Persistent Homology on RNA Secondary Structure Space

Adane Mamuye
University of Camerino
School of Science and Technology
Computer Science Division
adaneletta.mamuye@unicam.it

Matteo Rucco
University of Camerino
School of Science and Technology
Computer Science Division
matteo.rucco@unicam.it

ABSTRACT

In this ongoing work, we employ persistent homology to analyze the RNA suboptimal structure space and secondary structure space of 5s rRNA. Our basic idea is to measure the topological similarity of the suboptimal structures by sliding the minimum free energy window. The preliminary result shows that the more the energy window slides from the lowest minimum energy structure the more dissimilar the structures with the lowest one. In addition, persistent homology is also used to classify 5s rRNA of Archaea species into different family of classes. Our preliminary result shows that structural dissimilarity can be observed even for species that are grouped under the same genus.

Categories and Subject Descriptors

J.1 [**Computer Applications**]: Miscellaneous; J.4 [**Life and Medical Sciences**]: [Computational Biology- topological data analysis]

Keywords

Persistent homology, RNA secondary structure, topological invariant

1. INTRODUCTION

Ribonucleic Acid (RNA), single strand biological molecule, consists of a sequence of nucleotides: adenine (A), guanine (G), cytosine (C) and uracil (U). This sequence folds to itself; this folding leads to the formation of secondary structure. Secondary structures determine the functionality of RNA molecules. Numerous methods have been used to predict the optimal secondary structure (the minimum free energy structure). The optimal secondary structure is not always right structure; thus suboptimal structures, structures with free energy similar to the lowest free energy structure, provide structural alternatives to the optimal one. As a result, attempt had been made to obtain a holistic view of the near-optimal or even the complete folding space a given RNA sequence by generating suboptimal solutions [5]. Besides, Josef et al. [8] demonstrated that suboptimal structures are capable of capturing RNA homology (structural similarity in biological sense) even in divergent bacterial species. Moreover, Akihiro et al. [1] developed a method for classifying suboptimal secondary structures of a given RNA sequence and applied this metric to classify secondary structures derived from a set of mutant RNA sequences. Robert et al. [11] released a program called PaRNAss to search for novel ribosswitches. This program generates sample of suboptimal structures and then cluster them by computing the pairwise distance between them. On the other hand, all possible secondary structures (structure space) of a given sequence was considered to infer information about similarity of structures. Lorenz et al. [10] presented a tool for characterizing riboswitches by classifying all structures in a structure space into two reference structures based on their similarity distance. However, it would be significant to understanding the relevant similarity information hidden in the RNA secondary structure space. One of the methods to extract useful information from such space is persistent homology. Persistent homology, branch of algebraic topology, measures topological features of a given data space [3].

Persistent homology has been recently used as a promising tool for the topological simplification of complex data at a different spatial resolution. Xia et al. [15] presented a multiresolution persistent homology analysis to reveal the intrinsic topological invariants of an RNA tertiary structure. Besides, persistent homology has been proposed to the study of evolution by considering a set of genomes and calculate the genetic distance between each pair of sequences [2]. However, in the study of molecular evolution, the global relationships between sequence and structure are more relevant [4]; as a result, in addition to the sequence similarity, the structural similarity need to be investigated.

In this ongoing work, persistent homology is used to probe RNA structure space in two different views. In the first case, persistent homology is employed to scrutinize topological features from suboptimal RNA structure space of a sequence. This analysis suggests alternative set of conformations that are structurally similar to the lowest free energy structure. Moreover, analyzing the suboptimal structure space may give a clue about the structure-function relationship of the RNA molecule. Secondly, we use persistent homology to analyze the structure similarity among species of six families of archaea (known to be metabolically diverse organisms coexisting with Bacteria and Eukarya [6]) so as to know topological proximity among them. This information help identifying structural homology (in biological sense) among species. The paper is organized as follows. In Section 2, a theoretical background is given about persistent homology. Section 3 illustrates the persistent homology analysis result. Conclusion and future work is given in Section 4.

2. MATERIALS AND METHODS

2.1 Topological Data Analysis

Topological data analysis measures the features of a topological space based on topological invariants. Topological

BICT 2015, December 03-05, New York City, United States
Copyright © 2016 ICST
DOI 10.4108/eai.3-12-2015.2262543

invariants, algebraic objects which is invariant under homeomorphisms, are often preferable to analyze the topological space. The invariants can be computed from a simplicial complex of the topological space. A simplicial complex is a geometrical representation of a topological space which is realized as combination of simplices. In the following we will give a highlight about persistent homology, for a broad description we recommend [3, 16]

2.1.1 *Simplicial Complex*

The simplicial complex (K) is a k-simplex, σ^k, which is defined as the convex hull of $k+1$ affine independent points. The $k+1$ points are affinely independent iff the k vectors $v_i - v_0$, for $1 \leq i \leq k$, are linearly independent. If the convex hull consists the independent points $v_0, v_1, v_2, v_3, ..., v_k$; $\sigma^k = \{v_0, v_1, v_2, v_3, ..., v_k\}$, a k-simplex can be defined as:

$$\sigma^k = \{\lambda_0 v_0 + \lambda_1 v_1 + \ldots + \lambda_k v_k \mid \sum_{i=0}^{k} \lambda_i v_i; \sum_i \lambda_i = 1, i = 0, 1, \ldots, k\}$$

The dimension of the simplicial complex is the maximum dimension of any of its simplices.

In simplicial complex a vertex is a 0-simplex, a line segment connecting two vertices is a 1-simplex, a triangle is a 2-simplex , a tetrahedron is a 3-simplex and so on. In general, a simplicial complex K needs to fulfil two conditions: the faces of the simplex K should be also in K and the intersection of any two simplices in K is either empty or share common faces. Simplicial complexes of topological spaces are characterized by their corresponding topological invariants, such as persistent homology.

2.1.2 *Persistent Homology*

One can measure the topological features of the data by creating connections between proximate data points, varying the scale over which these connections are made, and looking for features that persist across scales; this is called persistent homology. The basic aim of persistent homology is to build an increasing family of simplicial complexes. The simplicial complex is indexed by a scale parameter around the data points while carefully keeping track of the appearance and disappearance of topological features, i.e, β_0 (connected components), β_1 (one dimensional holes), β_2 (two dimensional holes (voids)) and their higher-dimensional holes, at different spatial resolution [13].

2.1.3 *Vietoris-Rips Complex*

There are various ways of building the abstract simplicial complex from the point clouds (finite sets of points equipped with a distance function); however, we are interested only in Vietoris Rips. Vietoris-Rips complex is one of the most popular complexes in persistent homology and it constructs a simplex if the largest distance between any two points in a metric space is at most ϵ, i.e., the proximity parameter. Given a metric space X and a proximity parameter $\epsilon > 0$, the Vietoris-Rips simplicial complex $(X; \epsilon)$ has as its simplices the finite subsets of X of diameter less than ϵ. If each point in the metric space is the center of a ball; increasing simultaneously the radii results two or more intersecting balls. This intersection adds k-simplex any time we see a subset of k points with common intersection, see Figure 1.

2.2 Data and Similarity Measure

The suboptimal structures of an RNA sequence of 120 nucleotides are generated using RNAsubopt [9]. RNAsubopt, Vienna RNA package [1], reads RNA sequence and calculates all suboptimal secondary structures within a user defined energy range higher than the minimum free energy. Besides, 63 RNA sequences of 34 species of Archaea namely: Archaeoglobales, Halobacteriales, Methanobacteriales, Methanococcales, Methanomicrobiales and Methanosarcinales are taken from 5S rRNA database [12]. RNAfold [9] is used to compute the the minimum free secondary structure of the 63 sequences. RNAdistance [9] is used to compute the distance between secondary structures.

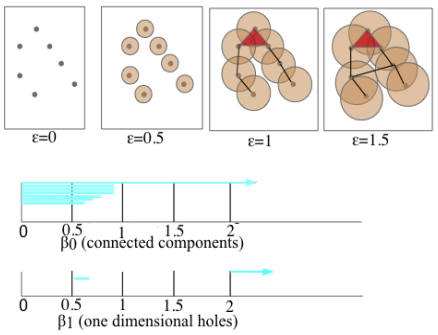

Figure 1: Example of the Vietoris-Rips complex construction. The top four figures illustrate the simplicial complex of 7 points at filtration values $\epsilon = 0$, $\epsilon = 0.5$, $\epsilon = 1$ and $\epsilon = 2$. The two barcodes correspond to β_0 and β_1 show the filtration evolutions at the specified ϵ values as presented on the horizontal axis of the barcodes.

3. RESULTS

3.1 Persistent Homology Analysis of Suboptimal structures

In this ongoing work, we employ topological techniques to classify suboptimal secondary structures within a small energy interval. Since the lowest free energy common structure is not always the correct structure, topological analysis help to suggest alternative set of conformations that are structurally similar to the lowest free energy structure. Here, RNA suboptimal structures are considered as a point could data coming from a sequence; then RNA distance is used to compute the similarity between each point. Since we obtained a distance function on the set of point clouds; Rips filtration is performed over them. A set of simplicial complexes generated throughout the filtration process provides information about the topological space of the suboptimal structures. In this ongoing study, all the filtrations are performed by using JavaPlex [13] and the topological invariant, Betti numbers, are visualized through barcodes. To select an appropriate similarity measuring technique, we used topological filtration on two different topological spaces that are constructed by using tree edit distance and string edit distance [9]. Then a sample structure space, selected from a large structure space with 220, 085 suboptimal structures, is analyzed.

3.1.1 *Selecting Appropriate Similarity Measure*

[1] http://www.tbi.univie.ac.at/RNA/

To infer the similarity between two or more RNA secondary structures, we use RNA distance. It calculates the tree edit distance and string edit distance half-matrices of the suboptimal structures. The proximity matrix entry at $[i, j]$ refers to the pairwise similarity distance between the i^{th} and j^{th} optimal structures. The smaller the value of the entry the higher the similarity. This metric is suffice to construct the Rips filtration.

The barcodes of the Rips filtration of the point cloud of RNA suboptimal structures at tree edit distance $\epsilon = 0 - 7$ are illustrated in Figure 2. The three panels, in the same Figure, represent β_0 (different clusters in our case), β_1 and β_2, from top to bottom. In this case, from $\epsilon = 0$ to $\epsilon < 2$, β_0 have 20 bars that indicate the total number of suboptimal structures considered. Then after $\epsilon = 2$ a transition occurs with $\beta_0 = 4$, $\beta_1 = 6$ and $\beta_0 = 0$. Suboptimal structures with tree edit distance less than or equal ϵ are contained in the same connected component; for instance, at $\epsilon = 2$ neighbor suboptimal structures are clustered as one family of structures. Thus, the four distinct connected components at β_0 refer to four different family of structures. Suboptimal structures are grouped in the same structure family, if the tree edit distance between them is equal to/smaller than ϵ. Conversely, if two suboptimal structures are from different classes they are far from each other (roughly speaking their similarity distance is greater than the threshold ϵ).

RNAsubop generates, in many cases, more than one structures of the same free energies. Suboptimal structures with the same energy values can be considered as neighboring structures [14]. Consequently, filtration at $\epsilon = 2$ tells us that all the existing structures with the same minimum free energy is grouped in the same family of structures. This indicates that suboptimal structures of the same minimum free energy are more similarity than others. On the other hand, one dimensional holes, $\beta_1 = 6$ are polygons formed by closed loops that are not triangles. One dimensional rings are created, if the similarity cost of nonadjacent suboptimal structures of the ring are beyond the filtration values.

When the filtration value set to $\epsilon = 5$, the 4 clusters are

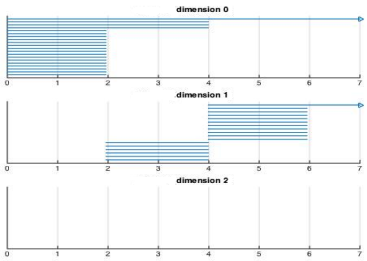

Figure 2: Persistent homology analysis on the tree edit distance matrix of RNA suboptimal structures. The horizontal axis refers to the filtration values. The three panels are barcodes for β_0 and β_1 and β_2, respectively.

merged at $\epsilon = 4$ into one cluster that persist to the end. Thus, increasing the filtration parameter to $\epsilon = 5$ results only one connected component with $\beta_1 = 11$ and $\beta_2 = 0$; all suboptimal secondary structures are put into a single cluster. At filtration $\epsilon = 7$ $\beta_1 = 1$ which implies the existence of suboptimal structures in the cluster with the tree edit distance threshold greater than 7.

On the other hand, RNAdistance, is used to compute the string edit distance between the RNA suboptimal structures. The result shows that even at the lowest string edit one connected component is constructed. From this, we can easily observe the difference between the tree edit distance and string edit distance methods. The string-based technique uses only the information whether a nucleotide is paired or unpaired, but neglect the connectivity information who pairs with whom [7]. Due to this limitation the persistent homology results over the string edit distance matrix are not interesting. As a result, in the rest of this ongoing work we use tree edit distance as a similarity measuring technique.

3.1.2 *Analysing the Suboptimal Structure Space*

The number of suboptimal structures grow exponentially within the defined energy range since there are many suboptimal structures with the same minimum free energy. However, all the suboptimal structures with the same free energy are not considered. As of our previous discussion, suboptimal structures with the same minimum free energy are grouped in the same structure family; hence, the first suboptimal structure from the sorted sequence of structures, with the same minimum free energy, is selected as a representative of the remaining structures. Accordingly, from the $220,085$ suboptimal structures, we draw 30 structures by sliding the energy window with $+0.3$ from the lowest free energy -52.70 to -44.30.

The persistent homology filtrations at $\epsilon = 5$, $\epsilon = 10$ and $\epsilon = 20$, are illustrated in Figure 3. At $\epsilon = 5$ there are 24 clusters ($\beta_0 = 24$); four of them consist of more similar suboptimal structures namely $(1, 2, 3)$, $(4, 6)$, $(5, 7)$ and $(11, 12, 14)$; while each of the remaining 20 connected components consists only one suboptimal structure. When the filtration parameter increases more similar clusters are merged as one cluster. This decreases the number of connected components; for instance, at $\epsilon = 10$ $\beta_0 = 6$, the connected components are reduced to 6 in which 27 of the suboptimal structures are clustered in three structure families. However, three suboptimal structures, i.e., 19, 28 and 29 exist as three clusters. Further increasing ϵ to 20 results only two clusters. The first 21 suboptimal structures including the lowest energy structure together with $23^r d$ and 24^{th} suboptimal structures are grouped in the same cluster, whereas the remming suboptimal structures grouped in the second family of structures.

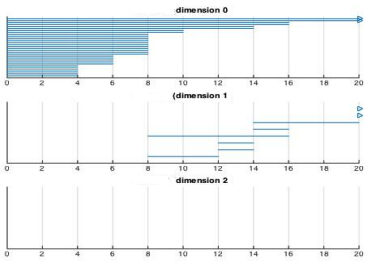

Figure 3: Persistent homology analysis on suboptimal structures.

3.2 Persistent homology Filtration on 5s rRNA

The ribosome is made up of dozens of distinct proteins as well as a few specialized RNA molecules known as ribosomal RNA (rRNA). The core function of the ribosome is

basically the same across different groups of organisms, but this does not mean the rRNAs are identical between species. The actual sequence of the nucleotides in rRNAs does vary between species. However, the mutation process may leads to large ensembles of RNA sequences that fold into the same secondary structure. In this ongoing work, we use persistent homology to capture 5s rRNAs structural similarities among species of the six families of Archaea namely: Archaeoglobales, Halobacteriales, Methanobacteriales, Methanococcales, Methanomicrobiales and Methanosarcinales that are taken from 5S rRNA database [12].

Our approach is based on the lowest free energy structures of the 63 RNA sequences of the species that are generated using RNAfold. To make the structural similarity comparison among the 63 structures, only the topological invariant β_0 is used.

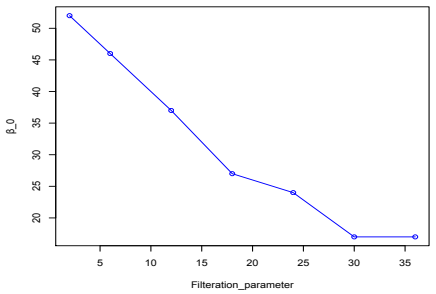

Figure 4: The vertical axis refers to the number of connected components β_0 at the filtration value ϵ of the horizontal axis.

At a particular scale ϵ, β_0 represents the number of different family of structures which contain similar secondary structure of different species. In the construction of simplicial complexes, two structures are considered as related (connected by a line) if the tree edit distance between them is smaller or equal to ϵ. Structurally similar species lie in neighborhood of each other on the topological space. Figure 4 shows the distribution of the 63 structures in different clusters (β_0), at $\epsilon = 2$ to $\epsilon = 36$ with increment of 4. At $\epsilon = 12$ two different Halobacterials, i.e, Holococcus Morrhuae species, are grouped together in the same cluster. On the other hand, three Halobacterials species, Haloferax Moditerranei, are found in different clusters. When we increase to $\epsilon = 30$ and then to $\epsilon = 36$; there is no change in the number of connected components; hence, the filtration value $\epsilon = 30$ can be taken as an optimal threshold for filtration. In general, from the filtration evolution, we can assert that being in the same species family doesn't mean they have the same secondary structures. The sequence difference among the species affect the structural similarity among species of the same family.

4. CONCLUSIONS

In this ongoing work, we study the application of persistent homology to capture topological features of two datasets namely the RNA suboptimal structure space and the optimal structure space of selected families of Archaea. Our preliminary result shows that sliding the energy window far from the lowest free energy results the increase in dissimilarity of structures with the lowest minimum free energy structure. Persistent homology is also used to captures important information from secondary structure space of species of different family; it clusters the species minimum free energy structure into different families. The persistent homology analysis shows that structural dissimilarity can be observed even for species that are classified under the same genus. This result is similar with the result of Josef et al. [8]; they mentioned that structural dissimilarity among 6S RNAs was observed even for closely related species.

In the future, concerning the suboptimal structure space we will use hamming distance or some others to measure the similarity between clusters. Concerning 5s rRNA, We will use our results to construct the phylogeny tree of the 63 species. we will also focus on extending the analysis on large datasets and construct a phylogenetic tree for species by taking into account both the structural similarities and sequence similarities.

5. REFERENCES

[1] N. Akihiro, Y. Akinori, and Y. Kenji. Classification of RNA Secondary Structures Using the Techniques of Cluster Analysis. *J. theor. Biol.*, 183:105–117, 1996.

[2] J. M. Chan, G. Carlsson, and R. Rabadan. Topology of Viral Evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571, 2013.

[3] H. Edelsbrunner and J. Harer. Persistent Homology a Survey. *Contemporary mathematics*, 453:257–282, 2008.

[4] D. Evers and R. Giegerich. RNA movies: Visualizing RNA Secondary Structure Spaces. *Bioinformatics*, 15(1):32–37, 1999.

[5] R. Giegerich, V. Björn, and R. Marc. Abstract Shapes of RNA. *Nucleic Acids ResNucleic Acids Res.*, 32(16):4843–4851, 2004.

[6] S. Gribaldo and C. Brochier-Armanet. The Origin and Evolution of Archaea: a State of the Art. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1470):1007–1022, 2006.

[7] I. L. Hofacker and P. F. Stadler. *RNA secondary structures*, volume volume 1, pages pages 439–489. Wiley-VCH, Weinheim, Germany, 2007.

[8] P. Josef, K. Libor, B. Jan, J. Edita, K. Jana, and V. Již Í. The Suboptimal Structures Find the Optimal RNAs: Homology Search for Bacterial Non-coding RNAs Using Suboptimal RNA Structures. *Nucleic Acids Research*, 39(8):3418–3426, 2011.

[9] R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, et al. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[10] R. Lorenz, C. Flamm, and I. Hofacker. 2D Projections of RNA Folding Landscapes. In *GI, Lecture Notes in Informatics, German Conference on Bioinformatics*, volume 157, pages 11–20, 2009.

[11] G. Robert, H. Dirk, and R. Marc. Prediction and Visualization of Structural Switches in RNA. In *Symposium on Biocomputing*, volume 4, pages 126–137, 1999.

[12] M. Szymanski, M. Z. Barciszewska, V. A. Erdmann, and J. Barciszewski. 5s Ribosomal RNA Database. *Nucleic Acids Research*, 30(1):176–178, 2002.

[13] A. Tausz, M. Vejdemo-Johansson, and H. Adams. Javaplex: A Research Software Package for Persistent (co) homology. *Software available at http://code. google. com/javaplex*, 2011.

[14] B. Voß. Advanced Tools for RNA Secondary Structure Analysis. 2004.

[15] K. Xia, Z. Zhao, and G.-W. Wei. Multiresolution Topological Simplification. *arXiv preprint arXiv:1504.00033*, 2015.

[16] A. J. Zomorodian. *Topology for Computing*, volume 16. Cambridge University Press, 2005.