APPENDIX

—

Structure and interaction prediction in prokaryotic RNA biology

Patrick R. Wright[1,*], Martin Mann[1,*] and Rolf Backofen[1,2,*]

[1]Bioinformatics Group, University of Freiburg, Freiburg, Germany
[2]Center for Biological Signaling Studies (BIOSS), University of Freiburg, Germany

[*]contributed equally

Contents

A Sparsification

Recapitulate a recursion to fill the Nussinov matrix $N_{i,j}$ shown in Eq. 1 and depicted in Fig. 1.

$$N_{i,j} = \max \begin{cases} 0 & : \text{if } (i + s_l) \geq j \\ N_{i+1,j-1} + 1 & : \text{if } R_i, R_j \text{ can form base pair} \\ \max_{i \leq k < j} \{N_{i,k} + N_{k+1,j}\} & : \text{decomposition} \end{cases} \quad (1)$$

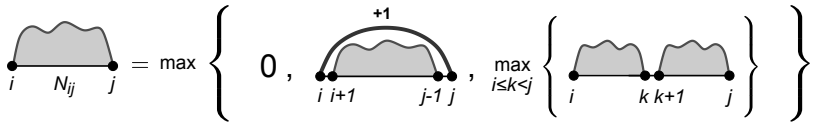

Figure 1: Graphical depiction of the Nussinov-like recursion from Eq. 1.

The time complexity of the Nussinov algorithm is $O(n^3)$. This is, however, still high for long RNA sequences as for instance large mRNAs, long non-coding RNAs and viral RNAs. For that reason, attempts have been made to reduce the overall time complexity on average to $O(n^2)$. When revisiting the recursion in Eq. 1, the split in the final decomposition case is the one causing the high complexity. Many of these splits will not lead to the optimal solution. This observation sparked the idea of sparsification techniques, first introduced by Ydo Wexler and colleagues [1], which is discussed in the following.

To this end, we first reformulate the Nussinov recursion by introducing two additional matrices $B_{i,j}$ and $D_{i,j}$, which handle the cases that $\mathcal{R}_i$ and $\mathcal{R}_j$ are paired (B; case 2 in Eq. 1), or the region is decomposed into two substructures (D; case 3 in Eq. 1). This gives rise to the modified recursion as depicted in Fig. 2a. For the decomposition case (i.e., the recursion for $D_{i,j}$), the exact copy of the Nussinov-like recursion from Eq. 1 would yield $D_{i,j} = \max_{q \in Q_{i,j}}(N_{i,q} + N_{q+1,j})$, where $Q_{i,j}$ contains all valid decomposition points q within the interval $i..j$. This decomposition is replaced by $D_{i,j} = \max_{q \in Q_{i,j}}(B_{i,q} + N_{q+1,j})$ to make the recursion non-ambiguous.

As already stated, the main complexity comes from the decomposition case, where $Q_{i,j}$ covers all interval indices, i.e. equals $\{i..j-1\}$ without sparsification. However, one can prove that many $q \in \{i..j-1\}$ are not required for the optimal solution. The idea is that if the best

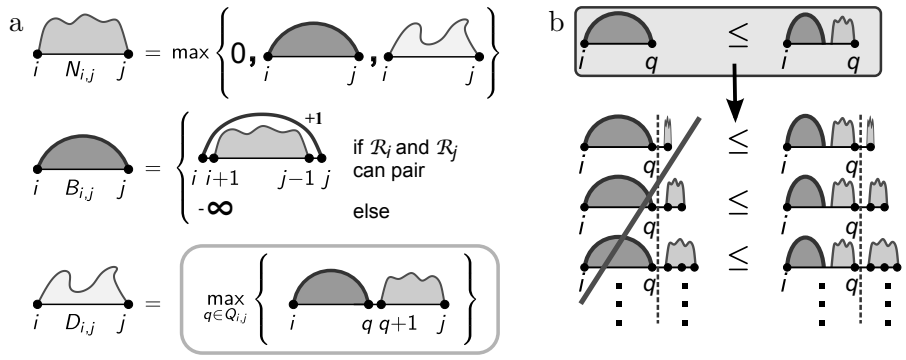

Figure 2: Sparsification of Nussinov's algorithm. a) Modified recursion scheme, where $N_{i,j}$ is calculated by two matrices $B_{i,j}$ and $D_{i,j}$, corresponding to the base pair and decomposition case, respectively. Sparsification takes place for the decomposition case. Without sparsification, $Q_{i,j}$ contains all elements of the interval $i..j$, which implies that the condition $q \in Q_{i,j}$ considers all possible decompositions. b) Analysis of the decomposition. If all conformations that contain the base pair (i, q) are not better than any conformation that decomposes the region $i..q$, then one can safely replace the sub conformation including the base pair (i, q) by a decomposition conformation when extending $i..q$ to the right. Thus, such positions q can be removed from Q for all regions $i..j$ with $q < j$, which reduces the runtime significantly.

conformation where i and q form a base pair is not better than a conformation stemming from a decomposition of $i..q$, then the base pair (i, q) is not required for the optimal conformation since it can always be replaced by this decomposition of $i..j$ without violating the condition of a nested structure (see Fig. 2b). This means that the position q is not required as a possible element of $Q_{i,j}$ for all $j > q$. Thus, $Q_{i,j}$ is set for each j to a current candidate list Q_i^{cand}, where an additional element q is only added to Q_i^{cand} if $\forall q' \in Q_{i,q} : B_{i,q} > B_{i,q'} + N_{q'+1,q}$. This gives rise to a time complexity of $O(n^2\psi(n))$, where $\psi(n)$ denotes the expected maximal size of a candidate list in a sequence of length n. As shown in [2], the candidate list size converges to a constant, which yields a quadratic time and space algorithm.

This approach has been extended in several ways. First, in [3] this approach was extended to the Sankoff approach [4], which is the co-folding of two homologous sequences that is discussed in the main text. In [5], it could be shown that sparsification of the Sankoff approach does not only reduce time, but also the space requirement. In [6] and [7], the idea of sparsification was extended to the problem of RNA-RNA interaction and RNA pseudoknot prediction, respectively. Furthermore, [3] introduced another technique to reduce the computational requirements for RNA structure prediction by using a variant of Vailant's method [8], who showed that parsing of a context-free grammar can be implemented by an optimized matrix multiplication.

Finally, the idea of sparsification for Sankoff-like approaches was extended in [9, 10] to a more *data-driven* approach. Here, one does not filter base pairs that provably do not lead to an optimal solution. Instead, one filters base pairs that have a low probability in the input sequences, which are thus not *likely* to yield an optimal solution in the co-folding of two sequences.

B Estimating p-values for interaction energies

Duplex energies of RNA-RNA interactions can not be directly used to make a combined prediction, because they are strongly influenced by the GC-content and dinucleotide frequency of the organism they are made for. Hence, the duplex energies need to be transformed to p-values, which are then comparable. In the following, we will introduce how p-values can be derived and how p-values from different organisms can be combined to enable comparative RNA-RNA

interaction prediction.

A p-value represents a statistical measure for the quality of a given prediction and, if correctly estimated, also enables comparability. Following the conclusions from extreme value theory [11], it is appropriate to regard the results of RNA-RNA interaction predictions as extreme value distributed (see Fig. 3).

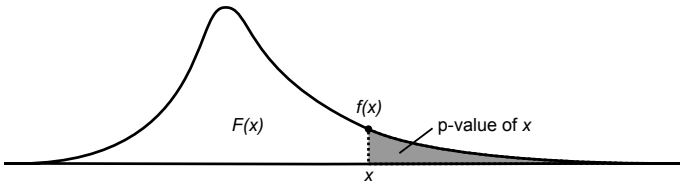

Figure 3: Schematic representation of a p-value for a given density function f of the background model (generalized extreme value distribution).

The density function f of the generalized extreme value (GEV) distribution is given in Eq. 2. The variable parameters are location (μ), scale (σ) and shape (ε). The location defines the center of the distribution while its width is governed by the scale parameter. The shape defines the character of the distribution's tails, e.g. a higher ε corresponds to a longer right tail.

$$f(x;\ \mu,\sigma,\varepsilon) = \frac{1}{\sigma}\left[1+\varepsilon\left(\frac{x-\mu}{\sigma}\right)\right]^{(-1/\varepsilon)-1} \exp\left(-\left[1+\varepsilon\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\varepsilon}\right) \tag{2}$$

A p-value is the probability that a certain event x or something more extreme ($\geq x$) is observed for a specific background model. Given the density function f of the events, a p-value can be computed by the integral $\int_x^\infty f(x)\mathrm{d}x$ (see Fig. 3). The cumulative distribution F for the GEV distribution (see Eq. 3) provides the integral $F(x) = \int_{-\infty}^x f(x)\mathrm{d}x$ for events $\leq x$, such that we can compute the p-value by $1 - F(x)$.

$$F(x;\ \mu,\sigma,\varepsilon) = \exp\left(-\left[1+\varepsilon\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\varepsilon}\right) \tag{3}$$

Since the *significant* p-values (very small) depend on the right tail of the distribution, the correct estimation of the respective parameters μ, σ, and ε is central to the quality of the final p-values. An appropriate volume of background predictions for estimating the parameters of the GEV can be obtained by dinucleotide shuffling sequences that are actually present in the real search space (e.g. putative target sequences that are present in the investigated genome) and predicting RNA-RNA interactions for these shuffled sequences.

Given that optimal interaction energies E are minimal energies, we have to use negated energy terms E_n for the estimation of the GEV and their p-value computation. Furthermore, a length normalization of E_n is needed (Eq. 4) if targets of different lengths are used, since longer targets often enable larger interaction patterns and thus lower energies.

$$E_n = \frac{-E}{\ln(mn)} \tag{4}$$

C General joint structure approaches

Dmitri D. Pervouchine introduced and applied the first intermolecular RNA interaction search (IRIS) method [12] that can predict *general duplex structures* incorporating the structural context of both interacting RNAs. As for single RNA structure prediction, we will present the approach for the simplified base pair maximization scheme. Here, IRIS utilizes Nussinov's

recursion matrix N (Eq. 1) to compute the maximal intramolecular base pair number without intermolecular interactions. In the following, we denote with $N^{\mathcal{R}^1}$ and $N^{\mathcal{R}^2}$ the according matrices for the interacting RNAs $\mathcal{R}^1$ and $\mathcal{R}^2$, respectively.

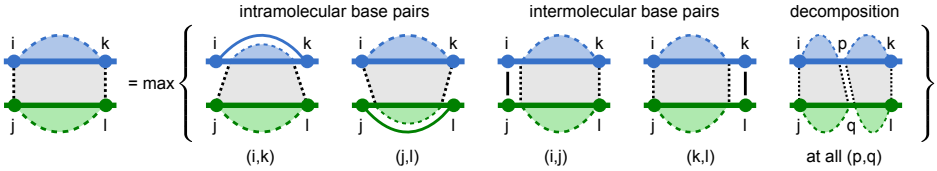

Figure 4: Depiction of the general joint structure prediction recursion of $M_{j..l}^{i..k}$ (Eq. 5), which handles intramolecular and intermolecular base pair extensions as well as a general decomposition case.

Entries $M_{j..l}^{i..k}$ in the duplex matrix M provide the maximal number of both intramolecular and intermolecular base pairs for the interacting subsequences $\mathcal{R}^1_{i..k}$ and $\mathcal{R}^2_{j..l}$. Here, entries with $i > k$ or $j > l$ correspond to the individual structure formation of $\mathcal{R}^1$ and $\mathcal{R}^2$, respectively, and are given by $N^{\mathcal{R}^1}$ and $N^{\mathcal{R}^2}$. The full recursion is provided in Eq. 5 and yields an algorithm with $O(n^6)$ time and $O(n^4)$ space complexity. A visual depiction of the recursion cases is given in Fig. 4.

$$M_{j..l}^{i..k} = \max \begin{cases} 0 & : \text{if both } j > l \text{ and } i > k \\ N_{i,k}^{\mathcal{R}^1}, \quad N_{j,l}^{\mathcal{R}^2} & : \text{no interaction considered} \\ & \quad \text{if } j > l \text{ or } i > k \\ M_{j..l}^{i+1..k-1} + 1, \quad M_{j+1..l-1}^{i..k} + 1 & : \text{intramolecular base pair} \\ & \quad \text{if } \mathcal{R}^1_i, \mathcal{R}^1_k \text{ or } \mathcal{R}^2_j, \mathcal{R}^2_l \text{ can pair} \\ M_{j+1..l}^{i+1..k} + 1, \quad M_{j..l-1}^{i..k-1} + 1 & : \text{intermolecular base pair} \\ & \quad \text{if } \mathcal{R}^1_i, \mathcal{R}^2_j \text{ or } \mathcal{R}^1_k, \mathcal{R}^2_l \text{ can pair} \\ \max\limits_{s,t} \left\{ M_{j..t}^{i..s} + M_{t+1..l}^{s+1..k} \right\} & : \text{decomposition of interaction site} \end{cases} \tag{5}$$

Note, for simplicity, the given recursion in Eq. 5 omits the minimal intramolecular base pair span s_l in the third case, which is enforced within the intramolecular folding algorithms. Furthermore, the recursion covers only interaction sites that are consecutive along the sequences. To allow for crossing interaction sites, the recursion can be extended [12]. As done by Zuker for the algorithm by Nussinov for single RNA structure prediction, Can Alkan and co-workers adapted the base pair maximization version of IRIS to derive an energy minimizing variant of equal complexity [13].

References

[1] Ydo Wexler, Chaya Zilberstein, and Michal Ziv-Ukelson. A study of accessible motifs and RNA folding complexity. *J Comput Biol*, 14(6):856–72, 2007. [PubMed:17691898] [doi:10.1089/cmb.2007.R020].

[2] Ydo Wexler, Chaya Ben-Zaken Zilberstein, and Michal Ziv-Ukelson. A study of accessible motifs and rna folding complexity. In Alberto Apostolico, Concettina Guerra, Sorin Istrail, Pavel A. Pevzner, and Michael S. Waterman, editors, *Proc. of the Tenth Annual International Conferences on Computational Molecular Biology (RECOMB'06)*, volume 3909 of *Lecture Notes in Computer Science*, pages 473–487. Springer, 2006. [PubMed:17691898] [doi:10.1089/cmb.2007.R020].

[3] Shay Zakov, Dekel Tsur, and Michal Ziv-Ukelson. Reducing the worst case running times of a family of RNA and CFG problems, using Valiant's approach. *Algorithms Mol Biol*, 6(1):20, 2011. [PubMed:21851589] [doi:10.1186/1748-7188-6-20].

[4] David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, 45(5):810–825, 1985. [doi:10.1137/0145048].

[5] Rolf Backofen, Dekel Tsur, Shay Zakov, and Michal Ziv-Ukelson. Sparse RNA folding: Time and space efficient algorithms. *J. Discrete Algorithms*, 9(1):12–31, 2011. [PubMed:7516239] [doi:10.1385/0-89603-276-0:267].

[6] Raheleh Salari, Mathias Möhl, Sebastian Will, S. Cenk Sahinalp, and Rolf Backofen. Time and space efficient RNA-RNA interaction prediction via sparse folding. In Bonnie Berger, editor, *Proc. of RECOMB 2010*, volume 6044 of *Lecture Notes in Computer Science*, pages 473–490. Springer-Verlag Berlin Heidelberg, 2010. [doi:10.1007/978-3-642-12683-3_31].

[7] Mathias Möhl, Raheleh Salari, Sebastian Will, Rolf Backofen, and S. Cenk Sahinalp. Sparsification of RNA structure prediction including pseudoknots. *Algorithms Mol Biol*, 5(1):39, 2010. [PubMed:21194463] [PubMed Central:PMC3161351] [doi:10.1186/1748-7188-5-39].

[8] L.G. Valiant. General context-free recognition in less than cubic time. *Journal of Computer and System Sciences*, 10:308–315, 1975. [doi:10.1016/S0022-0000(75)80046-8].

[9] Christina Schmiedl, Mathias Möhl, Steffen Heyne, Mika Amit, Gad M. Landau, Sebastian Will, and Rolf Backofen. Exact pattern matching for RNA structure ensembles. In *Proceedings of the 16th International Conference on Research in Computational Molecular Biology (RECOMB 2012)*, volume 7262 of *LNCS*, pages 245–260. Springer-Verlag, 2012. [doi:10.1007/978-3-642-29627-7_27].

[10] Sebastian Will, Christina Schmiedl, Milad Miladi, Mathias Möhl, and Rolf Backofen. SPARSE: Quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. In Minghua Deng, Rui Jiang, Fengzhu Sun, and Xuegong Zhang, editors, *Proceedings of the 17th International Conference on Research in Computational Molecular Biology (RECOMB 2013)*, volume 7821 of *LNCS*, pages 289–290. Springer Berlin Heidelberg, 2013. [PubMed:25838465] [PubMed Central:PMC4514930] [doi:10.1093/bioinformatics/btv185].

[11] E.J. Gumbel. *Statistics of extremes*. Columbia University Press, 1958.

[12] Dmitri D. Pervouchine. IRIS: intermolecular RNA interaction search. *Genome Inform*, 15(2):92–101, 2004. [PubMed:15706495].

[13] Can Alkan, Emre Karakoç, Joseph H. Nadeau, S. Cenk Sahinalp, and Kaizhong Zhang. RNA-RNA interaction prediction and antisense RNA target search. *J Comput Biol*, 13(2):267–82, 2006. [PubMed:16597239] [doi:10.1089/cmb.2006.13.267].