COMP564 Paper Review

Sequence-based identification of 3D structural modules in RNA with RMDetect

José Almeida Cruz and Eric Westhof

Jake Zhu

McGill

2019–04–02

Background

Problem: Provide Computational Search Tool for Structural RNA Module based on Sequence Information

- RNA is made up of Watson-Crick Pairs:
  1. Guanine–Cytosine
  2. Adenine–Uracil
- Structural RNA Modules are sets of ordered non-Watson-Crick Base Pairs embedded between Watson-Crick Pairs
- RNA Motifs, sets of Secondary Structure Elements, are Different from Sequence Motifs

Previous Solutions

- Motif Search Tools: `RNAMotif`, `MilPal` and `CMFinder`
- Requires Computer Resources
- Computer Expertise also Obligatory

Objectives

Objectives

- To Analyse Crystal Structures
- To Characterise a Module and its Interaction Network
- Interaction Networks depict
  1. Nucleotide Frequencies
  2. Base-Base Interactions

  for each Module Instance.

Challenges

- Only a few of the possible sequences are compatible with the given module are found in the crystal structure.
- Other Modules that are not covered by the Current Implementation of `RMDetect` exist.
- New Modules are likely to be Discovered

Methods

To capture all possible tertiary interactions and Base Pairs, `RMDetect` Relies on 3 Criteria:

1. Bayesian Network Models are Probabilistic Models where
   - Random Variables
   - Dependencies

   are represented as an Acylic Direct Graph
2. Base-Pair Probability Predictions
3. Positional Clustering of Candidates

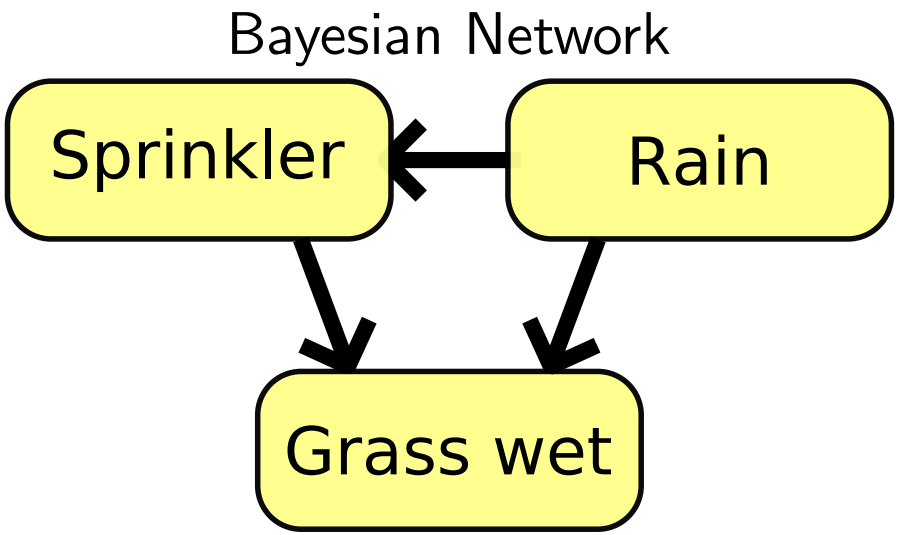


Figure: `https://en.wikipedia.org/wiki/Bayesian_network`

Methods

Single Sequence Search

- 15 Test Cases
- Previously Identified Modules in Crystal Structures
- Previously Obtained Reliable Sequence Alignments
- BPP = Base Pair Probability; M_{BN} = Bayesian Network Model; M_{GC} = Null Model; k, T = constants; FE = Free Energy
- $sp_{ij} = \{seq_i, seq_j\}$: pair of non-overlapping subsequences of S starting from i and j; WC_M = Set of all WC Pairs from M

$$score_{ij} = \log_2 \left(\frac{\Pr(sp_{ij}|M_{BN})}{\Pr(sp_{ij}|M_{GC})} \right), BPP_{ij} = \frac{\exp\left(-\frac{FE_{ij}}{kT}\right)}{\exp\left(-\frac{FE_{all}}{kT}\right)}$$

Multiple Sequence Search

- Same 15 Data Sets

Results

1. Mapped to Edges of the Bayesian Network:
   - All Watson-Crick Base Pairs
   - Most Non-Watson-Crick Base Pairs
   - Some Stacking Interactions
2. Single Sequence Search
   - True Positive Rate > 0.5
   - Exception: Tandem-GA Module where
     - False Discovery Rate > 0.5
3. Multiple Sequence Search
   - True Positive Rate $= 1$
   - 37 Instances
   - False Discovery Rate $= 0.23$
   - $\frac{5}{11}$ False Readings : Tandem-GA Module

Results

Why is it Significant?

- Fewer Computer Resources and Less Computer Expertise are needed to run `RMDetect`
- A Computational Tool for RNA Module Searching has been found.

Potential Pitfalls

1. Numerous False Positives in the Single Sequence Search
   - Small Size of Bayesian Networks
   - Short Alphabet of Nucleotides – due to the score equation
     1. Guanine
     2. Cytosine
     3. Adenine
     4. Uracil

Discussion

Significance

- Key Information on the Secondary Structure and Tertiary Fold can be provided by the Module in a RNA Sequence

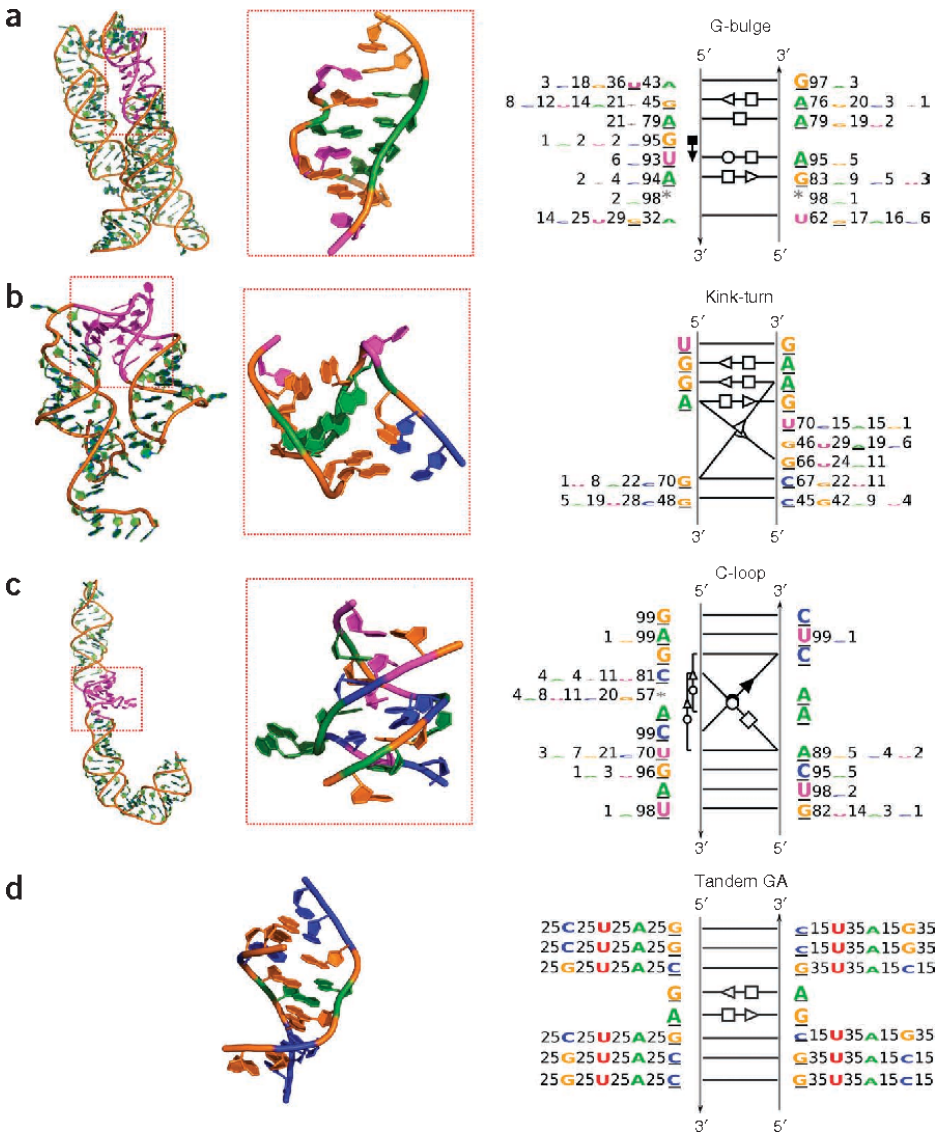
Impact – Can be Extended to:

- Searches in Public Databases
- Rfam Results
- Group I Intron Results
- Bacterial ncRNA Results

Conclusion and Future Work

- `RMDetect` (at the time) had issues, but it became an additional search tool.
- Today, the performance has been refuted: `JAR3D` and `BayesPairing` are two recent tools.
- According to the paper cited below, `RMDetect` does not allow the addition of more than a single model and "suffers of high computational costs and a minimal structural diversity in the modules considered due to its base pair probabilities scanning method."
- *Automated, customizable and efficient identification of 3D base pair modules with BayesPairing*. 4 March 2019. (Roman Sarrazin-Gendron[1], Vladimir Reinharz[2], Carlos G. Oliver[1], Nicolas Moitessier[3] and Jérôme Waldispühl[1,*])

Modules



Sequence Search Algorithms