structures for a given non-crossing structure G, making it incompatible with a partition function approach. Our novel algorithm, *CParty*, unambiguously decomposes each possible density-2 structure, ensuring a one-to-one mapping to calculating the partition function. CParty computes the CPF with respect to the input structure, which can be used to obtain conditional base pair probabilities and unveil RNA kinetic structure formation paths. We note that choosing input structure G is of particular importance, since it is impossible to reach the MFE structure when the fixed structure G is not part of the MFE structure. Previous work [12,32] identified promising techniques to reach the MFE structure by selecting energetically favourable initial stems, i.e., most stable non-crossing structures, or choosing stems compatible with chemical modification data (such as SHAPE reactivity).

Although a 'full' partition function is ultimately desirable, our focus on a CPF algorithm for density-2 pseudoknotted structures allows us to achieve cubic run time complexity. A CPF is particularly useful in cases where partial information is available for a given sequence. Different structure reactivity datasets may align for specific subparts of a functional structure, e.g., initial stems, while the rest of the structure is unknown [32]. As functional RNA structures are solved at higher resolution [27], significant differences remain in predictions (e.g., length dependent conformational flexibility [35]). Conditional base pair probabilities obtained via CParty implementation can enable analyses exploring key hierarchical folding paths.

Contributions. We introduce the CPF algorithm CParty and demonstrate that our algorithm completely and unambiguously decomposes the density-2 structure class. CParty is available at github.com/ltrinity/CParty. CParty can be utilized to identify high probability base pairs and unveil changes in base pair probability relative to the most likely initial folding of an RNA molecule. Computing the CPF for density-2 structures with CParty is also useful when the non-crossing structure G is unknown for a specific RNA, or when there are multiple candidate choices for G. The CPF obtained by CParty informs the choice of a best candidate for non-crossing structure G, explaining kinetic paths in a novel way. Integrating experimental data from various sources (e.g., SHAPE reactivity, crystallography, cryo-EM) within this CPF approach can enable further statistical analysis for the ensemble of density-2 secondary structures. We present an immediate application for the algorithm: analysis of the SARS-CoV-2 frameshift pseudoknot structure [14]. CParty provides the framework for additional functional structure path analyses, e.g., computing conditional base pairing probabilities towards therapeutic development in coronaviruses.

Outline. In Section 2, we define the CPF problem and relevant variants. Then, we provide a brief notation review. In Section 3, we proceed with a high level description of CParty and define the constrained partition subfunctions, proving correctness of the DP scheme. In Section 4 we provide details specific to implementation, validation, and testing of CParty. In Section 5, we discuss and contextualize how CParty can inform RNA structure formation analysis, providing an application to the SARS-CoV-2 frameshift pseudoknot sequence.

2 Preliminaries and problem statement

We define energy $E(G)$ of an RNA secondary structure G according to the realistic pseudoknot energy model of [10] with parameters from [2]. For an RNA sequence S, these energies induce a *Boltzmann distribution* of its structures that characterizes its equilibrium ensemble. In this distribution, the frequency of each structure G of S is proportional to its Boltzmann weight $B(E(G))$. This weight is defined as $B(x) = e^{x/(KT)}$ at a given temperature T; K is the universal gas constant. The sum over the Boltzmann weights of the structures of an RNA S is known as its *partition function* Z. More concretely, partition functions are defined for RNAs and specific structure classes Ω, e.g., all density-2 structures, as $Z = \sum_{s \in \Omega} B(E(s))$. Finally, the *ensemble probability* of a structure is $\Pr[G] = \frac{B(E(G))}{Z}$.

Problem Statement. Given an RNA sequence S and a non-crossing structure G, the *Conditional Partition Function problem (CPF problem)* is to compute the partition function Z for the set of all RNA secondary structures $G \cup G'$ of S, where G' is a non-crossing RNA structure that is disjoint from G and $G \cup G'$ is *density-2* (definition below). Equivalently: $\Pr[G'|G] = \frac{B(E(G' \cup G|G))}{Z}$, where Z is computed over all choices of G' for a given fixed structure G.

An interesting variant of the CPF problem forces all base pairs of G' to cross those of G. We expect this 'pseudoknot-only' variant of CParty can be derived relatively easily from our work.

Notation. We briefly review the required definitions from HFold to define and predict density-2 structures [13]. Density-2 structures are a subset of the *bi-secondary structure* class, which is the union of two disjoint pseudoknot-free secondary structures [33], e.g., $G \cup G'$. Here, we refer to bases in the given sequence by their indices, e.g., if the input is 'AUCG', base 0 is A, base 1 is U, and so on. If i is paired with j, then $i.j$ is a *base pair*. A *structure*, R, is a set of base pairs for an RNA sequence over alphabet {A, U, C, G}. The index of the base paired with base at index i in structure R is $bp_R(i)$.

To optimize computation and properly allow nested and non-nested base pairs, we define regions that are self-contained and can or cannot be bifurcated (split). Of key interest in the algorithm design are pseudoloops, which we formally define beginning with crossing base pairs and then carefully grouping these pseudoknotted base pairs into bands and associated regions. Density of a structure is calculated with respect to all pseudoloops in the structure. Abbreviated nomenclature included below.[4]

A **region** $[i, j]$ is a sequence of indices between i, the left border, and j, the right border, inclusive. A region is **weakly closed** if $\forall k \in [i, j]$ either $bp_R(k) \in [i, j]$ or $bp_R(k) = 0$ (unpaired). A region $[i, j]$ is **closed** if $[i, j]$ is weakly closed, and $\forall l \in [i, j-1]$, neither $[i, l]$ nor $[l+1, j]$ is weakly closed [24]. Base pair $i.j$ **covers** k if $i < k < j$ and $\nexists i'.j'$ such that $i < i' < k < j' < j$.

If $\exists i, i', j, j'$, such that $i.j$, $i'.j'$, and $i < i' < j < j'$, then $i.j$ **crosses** $i'.j'$. Now, $i.j$ is **pseudoknotted** if $\exists i'.j'$ such that $i.j$ crosses $i'.j'$. Pseudoknotted $i.j$ is **directly banded in** $i'.j'$, denoted $i.j \preceq i'j'$, if $i' \leq i < j \leq j'$ and $[i'+1, i-1]$ and $[j+1, j'-1]$ are weakly closed regions. Consider a maximal chain of $\preceq$, if $i.j$ is the outermost pair, and $i'.j'$ is the innermost pair, then $[i, i'] \cup [j'.j]$ is the **band** region. Bands, base pairs, or unpaired bases **associated with closed region $[i, j]$** are not in any closed region or band which are subregions of $[i, j]$ (cf. Fig. 1). A **pseudoloop** refers to the unpaired bases and base pairs associated with pseudoknotted closed region $[i, j]$, together with the closing base pairs of the bands associated with $[i, j]$.

Let L be a pseudoloop and $i.bp_R(i)$ and $bp_R(j).j$ be the closing base pairs of L. Band $[i_1, i'_1] \cup [j'_1, j_1]$ crosses k if $i_1 \leq k \leq j_1$. Let $\#B(L, k)$ be the number of bands associated with L that cross k. Then the **density** of L is: $max_{i \leq k \leq j}(\#B(L, k))$. It follows that the *density* of a structure, R, is the maximum density of L over all *pseudoloops* L of R. R is *density-k* if the density of R is at most k.

3 Efficient computation of the conditional partition function

HFold review. In the HFold hierarchical folding algorithm, $W(i, j)$ denotes the energy of the MFE density-2 secondary structure for a given subsequence $s_i s_{i+1} \ldots s_j$ and a given pseudoknot-free structure $G_{i,j}$ taken over all choices of pseudoknot-free $G'_{i,j}$ which is disjoint from $G_{i,j}$, and such that $G_{i,j} \cup G'_{i,j}$ is of density-2. W is computed via structure class $V(i, j)$ with $i.j$, and structure class $WMB(i, j)$, where i and j form a

[4] See [13] for full definitions.

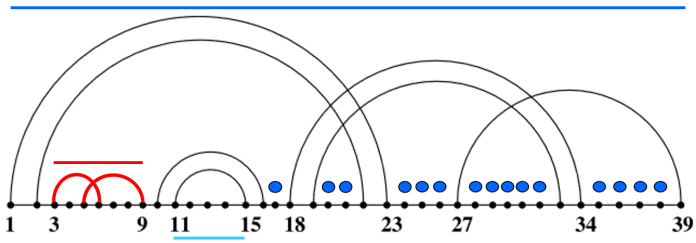

Fig. 1. The three *bands* $[1,2] \cup [22,23]$, $[18,19] \cup [33,34]$ and $[27,27] \cup [39,39]$ as well as unpaired bases not in closed subregions (c.f. blue dots •) are associated with closed pseudoknotted region $[1,39]$, i.e., density-2 pseudoloop $[1,39]$. Closed regions $[3,9]$ and $[10,16]$ are associated with pseudoloop $[1,39]$ but closed region $[11,15]$ is *not*. Bands $[3,3] \cup [6,6]$ and $[5,5] \cup [9,9]$ are associated with closed pseudoknotted region $[3,9]$, i.e., density-2 pseudoloop $[3,9]$. Note that bases $12-14$ are covered by 11.15 but not 10.16. Figure modified from [13].

pseudoloop. $WMB(i,j)$ is further classified to distinguish pseudoloops with rightmost band in G' as structure class $WMB'(i,j)$. Each successive band in a pseudoloop is computed via structure classes BE for bands in G, and VP for bands in G'. Following the approach of HFold, our CParty algorithm utilizes additional structure classes to ensure unambiguous decomposition of the density-2 structure class (cf. Table 1). We include intuition for runtime analysis of CParty structure classes, for formal proof refer to HFold [13].

We now introduce the CParty DP scheme in detail, conditions for terms in summation are shown in blue. CParty utilizes the DP09 energy model of HotKnots V2.0 [2], with parameters matching the pseudoknotted energy model of Dirks and Pierce [10] that were re-estimated using available pseudoknotted structures. Due to space constraints, energy parameter table (cf. Appendix Table 3) and complete structure class equations for PG', PG'^w, VP, VP^R, VP^L, VM, WM, WM^1, WM^P are described in the Appendix, Sections A.1-A.4. Structure classes with straightforward transition from HFold (WI, WI', V, BE) are also included in the Appendix, Sections A.5-A.8.

Table 1. CParty Structure Classes. The first column denotes each structure class abbreviation with description in second column. In the third column we delineate additional structure classes that need to be computed, e.g., W is computed via V where $i.j$, via P where i and j form a pseudoloop, and also by recursive call again to W.

Structure Class	*Description*	*Computed via*
W	General density-2 structures	W, V, P
V	$i.j$	VM
P	Pseudoknotted density-2 structures	PG', BE, WI
PG'	Pseudoknotted rightmost band in G'	PG', PG'^w, VP, BE, WI
PG'^w	Pseudoknotted rightmost band in G', weakly closed region between bands	PG', WI
WI	Weakly closed region inside pseudoloop	V, P, WI
WI'	Non-empty weakly closed region inside band	V, P, WI'
VP	$i.j$ in G' crosses base pair in G	VP, VP^R, VP^L, WI, WI'
VP^R	$i.bp(i)$ in G' crosses base pair in G, bp(i) $\neq j$	VP, WI'
VP^L	$bp(j).j$ in G' crosses base pair in G, bp(j) $\neq i$	VP
VM	$i.j$ close a multiloop	WM, WM^1, WM^P
WM	$i.j$ on a multiloop	WM, V, P, WM^P
WM^1	$i.j$ on a multiloop, terminal stem pseudoknot-free	WM^1, V
WM^P	$i.j$ on a multiloop, terminal stem pseudoknotted	WM^P, P
BE	$[i,i'] \cup [bp(i'), bp(i)]$ band region	BE, WI'

3.1 General density-2 structures

$Z_W(i,j)$ denotes the partition function over all 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) and i and j are not covered by a base pair in G, denoted as $\overline{isCovered(G,i)}$ and $\overline{isCovered(G,j)}$. Note that $Z_W(i,j)$ is weakly closed and can possibly be partitioned into two independent substructures. Base cases are as follows: if $i \geq j$, $Z_W(i,j) = 1$, since the substructure is empty and the only possibility for structure $R_{i,j}$ is the empty structure. In addition, $Z_W(i,j) = 0$, if $isCovered(G,i)$ or $isCovered(G,j)$, to enforce the exterior structure condition.

$Z_V(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed or empty and $i.j$. $Z_P(i,j)$ is the partition function over all pseudoknotted density-2 structures between i and j where the structure is closed (cf. Section 3.2).

$$Z_W(i,j) = \sum \begin{cases} (1) \sum\limits_{\substack{i \leq r < j \\ \overline{isCovered(r)}}} Z_W(i,r-1) \cdot Z_V(r,j) \\ (2) \quad Z_W(i,j-1) \\ (3) \sum\limits_{\substack{i \leq r < j \\ \overline{isCovered(r)}}} Z_W(i,r-1) \cdot Z_P(r,j) \cdot B(P_s) \end{cases} \tag{i}$$

Case (1) bifurcates such that $\exists r$, $i \leq r < j$, $bp_R(i) \leq r-1$ (i.e., i is either unpaired or paired with another base inside region $[i, r-1]$), and $bp_R(j) = r$ (i.e., j is paired with r).

Case (2) enforces j as unpaired: $bp_R(j) = 0$ and $bp_R(i) \leq r$ (i.e., i is either unpaired or paired with another base inside region $[i, j-1]$).

Case (3) bifurcates such that $bp_R(i) \leq r-1$ (i.e., i is either unpaired or paired with another base inside region $[i, r-1]$), and $\exists q, x$ such that $r < q < x < j$, $bp_R(r) = x$ and $bp_R(j) = q$ (i.e., $r.x$ crosses $j.q$). In this case, we add a penalty, P_s, for introducing an exterior pseudoknot.

Since Z_V handles pseudoknot-free loops, its time and space complexity are $O(n^3)$ and $O(n^2)$, respectively, where n is the length of the input sequence. Time and space complexity of Z_W, therefore, depend on the time and space complexity of Z_P, which handles pseudoloops.

Z_W fully decomposes the density-2 structure class (cf. Fig. 2), and is unambiguous to ensure compatibility with the partition function approach. The proof of correctness works by structural induction, showing the correctness of each single recurrence. Proving the correctness of $Z_W(i,j)$ serves as an example.

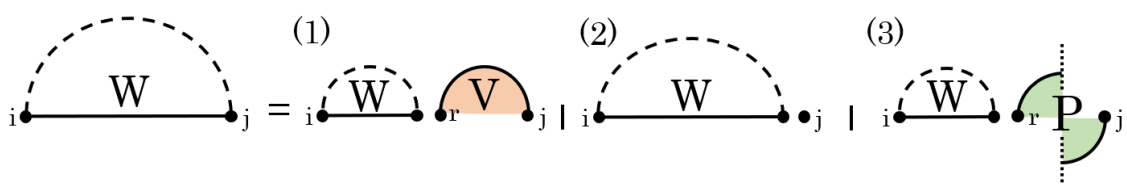

Fig. 2. $Z_W(i,j)$ structure class. Case (1) illustrating $r.j$ close a loop. Case (2) enforces j unpaired. Finally for Case (3) $r.j$ form a pseudoloop. Dotted arcs indicate possible structure, each solid arc represents a base pair. The dotted vertical line indicates an overlapping chain of bands can continue 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.

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

$Z_W(i,j)$ 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, j either pairs with k, $i \leq k < j$ such that $k.j$ closes a pseudoknot-free loop (1), j is unpaired (2), or j is the rightmost end of a chain of crossing base pairs (3), cf. Eq. i. These cases are disjoint, i.e., if j is unpaired it cannot also be paired; additionally, if j is paired and closes a pseudoknot-free

loop, it cannot also be paired in the rightmost band of a pseudoloop. Therefore, all three cases are disjoint, and 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). ■

3.2 Pseudoknotted density-2 structures

$Z_P(i,j)$ is the partition function over $[i,j]$ when $[i,j]$ is a pseudoknotted closed region (cannot be further partitioned), containing a chain of two or more successively overlapping bands that must alternate between $G_{i,j}$ and $G'_{i,j}$. The rightmost band of the pseudoloop may be in G or G', possibly with nested substructures interspersed throughout.

For $Z_P(i,j)$ base case, if $i \geq j$, then $Z_P(i,j) = 0$, since the substructure is empty. Otherwise, there are two cases (cf. Fig. 3), Case (1): j is paired in G, or Case (2): j is not paired in G. If j is paired in G, then in the MFE structure, some base l with $bp_G(j) < l < j$ must be paired in G', causing $bp_G(j).j$ to be pseudoknotted. We consider all possible choices of l. Once l is fixed, the inner base pair of the band whose outer base pair is $bp_G(j).j$ is also determined (e.g., $b_{(i,l)}$ or $b'_{(i,l)}$ cf. [13]). If j forms a pseudoloop with i and is not paired in G, we move to $Z_{PG'}$ Case (2), partition function over pseudoknotted density-2 structures with rightmost band in G'. Here we note these two cases are disjoint and Z_P is unambiguous, with j either paired in G or G'.

Since in the region $[i,j]$ for Z_P Case (1) we search over all possible choices of l, time complexity is $O(n^3)$. Therefore, time complexity of Z_P will be maximum of $O(n^3)$ and time complexity of $Z_{PG'}$.

$$Z_P(i,j) = \sum \begin{cases} (1) \displaystyle\sum_{\substack{bp_G(j)<l<j \\ bp_G(l)=0}} Z_{PG'}(i,l) \cdot Z_{BE}(b_{(i,l)}, b'_{(i,l)}) \cdot B(P_b) \cdot Z_{WI}(l+1, bp_G(b'_{(i,l)}) - 1) \\ (2) \quad Z_{PG'}(i,j) \end{cases} \tag{ii}$$

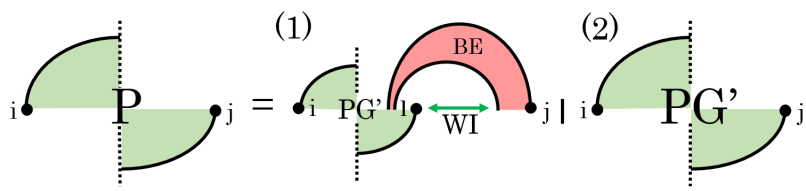

Fig. 3. Cases of Z_P. (1) j is paired in G and there must be some base, l, between $bp(j).j$ that is paired in G'. (2) j is not paired in G, then move directly to $Z_{PG'}$. Filled in circles show regions covered by specific structure classes, red for Z_{BE}, and green for Z_P and $Z_{PG'}$.

3.3 Pseudoknotted structures with rightmost band in G'

When i and j form a pseudoloop where the rightmost band of the pseudoloop is not in G, i.e., $bp_G(j) = 0$, it must be part of the G' structure.

To avoid allowing multiple adjacent weakly closed subregions in the pseudoloop, we introduce $Z_{PG'^w}(i,j)$, the partition function over all pseudoknotted density-2 structures between i and j where the rightmost band of the pseudoloop is part of the structure of G' and the structure is weakly closed.

Therefore, Cases $(1-2)$ of $Z_{PG'}$ (cf. Fig. 4) are used to account for the energy of the region spanned by the rightmost two bands using Z_{BE} and $Z_{VP}(l,j)$; and recursively continuing again to $Z_{PG'}$, or to $Z_{PG'^w}$ if there is a weakly closed region between the bands. For band border details see [13]. $Z_{PG'}$ Cases $(3-4)$ are end cases, where only one or two bands, respectively, need to be accounted for. For the base case where $i \geq j$, $Z_{PG'} = Z_{PG'^w} = 0$.

Time complexity for both $Z_{PG'}$ and $Z_{PG'^W}$ is $O(n^3)$, as in both cases we search over all values of l for a given region $[i,j]$.

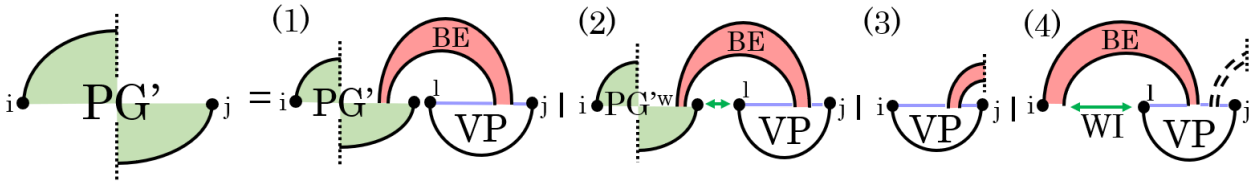

Fig. 4. Cases of $Z_{PG'}$. (1) handles two rightmost elements of the chain and continues. (2) is similar to case (1) except there is a weakly closed region between the bands, this will be handled by $Z_{PG'w}$ structure class to preserve the cubic time complexity. For the end cases we have (3) leftmost band of chain in G'; and (4) leftmost band in G. 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}.

3.4 $i.j$ in G' crosses base pair in 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. The energy of $R_{i,j}$ is the energy of all loops within $R_{i,j}$ that are not inside a band whose base pairs are in G and which crosses $i.j$. If $i \geq j$, i or j is paired in G', or $i.j$ does not cross any base pair of G, then $Z_{VP}(i,j) = 0$, otherwise $Z_{VP}(i,j)$ is computed as follows (cf. Fig. 5):

$Z_{VP}(i,j)$ Cases $(1-3)$ handle structures where there are no other base pairs in $[i,j]$ that cross the same band(s) $i.j$ crosses.[5] These cases are unambiguous: either i is covered, j is covered, or both. In Case (4), $(i+1).(j-1)$ forms a stacked pair (with its energy computed through e_{stP}). In Case (5), $i.j$ and $r.r'$ close an internal loop (with its energy computed through e_{intP}). In Cases $(6-9)$, we handle if there is a multiloop spanning a band. In Case (6), $r.(j-1)$ crosses a base pair in G and $[i+1, r-1]$ is a weakly closed non-empty region. In Case (7), $(i+1).r$ crosses a base pair in G and $[r+1, j-1]$ is a weakly closed non-empty region. In Case (8), $[i+1, r-1]$ is a weakly closed non-empty region, $r.bp(r)$ crosses a base pair in G, and $[bp(r)+1, j-1]$ is either empty or non-empty and weakly closed. Therefore, we introduce Z_{VP^R} (cf. Fig. 5), the partition

[5] Recurrences moved to Appendix for space, for band border lemmas see [13].

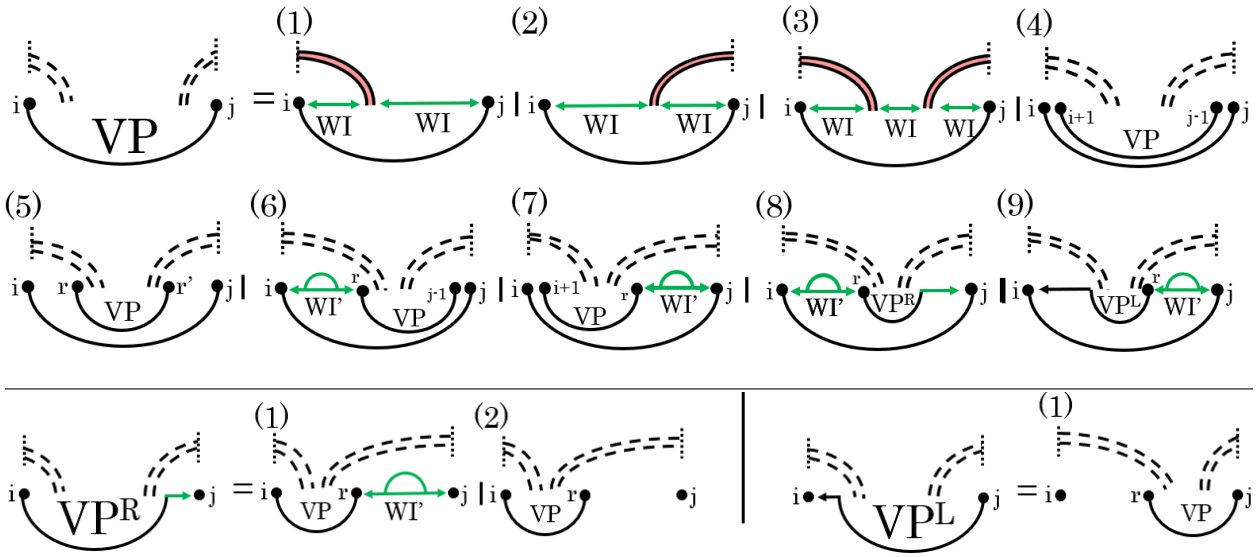

Fig. 5. Cases of VP, VP^R, and VP^L. Top: VP $(1-3)$ either two or three WI subregions (green) between i and j, band regions excluded. $(4-5)$, stacked pair and internal loop, respectively. $(6-9)$, $i.j$ closes a multiloop spanning a band. Bottom-left: VP^R, i.e., $i.bp(i)$ in G' crosses base pair in 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 $[i, r-1]$. 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.

function over all structures such that $i.bp(i) \in G'$ crosses base pair in G, and $bp(i) \neq j$ (distinct from Case (6)). Finally, in Case (9), $[r+1, j-1]$ is a weakly closed non-empty region, $r.bp(r)$ crosses a base pair in G, and $[i+1, bp(r)-1]$ is empty. Therefore, we introduce Z_{VP^L}, the partition function over all structures such that $bp(j).j \in G'$ crosses base pair in G, $bp(j) \neq i$ (distinct from Case (7)), and $[i+1, bp(j)-1]$ is empty.

Since G is a pseudoknot-free input structure, all base pair information in G, including borders of bands in G, can be computed in linear time. The time complexity of Z_{VP} is then dominated by the search over region $[i.j]$ to find value of r, and hence, is $O(n^3)$. For a high-level understanding: Cases $(6-9)$ of Z_{VP} are of critical importance to unambiguously handle a multiloop that spans a band or band(s). For a complete decomposition that preserves the $O(n^3)$ time complexity, Z_{VP^R} and Z_{VP^L} 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_{VP^R} 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_{VP^L} there is no possible case to introduce an additional structure adjacent to i. Thus, we avoid any ambiguity in Z_{VP} decomposition.

3.5 *i.j* closes a multiloop

$Z_{VM}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed and $i.j$ closes a multiloop (cf. Fig. 6). Otherwise, $Z_{VM}(i,j) = 0$

In Case (1) we have at least two branches, with a terminal pseudoknot-free stem. Z_{WM} is the partition function over all structures such that i and j are on a multiloop; and Z_{WM^1}, the partition function over all structures such that $i.bp(i)$ is a terminal stem on a multiloop and is pseudoknot-free. In Case (2) we have at least three branches, with a terminal pseudoknotted stem. Z_{WM^P} is the partition function over all structures such that $i.bp(i)$ is a terminal stem on a multiloop and forms a pseudoloop. Finally, in Case (3) we have exactly two branches, which form a pseudoloop.

i and j on a multiloop. $Z_{WM}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed, not empty, and i and j are on a multiloop (cf. Fig. 6). For base case where $i \geq j$, $Z_{WM}(i,j) = 0$. In Case (1), we have a pseudoknot-free initial stem, in Case (2), the initial stem is a pseudoloop. In Case (3), we have an intermediate pseudoknot-free branch that is not the initial stem. In Case (4), we have an intermediate pseudoloop that is not the the initial stem. In Case (5), j is unpaired.

4 Implementation

With CParty we compute the first CPF for density-2 structures, therefore, no other algorithm is available for a direct comparison to explicitly verify relative performance. To validate CParty, we first computed

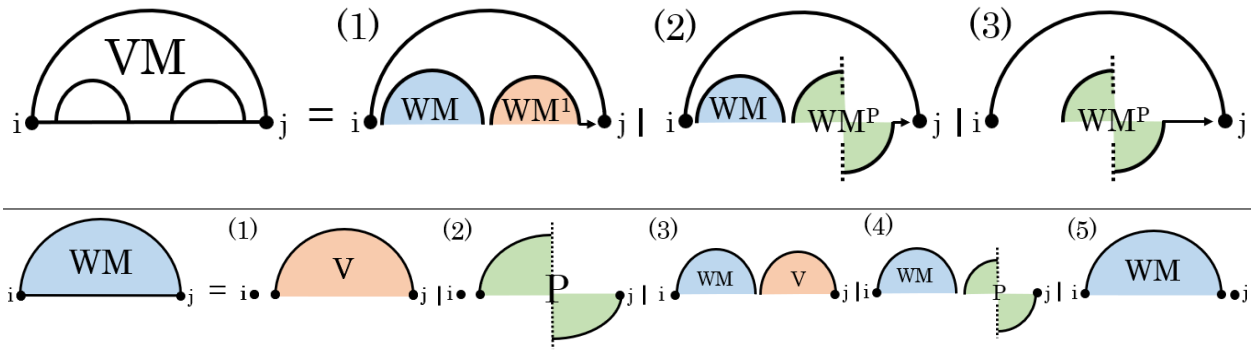

Fig. 6. Cases of VM and WM. Top: VM (1) terminal pseudoknot-free stem, (2) terminal pseudoknotted stem, with additional branches. (3) terminal pseudoknotted stem, no additional branches. Bottom: WM (1) initial stem pseudoknot-free, (2) initial stem pseudoknotted. (3) intermediate stem pseudoknot-free, additional branches. (4) intermediate stem pseudoknotted, with additional branches. (5), j is unpaired. Filled in circles show regions covered by specific structure classes, blue for Z_{WM}.

the partition function Z_{pkfree} for different test sequences giving no input structure as constraint. Note that, without an input G, the CParty partition function result will be pseudoknot-free. We then compared CParty Z_{pkfree} to the pseudoknot-free partition function result of RNAfold [15], Z'. For accurate comparison between Z_{pkfree} and Z', we used updated parameters [2,25] for RNAfold, with a dangle-free energy model (i.e., d_0). We found general agreement between pseudoknot-free CParty Z_{pkfree} (with empty input structure G) and RNAFold Z'. For further assessment of the pseudoknotted implementation of CParty, we compute the Boltzmann weight of the MFE structure obtained via HFold with a dangle-free energy model and an input structure G, and divide by CParty Z_{CPF} for the same sequence and structure pair. The resulting fraction is the proportion of the conditional ensemble attributed to the MFE structure, referred to as MFE^Z. We find MFE^Z is in the expected range, and explicitly confirm the completeness and correctness of Z_{CPF} for shorter sequences.

5 Discussion

Efficiently computing the partition function for RNA pseudoknot-free secondary structure is the problem solved in 1990 by the McCaskill algorithm and subsequent variants. McCaskill described a pseudoknot-free $O(n^4)$ DP algorithm to compute the total sum of energy in the system (partition function), further reduced to $O(n^3)$ using a simplified energy model [22]. Recently, a linear time approximation algorithm was proposed for pseudoknot-free partition function [38]. Computing partition function for pseudoknotted RNA secondary structure, is however, NP-hard. Existing pseudoknotted partition function methods require certain assumptions that limit the complexity of predicted pseudoknots [10]. These methods are too slow for long RNA sequences (e.g., NUPACK [37]). New approaches need to be adopted to enable more efficient partition function computation for RNA pseudoknots.

We presented CParty, a novel biologically motivated algorithm following the hierarchical folding hypothesis that efficiently computes the conditional partition function (CPF) for density-2 RNA secondary structures via an $O(n^3)$ DP scheme. CParty receives an RNA sequence and a pseudoknot-free structure, G, as input and calculates the CPF for $G \cup G'$, when G' is a pseudoknot-free structure disjoint from G, and $G \cup G'$ is a density-2 structure. Our algorithm removes the ambiguity of the dynamic programming algorithm HFold, enabling an efficient CPF approach that can sample from the conditional distribution, via stochastic analysis of the hierarchical distribution $\mathbb{P}(G) \times \mathbb{P}(G' \mid G)$.

Following the hierarchical folding hypothesis, by restricting the search space using the input structure G, we achieve a runtime complexity of $O(n^3)$ and space complexity of $O(n^2)$ matching the time and space complexity of pseudoknot-free partition function computation. Therefore, CParty represents a significant improvement in time and space complexity compared to the existing partition function algorithm for simple pseudoknots NUPACK [10,37], while also handling a broader class of structures including kissing hairpins and arbitrary depth of nested substructures. Computing the CPF can be a smart alternative to an unconditional pseudoknot partition function, because it is very fast and can still be informative in many scenarios.

We implemented CParty for stochastic analysis of pseudoknotted RNAs, enabling novel evaluation of hierarchical kinetics. An important factor in using our algorithm, is to correctly identify the input structure G. As previously identified [12,32], utilizing the most stable stems (lowest free energy or equivalently the highest probability stems) seems promising in identifying the MFE structure as well as low energy sub-optimal structures – structures that are energetically close to the MFE structure. Repeatedly sampling the hierarchical distribution with multiple choices of fixed structure G will be highly informative in identification of possible folding paths to the final structure. By incorporating structure constraints into sampling efforts, a CPF provides more accurate predictions of RNA secondary structures in specific biological contexts. With CParty we can obtain probabilistic structural information to assist identification of novel target sites and contribute to therapeutics.

Next, we describe how CParty enables probabilistic investigation to advance understanding of RNA structure formation. By applying CParty to the SARS-CoV-2 frameshift sequence we demonstrate initial folding of RNA significantly effects pseudoknot pairing propensity of the conditional ensemble. Uncovering hierarchical folding transitions between secondary structure motifs motivates future pan-coronaviral therapeutic RNA structure-function research. CParty can be similarly used in other applications to better harness available structure information, isolating key kinetic paths from initial non-crossing structure to pseudoknotted structure, with implications for subsequent tertiary structure modeling.

5.1 CParty for analyzing pseudoknot motifs in SARS-CoV-2

There has been widespread research predicting RNA structures for the SARS-CoV-2 frameshift sequence, including multiple viral genome reactivity probing experiments [11,18,36,39]. Employing CParty with different fixed input structures, G, we provide a stochastic view of suboptimal folding for the SARS-CoV-2 frameshift stimulating RNA structure ensemble. The frameshift stimulating pseudoknot in SARS-CoV-2 [14] possesses a pseudoknot structure which is density-2, making it an excellent use case for CParty.

Based on available SHAPE datasets and thermodynamic-based predictions for the frameshift sequence of SARS-CoV-2 [30,32], we identify the two most energetically favourable initial stems for the SARS-CoV-2 77 nucleotide frameshift pseudoknot sequence (cf. Table 2). With each of these initial stems as structure constraint, we compute Z_{CPF} via CParty. In addition, we calculate MFE^Z (cf. Section 4) for the sequence with each initial stem, respectively. With initial stem 2 as constraint, via CParty the conditional ensemble is more stable (larger Z_{CPF}) and diverse (smaller MFE^Z) than the conditional ensemble constrained by initial stem 1. For comparison, we computed the pseudoknot-free conditional partition function via RNAfold [15], referred to as Z', finding CParty Z_{CPF} is significantly larger than Z' due to the energy contribution of density-2 pseudoloops in the ensemble. Investigating the conditional ensemble constrained by initial stem 1, the ratio of RNAfold Z' to CParty Z_{CPF} is an order of magnitude smaller than the ratio of the conditional ensemble energies constrained by initial stem 2. We conclude that initial stem 2 folding leads to the most stable structures and significantly increased pseudoknot pairing propensity of the SARS-CoV-2 frameshift pseudoknot sequence.

Table 2. SARS-CoV-2 frameshift pseudoknot conditional ensemble analysis with most stable initial stems. First column: ID based on free energy rank. Second column: partial structure input for columns $3-6$, "()" indicate paired bases, "." show unpaired bases. Third column: free energy of structure in column two. Fourth column: MFE^Z, cf. Section 4. Bottom row: partial input sequence, accession: NC_045512.2, full sequence index $13466-13542$ [4].

ID	*Partial Stem Constraint*	*Kcal/mol*	MFE^Z	Z_{CPF}	Z'	$\frac{Z'}{Z_{CPF}}$
1	.....(((((((((((.....)))))))))))	-11.1	69%	$1.1*10^{14}$	$1.3*10^{12}$	$1.2*10^{-2}$
2	.(((((((.....))))))).....	-6.1	53%	$3.2*10^{14}$	$4.7*10^{11}$	$1.5*10^{-3}$
	AACGGGUUUGCGGUGUAAGUGCAGCCCGUCUUACACCGUG					

We investigate SARS CoV-2 frameshift conditional structure ensembles further by computing Z_{CPF} for decreasing sequence lengths with initial stems 1 and 2 as constraint, in order to simulate the effects of the translocating ribosome [3,9] (cf. Fig. 7). We observed the hairpin formed by initial stems 1 and 2 folding (referred to as 3_3 structure motif cf. [29,34], Fig. 7-b top arc diagram) dominates the conditional ensemble until a transition occurs (cf. Fig. 7, red rectangles) to the native-type structure [14] (3_6 motif [29], Fig. 7-b bottom arc diagram), which dominates the shorter length ensemble. Based on the stable transition of ensemble energies from 3_3 to 3_6 motif we hypothesize that destabilization of initial stem 2 *facilitates* subsequent refolding of the native-type pseudoknot.

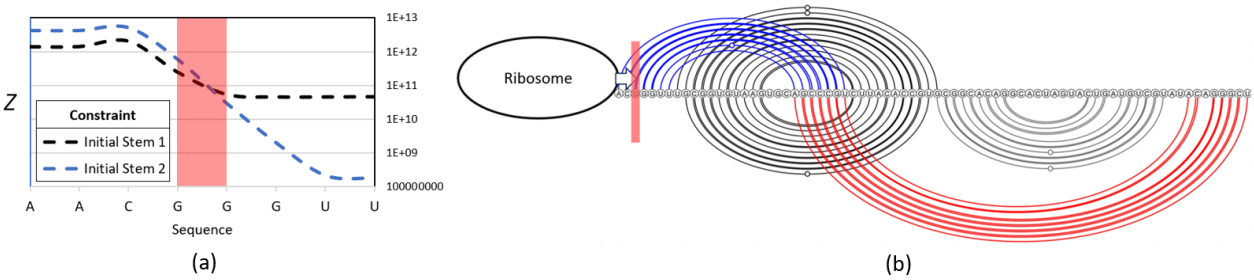

Fig. 7. SARS-CoV-2 secondary structure motif transition. (a): Z_{CPF} for decreasing SARS-CoV-2 sequence length, Z_{CPF} is divided by each sequence length for normalization. (b): Top arc diagram: 3_3 motif [29], initial stem 1 in black, initial stem 2 in blue. Bottom arc diagram: 3_6 motif. (a & b): Red rectangles highlight the location of a transition from the 3_3 motif to the 3_6 motif.

References

1. Akutsu, T.: Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. Discrete Appl. Math. **104**(1-3), 45–62 (2000)
2. Andronescu, M.S., Pop, C., Condon, A.E.: Improved free energy parameters for RNA pseudoknotted secondary structure prediction. RNA **16**(1), 26–42 (2010)
3. Atkins, J.F., Loughran, G., Bhatt, P.R., Firth, A.E., Baranov, P.V.: Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. Nucleic Acids Res. **44**(15), 7007–7078 (2016)
4. Brister, J.R., Ako-Adjei, D., Bao, Y., Blinkova, O.: NCBI viral genomes resource. Nucleic Acids Res. **43**(D1), D571–D577 (2015)
5. Chen, G., Chang, K.Y., Chou, M.Y., Bustamante, C., Tinoco, I.: Triplex structures in an RNA pseudoknot enhance mechanical stability and increase efficiency of–1 ribosomal frameshifting. PNAS **106**(31), 12706–12711 (2009)
6. Chen, H.L., Condon, A., Jabbari, H.: An o (n^5) algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids. J. Comput. Biol. **16**(6), 803–815 (2009)
7. Cho, S.S., Pincus, D.L., Thirumalai, D.: Assembly mechanisms of RNA pseudoknots are determined by the stabilities of constituent secondary structures. PNAS **106**(41), 17349–17354 (2009)
8. Ding, Y., Lawrence, C.E.: A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. **31**, 7280–7301 (Dec 2003)
9. Dinman, J.D.: Mechanisms and implications of programmed translational frameshifting. Wiley Interdiscip. Rev. RNA **3**(5), 661–673 (2012)
10. Dirks, R.M., Pierce, N.A.: A partition function algorithm for nucleic acid secondary structure including pseudoknots. J. Comput. Chem. **24**(13), 1664–1677 (2003)
11. Huston, N.C., Wan, H., Strine, M.S., Tavares, R.d.C.A., Wilen, C.B., Pyle, A.M.: Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. Mol. Cell **81**(3), 584–598 (2021)
12. Jabbari, H., Condon, A.: A fast and robust iterative algorithm for prediction of RNA pseudoknotted secondary structures. BMC Bioinf. **15**(1), 147 (2014)
13. Jabbari, H., Condon, A., Pop, A., Pop, C., Zhao, Y.: HFold: RNA pseudoknotted secondary structure prediction using hierarchical folding. In: International Workshop on Algorithms in Bioinformatics. pp. 323–334. Springer Berlin Heidelberg, Philadelphia, PA, USA (2007)
14. Kelly, J.A., Olson, A.N., Neupane, K., Munshi, S., San Emeterio, J., Pollack, L., Woodside, M.T., Dinman, J.D.: 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)
15. Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L.: Viennarna package 2.0. Algorithms Mol. Biol **6**, 1–14 (2011)
16. Lorenz, R., Hofacker, I.L., Stadler, P.F.: RNA folding with hard and soft constraints. Algorithms Mol. Biol. **11**(1), 1–13 (Dec 2016)
17. Lyngsø, R.B., Pedersen, C.N.: Pseudoknots in RNA secondary structures. In: Proceedings of the fourth annual international conference on Computational molecular biology. pp. 201–209. Association for Computing Machinery, New York, NY, United States (2000)
18. Manfredonia, I., Nithin, C., Ponce-Salvatierra, A., Ghosh, P., Wirecki, T.K., Marinus, T., Ogando, N.S., Snijder, E.J., van Hemert, M.J., Bujnicki, J.M., et al.: Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. Nucleic Acids Res. **48**(22), 12436–12452 (2020)
19. Mathews, D.H.: Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA **10**(8), 1178–1190 (2004)
20. Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H.: Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288**(5), 911–940 (1999)
21. Mathews, D.H., Turner, D.H.: Prediction of RNA secondary structure by free energy minimization. Curr. Opin. Struct. Biol. **16**(3), 270–278 (2006)
22. McCaskill, J.S.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers: Original Research on Biomolecules **29**(6-7), 1105–1119 (1990)
23. Ponty, Y., Saule, C.: A combinatorial framework for designing (pseudoknotted) RNA algorithms. In: T. Przytycka, M.F.S. (ed.) Algorithms in Bioinformatics. pp. 250–269. No. 6833 in LNBI, Springer, Saarbrücken, Germany (Jan 2011)
24. Rastegari, B., Condon, A.: Parsing nucleic acid pseudoknotted secondary structure: algorithm and applications. J. Comput. Biol. **14**(1), 16–32 (2007)
25. Ren, J., Rastegari, B., Condon, A., Hoos, H.H.: Hotknots: heuristic prediction of RNA secondary structures including pseudoknots. RNA **11**(10), 1494–1504 (2005)
26. Rivas, E., Eddy, S.R.: A dynamic programming algorithm for RNA structure prediction including pseudoknots. J. Mol. Biol. **285**(5), 2053–2068 (1999)
27. Roman, C., Lewicka, A., Koirala, D., Li, N.S., Piccirilli, J.A.: The SARS-CoV-2 programmed- 1 ribosomal frameshifting element crystal structure solved to 2.09 å using chaperone-assisted RNA crystallography. ACS Chem. Biol. **16**(8), 1469–1481 (2021)

28. Sato, K., Kato, Y., Hamada, M., Akutsu, T., Asai, K.: IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinf. **27**(13), i85–i93 (2011)
29. Schlick, T., Zhu, Q., Dey, A., Jain, S., Yan, S., Laederach, A.: To knot or not to knot: multiple conformations of the SARS-CoV-2 frameshifting RNA element. J. Am. Chem. Soc. **143**(30), 11404–11422 (2021)
30. Schlick, T., Zhu, Q., Jain, S., Yan, S.: Structure-altering mutations of the SARS-CoV-2 frameshifting RNA element. Biophys. J. **120**(6), 1040–1053 (2021)
31. Tinoco Jr, I., Bustamante, C.: How RNA folds. J. Mol. Biol. **293**(2), 271–281 (1999)
32. Trinity, L., Wark, I., Lansing, L., Jabbari, H., Stege, U.: Shapify: Paths to SARS-CoV-2 frameshifting pseudoknot. PLoS Comput. Biol. **19**(2), e1010922 (2023)
33. Witwer, C., Hofacker, I.L., Stadler, P.F.: Prediction of consensus RNA secondary structures including pseudoknots. IEEE/ACM Trans. Comput. Biol. Bioinf. **1**(2), 66–77 (2004)
34. Yan, S., Zhu, Q., Hohl, J., Dong, A., Schlick, T.: Evolution of coronavirus frameshifting elements: Competing stem networks explain conservation and variability. PNAS **120**(20), e2221324120 (2023)
35. Yan, S., Zhu, Q., Jain, S., Schlick, T.: Length-dependent motions of SARS-CoV-2 frameshifting RNA pseudoknot and alternative conformations suggest avenues for frameshifting suppression. Nat. Commun. **13**(1), 4284 (2022)
36. Yang, S.L., DeFalco, L., Anderson, D.E., Zhang, Y., Aw, J.G.A., Lim, S.Y., Lim, X.N., Tan, K.Y., Zhang, T., Chawla, T., et al.: Comprehensive mapping of SARS-CoV-2 interactions in vivo reveals functional virus-host interactions. Nat. Commun. **12**(1), 1–15 (2021)
37. Zadeh, J.N., Steenberg, C.D., Bois, J.S., Wolfe, B.R., Pierce, M.B., Khan, A.R., Dirks, R.M., Pierce, N.A.: NUPACK: analysis and design of nucleic acid systems. J. Comput. Chem. **32**(1), 170–173 (2011)
38. Zhang, H., Zhang, L., Mathews, D.H., Huang, L.: LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. Bioinf. **36**(Supplement_1), i258–i267 (2020)
39. Zhang, K., Zheludev, I.N., Hagey, R.J., Haslecker, R., Hou, Y.J., Kretsch, R., Pintilie, G.D., Rangan, R., Kladwang, W., Li, S., 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)
40. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. **9**(1), 133–148 (1981)

A Appendix

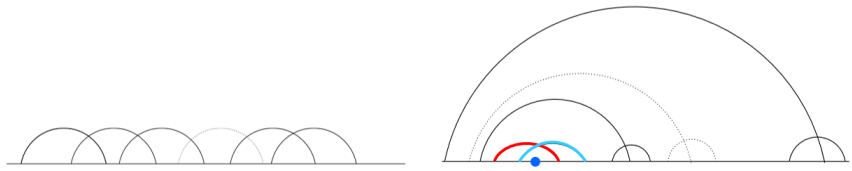

Fig. 8. Arc diagram representation of two density-2 structures, each structure contains an arbitrary number and depth of bands [13]. Blue dot covered by red and light blue bands indicates how association with closed region maintains the density-2 property. Figure modified from [13].

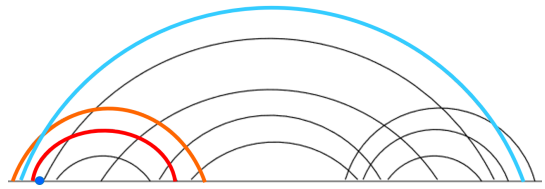

Fig. 9. Example of a *bi-secondary structure* that is not a density-2 structure [13]. Blue dot covered by red, orange, and light blue bands indicate density-2 property not satisfied. Figure modified from [13].

Table 3. ENERGY PARAMETERS. All parameters were derived at 37 degrees celsius and 1 M salt (NaCl) concentration or extrapolated from experimental values cf. [13,2,25].

Name	*Description*	*Value (kcal/mol)*
P_s	Exterior pseudoloop initiation penalty	-1.38
P_{sm}	Penalty for introducing pseudoknot inside a multiloop	10.07
P_{sp}	Penalty for introduce pseudoknot inside a pseudoloop	15.00
P_b	Band penalty	2.46
P_{up}	Penalty for unpaired base in a pseudoloop	0.06
P_{ps}	Penalty for closed subregion inside a pseudoloop	0.96
$e_H(i,j)$	Energy of a hairpin loop closed by $i.j$	
$e_S(i,i+1,j-1,j)$	Energy of a stacked pair closed by $i.j$	
$e_{stP}(i,i+1,j-1,j)$	Energy of a stacked pair that spans a band	$0.89 \times e_S(i,j)$
$e_{int}(i,r,r',j)$	Energy of a pseudoknot-free internal loop	
$e_{intP}(i,r,r',j)$	Energy of an internal loop that spans a band	$0.74 \times e_{int}(i,d,e,j)$
a	Multiloop initiation penalty	3.39
b	Multiloop base pair penalty	0.03
c	Penalty for unpaired base in a multiloop	0.02
a'	Penalty for introducing a multiloop that spans a band	3.41
b'	Base pair penalty for a multiloop that spans a band	0.56
c'	Penalty for unpaired base in a multiloop that spans a band	0.12

A.1 Pseudoknotted structures with rightmost band in G'

When i and j form a pseudoloop where the rightmost band of the pseudoloop is not in G, i.e., $bp_G(j) = 0$, it must be part of the G' structure.

To avoid allowing multiple adjacent weakly closed subregions in the pseudoloop, we introduce $Z_{PG'^w}(i,j)$, the partition function over all pseudoknotted density-2 structures between i and j where the rightmost band of the pseudoloop is part of the structure of G' and the structure is weakly closed.

Therefore, Cases $(1-2)$ of $Z_{PG'}$ (cf. Fig. 4) are used to account for the energy of the region spanned by the rightmost two bands using Z_{BE} and $Z_{VP}(l,j)$; and recursively continuing again to $Z_{PG'}$, or to $Z_{PG'^w}$ if there is a weakly closed region between the bands. For band border details see [13]. $Z_{PG'}$ Cases $(3-4)$ are end cases, where only one or two bands, respectively, need to be accounted for. For the base case when $i \geq j$, $Z_{PG'} = Z_{PG'^w} = 0$.

$$Z_{PG'}(i,j) = \sum \begin{cases} (1) \text{ if } bp_G(j) = 0 \\ \displaystyle\sum_{\substack{i<l<b_{(i,l)} \\ isCovered(G_{i,j},l)}} Z_{PG'}(i,l-1) \cdot Z_{BE}(b_{(i,l)}, b'_{(i,l)}) \cdot Z_{VP}(l,j) \cdot B(2P_b) \\ (2) \text{ if } bp_G(j) = 0, \text{ and } bp_G(l-1) < (l-1) \\ \displaystyle\sum_{\substack{i<l<b_{(i,l)} \\ isCovered(G_{i,j},l)}} Z_{PG'^w}(i,l-1) \cdot Z_{BE}(b_{(i,l)}, b'_{(i,l)}) \cdot Z_{VP}(l,j) \cdot B(2P_b) \\ (3)\ Z_{VP}(i,j) \cdot B(P_b) \\ (4) \text{ if } 0 = bp_G(j) < bp_G(i) \\ \displaystyle\sum_{i<l<bp_G(i)} Z_{BE}(b_{(i,l)}, b'_{(i,l)}) \cdot B(2P_b) \cdot Z_{VP}(l,j) \cdot Z_{WI}(b'_{(i,l)}+1, l-1) \end{cases} \tag{iii}$$

$$Z_{PG'^w}(i,j) = \sum_{\substack{i<l<j \\ cover(l)=cover(j)}} Z_{PG'}(i,l) \cdot Z_{WI}(l+1,j) \tag{iv}$$

A.2 $i.j$ in G' crosses base pair in 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. The energy of $R_{i,j}$ is the energy of all loops within $R_{i,j}$ that are not inside a band whose base pairs are in G and which crosses $i.j$. If $i \geq j$, i or j is paired in G', or $i.j$ does not cross any base pair of G, then $Z_{VP}(i,j) = 0$, otherwise $Z_{VP}(i,j)$ is computed as follows:

$Z_{VP}(i,j)$ Cases $(1-3)$ handle structures where there are no other base pairs in $R_{i,j}$ that cross the band(s) $i.j$ crosses. These cases are unambiguous: either i is covered, j is covered, or both. In Case (4), $(i+1).(j-1)$ forms a stacked pair (with its energy computed through e_{stP}). In Case (5), $i.j$ and $r.r'$ close an internal loop (with its energy computed through e_{intP}). In Case (6), $r.(j-1)$ crosses a base pair in G and $[i+1, r-1]$ is a weakly closed non-empty region (multiloop spanning band initiation, base pair, and unpaired base penalties, a', b', and c', respectively). In Case (7), $(i+1).r$ crosses a base pair in G and $[r+1, j-1]$ is a weakly closed non-empty region. In Case (8), $[i+1, r-1]$ is a weakly closed non-empty region, $r.bp(r)$ crosses a base pair in G, and $[bp(r)+1, j-1]$ is either empty or non-empty and weakly closed. Therefore, we introduce Z_{VP^R} (cf. Fig. 5), the partition function over all structures such that $i.bp(i) \in G'$ crosses base pair in G, and $bp(i) \neq j$ (distinct from Case (6)). Finally, in Case (9), $[r+1, j-1]$ is a weakly closed non-empty region, $r.bp(r)$ crosses a base pair in G, and $[i+1, bp(r)-1]$ is empty. Therefore, we introduce Z_{VP^L}, the partition function over all structures such that $bp(j).j \in G'$ crosses base pair in G, $bp(j) \neq i$ (distinct from Case (7)), and $[i+1, bp(j)-1]$ is empty.

$$Z_{VP}(i,j) = \sum \begin{cases} (1) \text{ if } isCovered(G,i), \text{ and } \overline{isCovered(G,j)} \\ \quad Z_{WI}(i+1, B'_{(i,j)}-1) \cdot Z_{WI}(B_{(i,j)}+1, j-1) \\ (2) \text{ if } \overline{isCovered(G,i)}, \text{ and } isCovered(G,j) \\ \quad Z_{WI}(i+1, b_{(i,j)}-1) \cdot Z_{WI}(b'_{(i,j)}+1, j-1) \\ (3) \text{ if } isCovered(G,i), \text{ and } isCovered(G,j) \\ \quad Z_{WI}(i+1, (B'_{(i,j)}-1)) \cdot Z_{WI}(B_{(i,j)}+1, b_{(i,j)}-1) \cdot Z_{WI}(b'_{(i,j)}+1, j-1) \\ (4) \text{ if } (bp_G(i+1)=0, \text{ and } bp_G(j-1)=0) \\ \quad B(e_{stP}(i,i+1,j-1,j)) \cdot Z_{VP}(i+1,j-1) \\ (5) \text{ if } cover(G,i) = cover(G,r) \text{ and } cover(G,j) = cover(G,r') \\ \quad \displaystyle\sum_{\substack{i<r<\min(B'_{(i,j)}, b_{(i,j)}) \\ \max(b'_{(i,j)}, B_{(i,j)})<r'<j}} B(e_{intP}(i,r,r',j)) \cdot Z_{VP}(r,r') \\ (6) \displaystyle\sum_{\substack{i<r<\min(B'_{(i,j)}, j) \\ bp_G(r)=0}} Z_{VP}(r,j-1) \cdot B(a'+2b') \cdot Z_{WI'}(i+1,r-1) \\ (7) \displaystyle\sum_{\substack{max(i, b'_{(i,j)})<r<j \\ bp_G(r)=0}} Z_{VP}(i+1,r) \cdot B(a'+2b') \cdot Z_{WI'}(r+1,j-1) \\ (8) \displaystyle\sum_{\substack{i<r<\min(B'_{(i,j)}, j) \\ bp_G(r)=0}} Z_{VP^R}(r,j-1) \cdot B(a'+2b') \cdot Z_{WI'}(i+1,r-1) \\ (9) \displaystyle\sum_{\substack{max(i, b'_{(i,j)})<r<j \\ bp_G(r)=0}} Z_{VP^L}(i+1,r) \cdot B(a'+2b') \cdot Z_{WI'}(r+1,j-1) \end{cases} \tag{v}$$

$i.bp(i)$ in G' crosses base pair in G, $bp(i) \neq j$

$$Z_{VP^R}(i,j) = \sum \begin{cases} (1) \displaystyle\sum_{max(i, b'_{(i,j)})<r<j} Z_{VP}(i,r) \cdot Z_{WI'}(r+1,j) \\ (2) \text{ if empty}(G, [(r+1), j]) \\ \quad \displaystyle\sum_{max(i, b'_{(i,j)})<r<j} Z_{VP}(i,r) \cdot B(c'(j-r)) \end{cases} \tag{vi}$$

$bp(j).j$ in G' crosses base pair in G, $bp(j) \neq i$

$$Z_{VP^L}(i,j) = \begin{array}{c} \text{if empty}(G, [i, (r-1)]) \\ \displaystyle\sum_{i<r<min(B'_{(i,j)}, j} B(c'(r-i)) \cdot Z_{VP}(r,j) \end{array} \tag{vii}$$

A.3 *i.j* closes a multiloop

$Z_{VM}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed and $i.j$ closes a multiloop. Otherwise, $Z_{VM}(i,j) = 0$

$$Z_{VM}(i,j) = \sum \begin{cases} (1) \sum\limits_{(i+1)<h\leq(j-1)} Z_{WM}(i+1,h-1)\cdot B(a+b)\cdot Z_{WM^1}(h,j-1) \\ (2) \sum\limits_{(i+1)<h\leq(j-1)} Z_{WM}(i+1,h-1)\cdot B(a+b)\cdot Z_{WM^P}(h,j-1) \\ (3) \sum\limits_{i<r<(j-1)} Z_{WM^P}(r,j-1)\cdot B(a+b)\cdot B(c(r-i-1)) \end{cases} \tag{viii}$$

A.4 i and j on a multiloop.

$Z_{WM}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed, not empty, and i and j are on a multiloop. For base case where $i \geq j$, $Z_{WM}(i,j) = 0$.

$$Z_{WM}(i,j) = \sum \begin{cases} (1) \sum\limits_{i\leq r<j} B(c(r-i))\cdot Z_V(r,j)\cdot B(b) \\ (2) \sum\limits_{i\leq r<j} B(c(r-i))\cdot Z_P(r,j)\cdot B(P_{sm}+b) \\ (3) \sum\limits_{i<r<(j-1)} Z_{WM}(i,r)\cdot Z_V(r+1,j)\cdot B(b) \\ (4) \sum\limits_{i<r<(j-1)} Z_{WM}(i,r)\cdot Z_P(r+1,j)\cdot B(P_{sm}+b) \\ (5) \quad Z_{WM}(i,j-1)\cdot B(c) \end{cases} \tag{ix}$$

i and j on a multiloop, terminal stem pseudoknot-free $Z_{WM^1}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed, not empty, and i and j are on a multiloop (terminal stem, cf. Fig. 10). With Case (1), $i.j$ form a pseudoknot-free loop, and Case (2), j is unpaired. For base case where $i \geq j$, $Z_{WM^1}(i,j) = 0$.

$$Z_{WM^1}(i,j) = \sum \begin{cases} (1) \quad Z_V(i,j)\cdot B(b) \\ (2) \quad Z_{WM^1}(i,j-1)\cdot B(c) \end{cases} \tag{x}$$

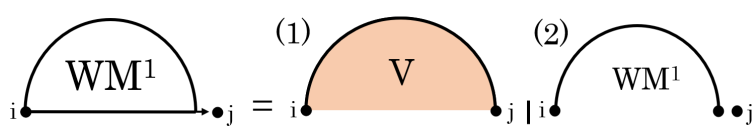

Fig. 10. Cases of WM^1. (1) terminal stem pseudoknot-free, (2) j is unpaired.

i and j on a multiloop, terminal stem pseudoknotted $Z_{WM^P}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed, not empty, and i and j are on a multiloop (pseudoknotted terminal branch, cf. Fig. 11). With Case (1), $i.j$ form a pseudoloop, and Case (2), j is unpaired. For base case where $i \geq j$, $Z_{WM^P}(i,j) = 0$.

$$Z_{WM^P}(i,j) = \sum \begin{cases} (1) \quad Z_P(i,j)\cdot B(P_{sm}+b) \\ (2) \quad Z_{WM^P}(i,j-1)\cdot B(c) \end{cases} \tag{xi}$$

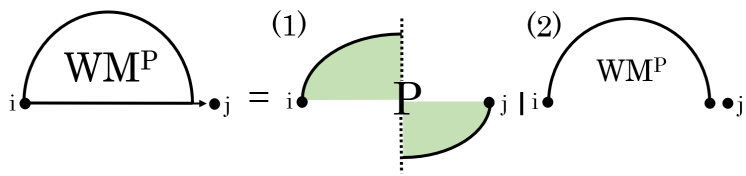

Fig. 11. Cases of WM^P. (1) terminal stem pseudoknotted, (2) j is unpaired.

A.5 Weakly closed subregions inside pseudoloop

$Z_{WI}(i,j)$ is the partition function over all structures $R_{i,j}$ given that $[i,j]$ is weakly closed ($cover(i) = cover(j) \neq 0$), and $R_{i,j}$ is inside a pseudoloop. If $i = j$ and $bp_G(i) = 0$, $[i,j]$ is empty and $Z_{WI}(i,j) = P_{up}$, i.e., penalty for unpaired base in a pseudoloop. $Z_{WI}(i,j) = 0$, if $i > j$. Otherwise, $Z_{WI}(i,j) = 0$ ($cover(i) \neq cover(j)$, subregion not weakly closed). Z_{WI} is similar to Z_W with additional penalties for base pair, unpaired bases, and pseudoknot initiation inside a pseudoloop.

$$Z_{WI}(i,j) = \sum \begin{cases} (1) \text{ if } r.j \in G, \text{or } (bp_G(r) = 0 \text{ and } bp_G(j) = 0) \\ \quad \sum_{i \leq r < j} Z_{WI}(i, r-1) \cdot Z_V(r,j) \cdot B(P_{ps}) \\ (2) \quad Z_{WI}(i, j-1) \cdot B(P_{up}) \\ (3) \sum_{i \leq r < j} Z_{WI}(i, r-1) \cdot Z_P(r,j) \cdot B(P_{sp} + P_{ps}) \end{cases} \tag{xii}$$

A.6 Non-empty weakly closed subregion inside band

$Z_{WI'}(i,j)$ is the partition function over all nonempty structures $R_{i,j}$, if $[i,j]$ is weakly closed with respect to G, given that $R_{i,j}$ is inside a band. Otherwise, $Z_{WI'}(i,j) = 0$. $Z_{WI'}$ is similar to Z_{WI} with additional penalties for base pair or unpaired base in multiloop spanning a band, and pseudoknot initiation inside a multiloop.

$$Z_{WI'}(i,j) = \sum \begin{cases} (1) \text{ if } r.j \in G, \text{or } (bp_G(r) = 0 \text{ and } bp_G(j) = 0) \\ \quad \sum_{i \leq r < j} Z_{WI'}(i, r-1) \cdot Z_V(r,j) \cdot B(b') \\ (2) \text{ if } bp_G(j) = 0 \\ \quad Z_{WI'}(i, j-1) \cdot B(c') \\ (3) \sum_{i \leq r < j} Z_{WI'}(i, r-1) \cdot Z_P(r,j) \cdot B(P_{sm} + b') \end{cases} \tag{xiii}$$

A.7 $i.j$ close a loop

$Z_V(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed or empty and $i.j$. Otherwise $Z_V(i,j) = 0$. This subfunction and $Z_{VBI}(i,j)$ to follow are unchanged from Mathews et al. [20] pseudoknot-free algorithm, i.e., penalties for hairpin loop, stacked base pairs, internal/bulge loop, or multiloop.

***i.j* close an internal/bulge loop** $Z_{VBI}(i,j)$ is the partition function over all structures $R_{i,j}$ for region $[i,j]$, if $[i,j]$ is weakly closed or empty and $i.j$ closes a bulge or internal loop. Otherwise $Z_{VBI}(i,j) = 0$ [20].

A.8 $[i, i'] \cup [bp(i'), bp(i)]$ band region

$Z_{BE}(i, i')$ is the partition function over the band $[i', i] \cup [bp_G(i), bp_G(i')]$, if $i \leq i' < bp_G(i') \leq bp_G(i)$; otherwise $Z_{BE}(i, i') = 0$. In Case (1), $bp(i+1).bp(i) - 1$ form a stacked pair in G. In Case (2), $i.bp(i)$ and $l.bp(l)$ in G close an internal loop. In Case (3), $[i+1, l-1]$ and $[bp(l)+1, bp(i)-1]$ are both weakly closed non-empty region. In Case (4), $[i+1, l-1]$ is a weakly closed non-empty region and $[bp(l)+1, bp(i)-1]$ is empty. In Case (5), $[bp(l)+1, bp(i)-1]$ is a weakly closed non-empty region and $[i+1, l-1]$ is empty (cf. Fig. 12). For base case, $Z_{BE}(i, i) = 0$ if $i < bp_G(i)$.

$$Z_{BE}(i,j) = \sum \begin{cases} (1) \text{ if } bp_G(i+1) = bp_G(i) - 1 \\ \quad B(e_{stP}(i, bp_G(i))) \cdot Z_{BE}(i+1, i') \\ (2) \text{ if } bp_G(l) > 0,\ \text{empty}(G, [i+1, l-1]),\ \text{empty}(G, [bp_G(l)+1, bp_G(i)-1]), \\ \quad\quad i < l \leq i',\ \text{and } (bp_G(i') \leq bp_G(l) < bp_G(i)) \\ \quad B(e_{intP}(i, l, bp_G(l), bp_G(i))) \cdot Z_{BE}(l, i') \\ (3) \text{ if } bp_G(l) > 0,\ \text{weakly closed}(G, [i+1, l-1]), \\ \quad\quad \text{weakly closed}(G, [bp_G(l)+1, bp_G(i)-1]), i < l \leq i',\ \text{and } bp_G(i') \leq bp_G(l) < bp_G(i) \\ \quad Z_{WI'}(i+1, l-1) \cdot Z_{BE}(l, i') \cdot Z_{WI'}(bp_G(l)+1, bp_G(i)-1) \cdot B(a' + 3b') \\ (4) \text{ if } bp_G(l) > 0, \text{weakly closed}(G, [i+1, l-1]),\ \text{and empty}(G, [bp_G(l)+1, bp_G(i)-1]), \\ \quad\quad i < l \leq i',\ \text{and } bp_G(i') \leq bp_G(l) < bp_G(i) \\ \quad Z_{WI'}(i+1, l-1) \cdot Z_{BE}(l, i') \cdot B(a' + 2b' + c'(bp_G(i) - bp_G(l) + 1)) \\ (5) \text{ if } bp_G(l) > 0,\ \text{empty}(G, [i+1, l-1]), \text{weakly closed}(G, [bp_G(l)+1, bp_G(i)-1]), \\ \quad\quad i < l \leq i',\ \text{and } bp_G(i') \leq bp_G(l) < bp_G(i) \\ \quad Z_{BE}(l, i') \cdot Z_{WI'}(bp_G(l)+1, bp_G(i)-1) \cdot B(a' + 2b' + c'(l - i - 1)) \end{cases} \tag{xiv}$$

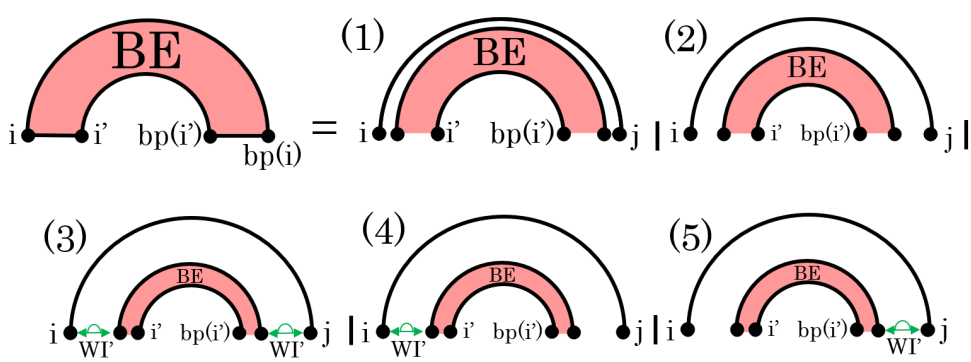

Fig. 12. Cases of BE. (1) stacked pair in G; (2) internal loop; (3) initial and terminal weakly closed non-empty regions. (4) initial weakly closed non-empty region and terminal loop. (5) initial loop and terminal weakly closed non-empty region.