

Simplicial Complexes and Graph Rewriting for Modelling RNA Folding

Adane Mamuye

Supervisor: Prof. Emanuela Merelli

Co-Supervisor: Dr. Luca Tesei

July 25, 2017



OUTLINE

Introduction

RNA Folding

Methodology

Results Discussion

Conclusion

Acknowledgement

COMPLEX SYSTEMS

A complex system displays collective behavior that **does not** trivially inferred from the behaviors of the individual components.

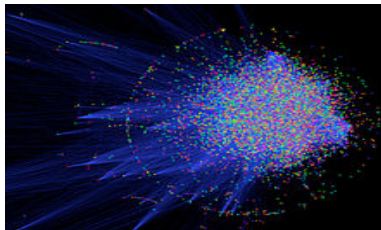


Figure: Protein interaction network (Wikibooks)

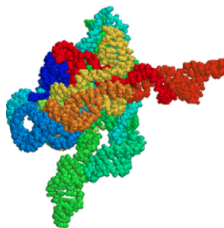
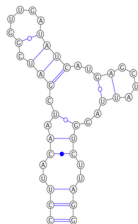
COMPLEX SYSTEMS MODELING CHALLENGES

Understanding of complex systems has become a **daunting challenge** due to:

- ▶ Unique properties, namely **nonlinearity, emergence and self-organization**
- ▶ Data voluminous

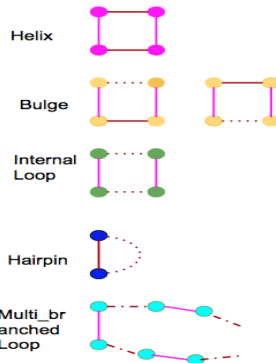
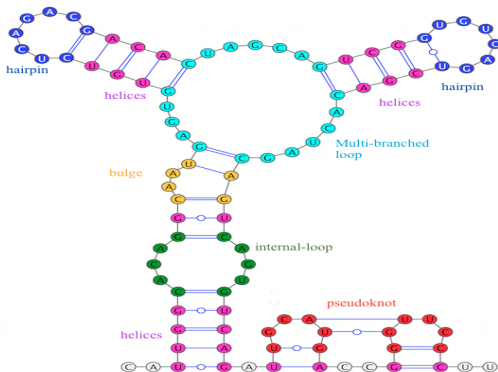
RNA

RNA is a complex biomolecule.



RNA SS SUB-STRUCTURES

RNA SS composed of sub-structures.



RNA FOLDING PROBLEM

Analyzing RNA SS is invaluable in **understanding genetic diseases and creating new drugs**.

Predicting pseudoknot free structure is **inaccurate** and predicting pseudoknotted structure is **NP-hard**.

Accurate prediction of RNA SSs is still an **ongoing challenge** in computational biology.

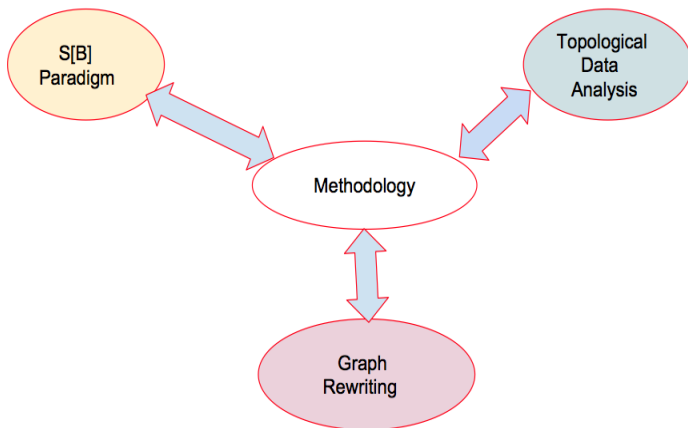
RNA FOLDING PROBLEM

The main goal of this thesis is to develop a new RNA secondary structure prediction model.

- ▶ Self adaptability behaviour

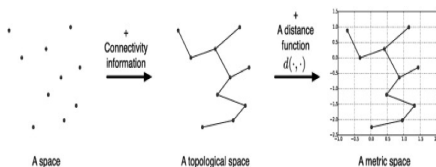
METHODOLOGY

Our methodology is based on three pillars.



TOPOLOGY

Topology is the branch of mathematics concerned with the **study of shapes up to different notions of equivalence.**



SIMPLICIAL COMPLEX

A simplicial complex is a geometrical representation of a topological space.

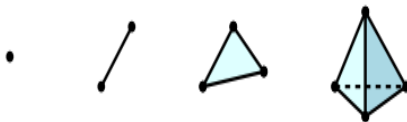


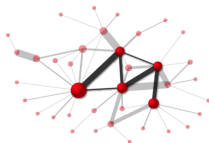
Figure: From left to right, 0-simplex, 1-simplex, 2-simplex and 3-simplex, respectively

BUILDING SIMPLICIAL COMPLEXES

- ▶ Point clouds: Cech and Vietoris-Rips



- ▶ Graphs: Clique



TOPOLOGICAL INVARIANTS-HOMOLOGY

Topology **classifies the topological space** using topological invariants, such as, homology and homotopy.

Homology groups **detect holes and loops indirectly by looking at the space surrounding them.**

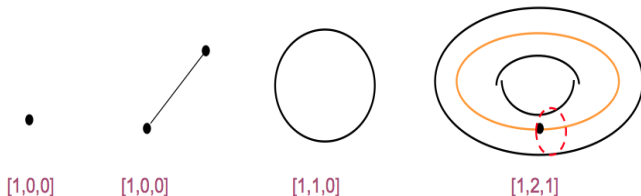
The *kth* homology group is

$$H_k = Z_k/B_k = \ker \partial_k / \operatorname{im} \partial_{k+1}$$

TOPOLOGICAL INVARIANTS-BETTI NUMBERS

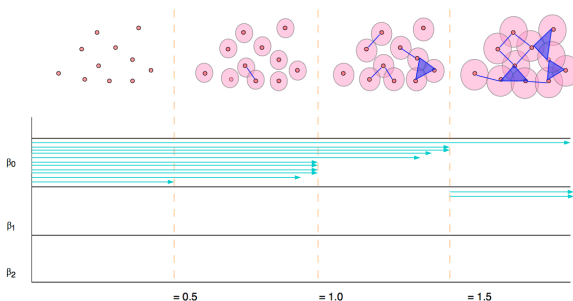
Betti numbers giveaway to count the number of distinct loops and holes.

- ▶ $Betti_0$ - connected components
- ▶ $Betti_1$ - one dimensional hole
- ▶ $Betti_2$ - two dimensional hole



PERSISTENT HOMOLOGY FILTRATION

Persistent homology studies how the invariants of the shapes evolve along the filtration.



TOPOLOGICAL DATA ANALYSIS

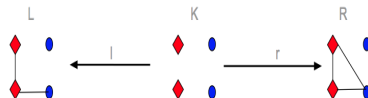
The aim of TDA is to infer relevant, qualitative and quantitative topological information from large, high- dimensional, and diverse datasets.

TDA measures the features of topological space based on topological invariants, such as Betti numbers .

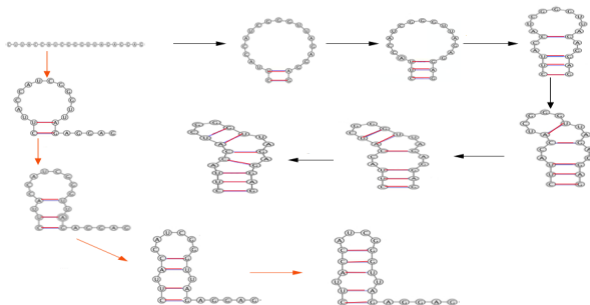
Persistent homology is a method in TDA appealing for various applications

GRAPH REWRITING

Graph rewriting is the **rule-based modification** of graphs, where each application of a graph rule leads to a graph rewriting



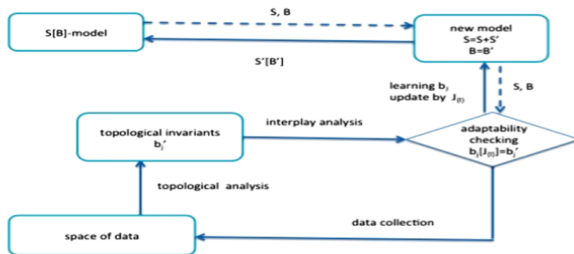
GRAPH REWRITING



THE S[B] PARADIGM

A computational model for self-adaptive complex systems with two coupled levels:

- Behavioral level:- B
- Structural level:- S

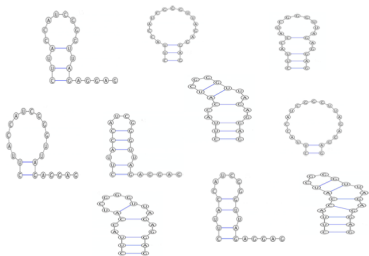


SIMPLICIA DATA ANALYSIS OF RNA

The folding space for RNA secondary structure with n nucleotides 1.8^n (Zuker & Sankoff, 1984)

Enumerating all possible RNA conformation **has been shown to be impractical for sequence longer than 40 nucleotides** (Cupal et. al., 1997)

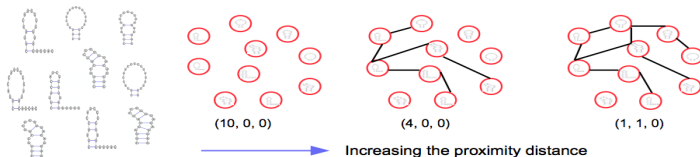
The folding space is **corresponding to multidimensional space** (Huang et. al., 2012)



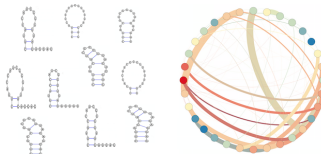
HOW TO BUILDING A SIMPLICIAL COMPLEX FROM THE FOLDING SPACE

We follow two different ways to construct a simplicial complex from the folding space.

- ▶ Folding space as a point cloud: using Vietoris-Rips

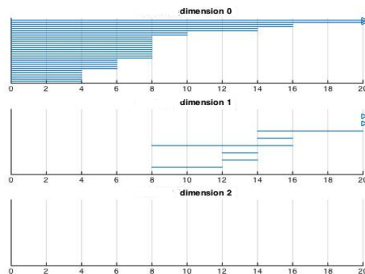


- ▶ Folding space as a complex graph: using CWRPH



RESULT ACHIEVED-I

► Persistent Homology Analysis of suboptimal structure



Further increasing ϵ to 20 results only two clusters.

RESULT ACHIEVED-I

- Persistent homology Filtration on 5s rRNA



Persistent Homology on RNA Secondary Structure Space

Full Text:  PDF  [Get this Article](#)

Authors: [Adane Mamuye](#) [University of Camerino, School of Science and Technology, Computer Science Division](#)

[Matteo Rucco](#) [University of Camerino, School of Science and Technology, Computer Science Division](#)

RESULT ACHIEVED-II

Betti number generators are used to derive five classes of RNA structural elements from the folding space.

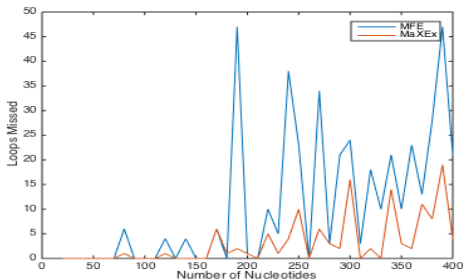


CWRP

Betti ₁ Generators		Classes of persistent loops
[3,8] [3,4] [4,5] [6,7] [8,7] [5,6] ; [20,19] [21,20] [17,21] [17,18] [18,19]		Hairpin
[3,8] [2,9] [2,3] [9,8] ; [22,21] [17,21] [16,17] [16,22] ; [2,9] [10,9] [1,2] [1,10] ; [21,29] [30,29] [21,20] [20,30] ; [15,16] [15,23] [23,22] [16,22] ; [14,15] [14,24] [15,23] [24,23]		Helix
[21,29] [31,30] [30,29] [17,21] [18,31] [17,18] ; [21,29] [17,21] [29,28] [22,28] [16,17] [16,22] ; [15,16] [24,23] [24,25] [25,26] [15,23] [22,28] [28,27] [26,27] [16,22] ; [21,29] [31,30] [30,29] [17,21] [17,32] [32,31]		Noise

RESULT ACHIEVED-II

The TDA results are compared with two prediction tools



RESULT ACHIEVED-II

We estimated the underlying linear relationship between the RNA sequence length and the number of persistent loops missed, by linear regression.



Mol. Based Math. Biol. 2016; 4:14–25

DE GRUYTER OPEN

Research Article

Open Access

Adane L. Mamuye*, Matteo Rucco, Luca Tesei, and Emanuela Merelli

Persistent Homology Analysis of RNA

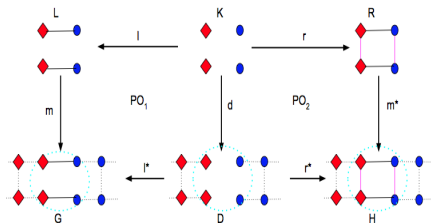
GRAPH REWRITING FOR MODELING RNA FOLDING

Graphical representation of RNA secondary structure provides the basis for modeling its folding process via graph rewriting.

Such transformations are expressed by derivations of a given graph grammar.

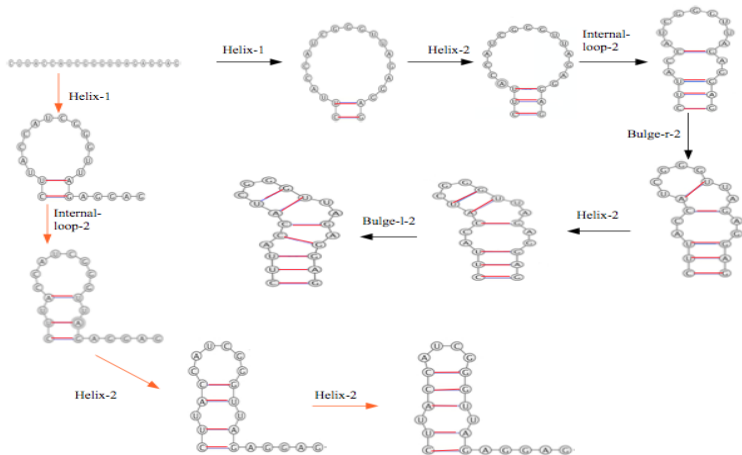
RNA graph grammar: $G_{RNA} = (\{p : (L \leftarrow K \rightarrow R)\}_{p \in P}, G_0)$

p (Helix - 1):

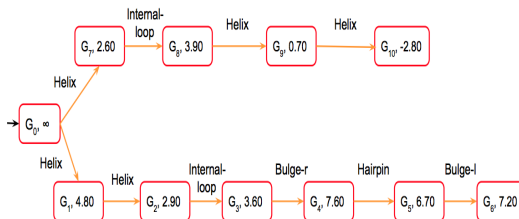


RNA FOLDING AS GRAPH REWRITING: A TOY EXAMPLE

Two possible folding transformations of an RNA molecule with 21 nucleotides



THE FOLDING PROCESS AS LTS



A Graph Grammar for Modelling RNA Folding

Adane Letta Mamuye

School of Science and Technology
University of Camerino
Via del Bastione 1 62032
Camerino, Italy
adaneletta.mamuye@unicam.it

Emanuela Merelli

School of Science and Technology
University of Camerino
Via del Bastione 1 62032
Camerino, Italy
emanuela.merelli@unicam.it

Luca Tesei

School of Science and Technology
University of Camerino
Via del Bastione 1 62032
Camerino, Italy
luca.tesei@unicam.it

RNA S[B]- CHALLENGES

Challenges:

- ▶ Graph transformation can be used to model the folding process of RNA, but not able to identify the minimum free energy secondary structure.
- ▶ Searching for the minimum free energy structure is computational demanding

REVERSIBLE MODEL FOR RNA FOLDING

$$p : L \leftarrow K \rightarrow R$$

We can obtain the reverse rule

$$p^{-1} : R \leftarrow K \rightarrow L$$

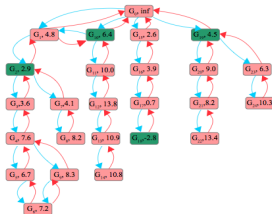
We used the S[B] paradigm

- ▶ **B (behavioural level)**: models the folding process
- ▶ **S (structural level)**: controls the process by imposing constraints over B.

RNA S[B]- BEHAVIORAL LEVEL

The behavioral level $B = \langle Q, \rightarrow_B, \mathcal{L}, q_0 \rangle$ is a LTS .

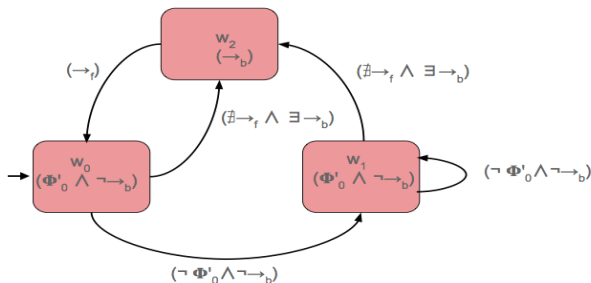
- ▶ G : the secondary structure derived at each transformation (translation);
- ▶ e : energy of the derived secondary structure in kcal/mol;
- ▶ $\rightarrow_B$: the transition type, $\rightarrow_B = \rightarrow_f \cup \rightarrow_b$ where $\rightarrow_b = (\rightarrow_f)^{-1}$.



RNA S[B]- STRUCTURAL LEVEL

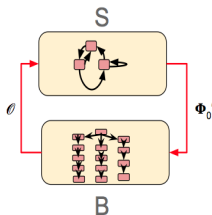
The structural level $S = \langle W, \rightarrow_S, \mathcal{O}, \mathcal{A}, w_0 \rangle$

The structural level S controls the adaptation dynamics of the B level towards the lowest MFE secondary structure.



$$\Phi'_0 = \exists q_{\text{next}} \in \text{forward}(q) : e(q_{\text{next}}) \leq e(q)$$

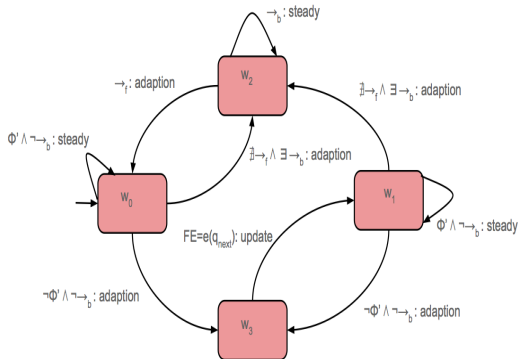
RNA S[B]- CHALLENGES



This model faces four main challenges, namely

- ▶ Livelock
- ▶ Non-terminating
- ▶ Memorylessness
- ▶ State space explosion

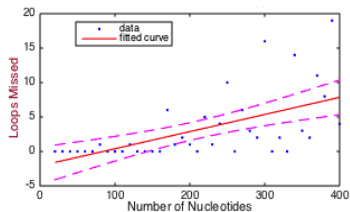
RNA S[B]- NEW STRUCTURAL LEVEL



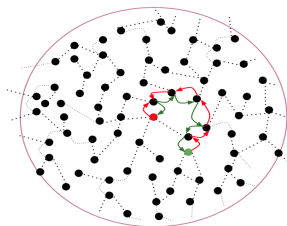
$$\Phi'_0 = \exists q_{next} \in \text{forward}(q) : (e(q_{next}) \leq e(q)) \wedge (\text{explore}(q_{next}) == 0)$$

ANALYSIS OF SIMPLICES FOR STATE SPACE REDUCTION

Confidence intervals are used to get sense of whether the prediction estimates are any good.



ANALYSIS OF SIMPLICES FOR STATE SPACE REDUCTION



The states in the big circle represent the subspace of the B-level as predicted by the upper bound regression line.

CONCLUSION

The proposed approach combines data driven techniques with formal approaches for modeling the RNA secondary structure prediction.

The proposed approach, only focus on theoretical modeling; thus, developing RNA secondary structure prediction tool is the work to be done.

PUBLICATIONS

Articles

- ▶ Adane L. Mamuye, Matteo Rucco, Luca Tesei, Emanuela Merelli. Persistent Homology Analysis of RNA. Journal of Mol. Based Math. Bio., 4:14–25, 2016.

Papers on Proceedings

- ▶ A. L. Mamuye, E. Merelli, L. Tesei. A Graph Grammar for Modeling RNA Folding. Electr. Notes Theor. Comput. Sci. 2nd Graphs as Models workshop (GaM 2016) EPTCS 231, 2016, pp. 31-41.
- ▶ Matteo Rucco, Adane Letta Mamuye, Marco Piangerelli, Michela Quadrini, Luca Tesei, Emanuela Merelli. Survey of TOPDRIM Applications of Topological Data Analysis. Proceedings of the 2nd International Workshop on Knowledge Discovery on the WEB-KDWEB, Cagliari, 2016.

PUBLICATIONS

Papers on Proceedings

- ▶ A. Mamuye, E. Merelli. and L. Tesei. *Towards a Shape Language for Interpreting RNA Folding*. BICT'15 Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), pp. 185-188, 2015.
- ▶ A. Mamuye and M. Rucco. *Persistent Homology on RNA Secondary Structure Space*. BICT'15 Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), pp. 189-192, 2015.

ACKNOWLEDGEMENT

- ▶ Our study is financed by the **European project: TOPology DRIVEN Methods for complex systems (TOPDRIM)** - FP7-ICT/ICT-2011.9.7. <http://www.topdrim.eu>
- ▶ **University of Gondar** (<http://www.uog.edu.et/en/>)

Thank You!