



Investigating the RFAM paradox: The pseudoknot explanation

Stefan Janssen, Yann Ponty, Balaji Raman, Saad Sheikh, Jean-Marc Steyaert,
Peter Clote

► **To cite this version:**

Stefan Janssen, Yann Ponty, Balaji Raman, Saad Sheikh, Jean-Marc Steyaert, et al.. Investigating the RFAM paradox: The pseudoknot explanation. Fifth Indo-French Bioinformatics Meeting, Mar 2011, Hyderabad, India. hal-00585647

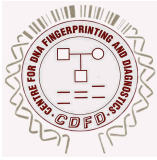
HAL Id: hal-00585647

<https://hal.archives-ouvertes.fr/hal-00585647>

Submitted on 13 Apr 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Investigating the RFAM paradox: The pseudoknot explanation

Stefan Janssen^{1,*}, Yann Ponty^{2,*}, Balaji Raman^{2,*}, Saad Sheikh²,
Jean-Marc Steyaert², Peter Clote^{2,3}

¹Faculty of Technology, Bielefeld University, Germany,

²Laboratoire d'informatique, École Polytechnique, Palaiseau, France,

³Biology Department, Boston College, Chestnut Hill, MA, USA.

(* equal contribution)

balaji@lix.polytechnique.fr

Motivation

Given the high cost and low-throughput of experimental methods to determine the structure of RNA, computational methods based on free-energy minimization (MFE), such as RNAfold [1] are routinely used for the *ab initio* prediction of secondary structures of RNA. Such methods are indeed substantially successful at predicting the secondary structure, and were shown to recover about 73% of base-pairs for RNAs of length less than 700 bps [2] on established benchmarks.

The RFAM database [3] groups most available RNA sequences into families sharing functional characteristics. Manually-curated multiple sequence alignments allow for the prediction of conserved structural elements through a mixture of semi-automated comparative prediction methods and experimental evidences. One of the denoted features of RFAM is the availability of a consensus secondary structure for each family, which is widely regarded as reliable, if slightly conservative. Therefore, it is expected that separate MFE predictions for family members largely overlap with their corresponding RFAM consensus.

The RFAM paradox

To confirm this intuition, we predicted MFE secondary structures for all 26,704 sequences of the 1,446 RFAM *seed* alignments. We systematically compared these predictions to their associated RFAM consensus secondary

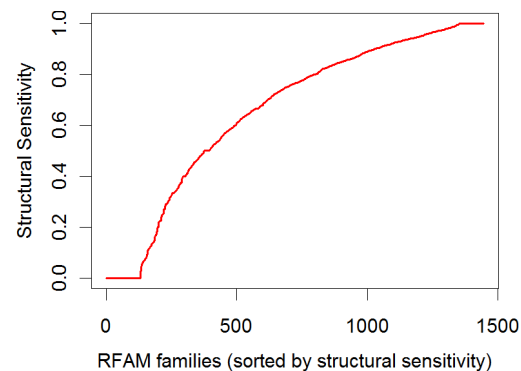
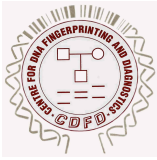


Figure 1: Average structural sensitivity of RNAfold predictions on families.

structure. Averaging over each family, we surprisingly observed that a large majority of families exhibited very little overlapping base-pairs between predicted structures and family consensus (Figure 1). For instance, RNAfold had absolute zero average sensitivity for around 12 % percent of the RFAM families.

The PK-oblivious algorithm explanation

As a first step towards a more comprehensive analysis of possible origins for such a discrepancy, we hypothesized the presence of complex structural features, *pseudoknots* (PK), as responsible for at least some of the apparent shortcomings of MFE approaches. Since PKs are typically absent from the search space of structure prediction methods, due to computational complexity reasons, the energy minimization scheme may be diverted toward structures having low-energy, yet sharing no resemblance with the RFAM consensus. Unfortunately, existing *ab initio* prediction



methods, including PKs, are either based on heuristics, or extremely time-consuming. Given the large scale of our experiment, this time-efficiency was especially critical. Therefore we designed a new method based on a recent contribution by one of the authors (pKiss [4]), combining an exact exploration of a restricted search space with a low time-complexity.

Pseudoknot detection

Our approach uses the pKiss software to compute the MFE conformation obtained by either excluding PKs altogether, or enforcing the presence of *simple canonical* or *kissing hairpins* PKs, the two predominant naturally-occurring PKs. For both types of PKs, the MFE differences (Δ_{PK} and Δ_{KH} respectively) between conformation spaces enforcing and precluding presence of pseudoknots, renormalized by the sequence length were computed (see Figure 2). If this difference in MFE is significantly large (or crosses some threshold), then the presence of PKs is suspected for the input RNA sequence. A similar approach was used for a prediction of ribosomal frameshift sites (KnotInFrame [5]). First we established the capacity of these indicators to discriminative PK families within RFAM. Each family in RFAM was annotated as pseudoknotted or not, depending on the presence of PKs in the consensus structure. ROC curves were computed for Δ_{PK} , Δ_{KH} and an optimal linear combination of both, yielding areas under the curve (AUC) respectively of 84.1%, 61% and 84.5% respectively. Moreover a large majority (88.3%) of families annotated within RFAM as featuring pseudoknots turned out to be associated with positive Δ_{PK} or Δ_{KH} values.

Towards a reannotation of PK families

Turning to our initial investigation of the responsibility of PKs for RNAfold's disagreement with RFAM's consensus, we investigated the 71 families having pseudoknot annotations in RFAM; the mean structural sensitivity of the 71 families is

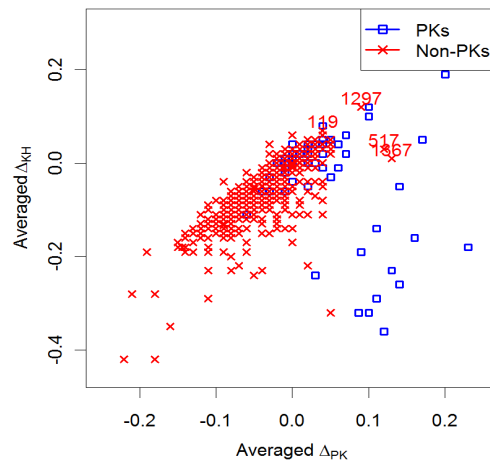


Figure 2: Plot of the two indicators Δ_{PK} and Δ_{KH} in presence/absence (blue/red) of PKs.

0.57. We found 509 RFAM families having positive Δ_{KH} or Δ_{PK} values in contrast to just 71 families in RFAM with pseudoknot annotations. We also manually investigated a short list of 15 of the 509 RFAM families. We found evidences of pseudoknots in the literature for at least 11 of these families. A plausible origin for the absence of such annotations may be found in the inherent conservatism of consensus approaches.

Many families remain associated with low RNAfold/RFAM overlap and low Δ_{KH} , Δ_{PK} values, pointing toward other explanations, such as the presence of non-canonical base-pairs (also excluded from MFE approaches) or multi-stable structure (for which the single-consensus view of RFAM may be overly restrictive). Investigating these complementary explanations should help gain insight on the shortcomings of *ab initio* folding methods, and help circumvent their current limitations.

References

- [1] Hofacker IL, et al. *Fast folding and comparison of RNA secondary structures*. Monatsh. Chem. 1994;125:167-188.
- [2] Mathews DH, et al. *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure*. J Mol Biol. 1999; May 21;288(5):911-40.
- [3] Gardner PP, et al. *Rfam: updates to the RNA families database*. NAR 2008.
- [4] Theis C, et al. *Prediction of RNA Secondary Structure Including Kissing Hairpin Motifs*. WABI 2010, LNBI 6293.
- [5] Theis C, et al. *KnotInFrame: prediction of -1 ribosomal frameshift events*. NAR 36(18), Pages: 6013-20, 2008