Large RNA Secondary Structure Conservation Annotation Using Secondary Structure-Based MSA

Jan Pešek, David Hoksza*
Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.

* Corresponding author. email: david.hoksza@mff.cuni.cz
Manuscript submitted December 4, 2015; accepted January 18, 2016.
doi: 10.17706/ijbbb.2016.6.1.18-25

Abstract: Identification of conserved regions of a set of RNA secondary structures is currently an open research problem when dealing with large RNA molecules such as ribosomal RNA. We designed and implemented a method for conservancy annotation of a set of RNA molecules using their secondary structures. The method first converts secondary structures into linear representations, which are then forwarded into multiple sequence alignment (MSA). The resulting secondary structure-based MSA is subsequently passed into a conservancy identification procedure which uses a sliding window technique to identify conserved position in the MSA and assign them a score based on the secondary structure content of the window. The algorithm can be used to rank overall conservancy of the structures, which generally denotes evolutionary distance, as well as to assign conservancy to individual bases to identify high- or low-conservancy regions. We tested the algorithm for correlation with evolutionary distance, where it matches the expectations. The method is freely available as a stand-alone tool implemented in the Python programming language.

Key words: RNA secondary structure, conservation, multiple sequence alignment.

1. Introduction

Ribonucleic acids (RNA) play an important role in many processes related to protein synthesis or regulation of genetic expression [1]. Function of an RNA molecule is determined by its structure, i.e. a shape that is formed by the molecule. There are several levels of description of this structure. The simplest, primary structure, expresses an RNA molecule as a linear sequence of nucleotides. Tertiary structure, on the other hand, describes the full three-dimensional structure of the molecule. The tertiary structure is a labeled set of 3D coordinates, thus representing the most informative model of an RNA molecule. However, obtaining the tertiary structure of a molecule is a complex problem and therefore, due to the progress in sequencing technology, for most of RNA molecules the sequence information is available while the corresponding three-dimensional structure is not. In between, secondary structure does not tell atom positions in three-dimensional space, but rather describes the pattern of hydrogen bonds between bases (nucleotides). Two bases that are part of a hydrogen bond are called base pairs. The base pairs are frequently organized into specific substructures such as helices, loops, junctions, hairpins, overhangs bulges or pseudoknots. These are called structural features or motifs and are considered to be the "building blocks" of RNA secondary structures. Secondary structure information is a relatively good approximation of the tertiary structure, since it shows which parts of the sequence get near each other in the resulting 3D fold. While the computational prediction of tertiary structure is currently not possible, fortunately, there

are algorithms which can predict secondary structure of an RNA molecule from its sequence for even the largest RNAs, such as ribosomal RNAs [2].

One of the common tasks when analyzing RNA molecules is to study the conservancy at the level of entire molecules, or even at the level of individual segments, motifs or even residues. The goal of conserved regions identification is to reveal functionally important parts of a molecule. Or, inversely, knowledge of unconserved regions can help to discover organism-specific portions of molecules and thus drive the consequent analysis. An example of such important regions can be expansion segments of ribosomal RNAs [3].

Conservancy expresses the level of homology of molecules or their parts and is traditionally associated with sequences, i.e. molecules are considered conserved if they show high sequence similarity. However, since conservancy is an evolutionarily-based term, it is actually related to the underlying structures. So we can deduce conservancy of molecules, or their parts, based on the similarity of their tertiary or secondary structures, i.e. similarity in base pairing patterns. Measuring conservancy on the structural level can reveal conserved regions that are not obvious from the sequence itself: very similar secondary (and tertiary) structure can be formed by relatively different sequences of nucleotides, and as the biological function of a molecule depends more heavily on its structure than sequence, measuring conservation at the sequence level only can obscure functional similarities.

The availability of reliable algorithms for secondary structure prediction of large RNA molecules calls for tools enabling measurement of conservancy of even such large structures. Although there exist several approaches for measuring conservancy of entire structures, a tool which would enable measuring conservancy of individual nucleotides for large structures based on the secondary structure is basically missing. We call this task conservancy annotation, since such a method can annotate each position with a secondary structure-based conservation level. A nice overview of approaches to secondary structure conservation identification of whole molecules can be found in [4], where the approaches are divided into three classes: (i) comparison of predicted minimum free energies, (ii) comparison of single structures, (iii) comparison of ensembles of structures representing the whole folding space. With the exception of comparison based on single structures, all the methods are based on energy computations which tend to be computationally demanding. This is reflected in the size of the structures in the test sets; for example, a widely used library Bralibase [5] contains sequences having only few tens of bases. However, the sizes of ribosomal RNA subunits span from few hundreds to few thousands bases. Moreover, the methods mentioned above are meant to compute conservation of whole structures and their adjustment to annotating the conservation of individual parts of a structure is far from obvious. Annotation on the level of individual bases enables to visually express the conservation as shown in the experimental section. Another possibility, apart from adjusting a whole structure conservation approach, is to utilize some of the methods for prediction of consensus secondary structure, such as LocaRNA [6] or Infernal [7], and transfer that information back to the structures from which the consensus structure was built, but it is again not obvious how to do that. Both of the approaches are tedious and therefore we introduce here a fast, straightforward method and its freely available implementation which enables conservation annotation of a set of RNA sequences with conservancy levels based on the available secondary structure information.

The method first converts the input secondary structures into dot-parenthesis representation which is then forwarded into multiple sequence alignment (MSA). The resulting secondary structure-based MSA is subsequently passed into a conservancy identification procedure which uses sliding window technique to identify conserved positions in the MSA and assign them a score based on the secondary structure content of the window. The following section details the procedure and discusses its implementation.

2. Secondary Structure Conservation Annotation Assessment

To be able to conduct the multiple sequence alignment, the two-dimensional secondary structures need to be converted into one-dimensional sequences. In order to do so, we utilize the so-called dot-bracket notation. In this representation, each character represents a base. Matching parenthesis at positions *i* and *j* indicate a base-pair while dot is used to represent an unpaired base. Fig. 1 shows an example of an RNA sequence, its secondary structure in dot-bracket notation and corresponding visual representation.

Obviously, the dot-bracket representation of secondary structure is a sequence and thus one can easily use sequence alignment to compare the structures. The advantage of such an approach is that the problem of sequence alignment is well established in the bioinformatics community, many tools for this task exist and basically any existing multiple sequence alignment (MSA) tool can be used for our purpose [8].

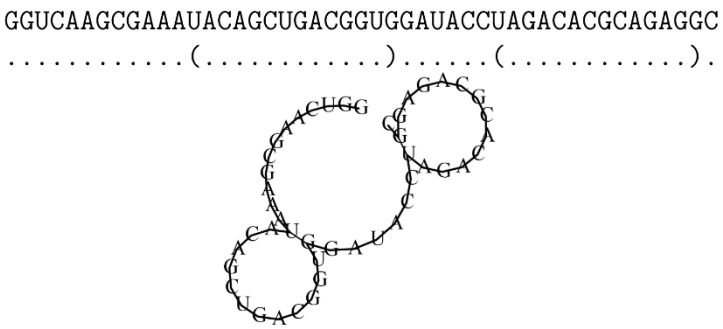

Fig. 1. Example of seconday structure in dot-bracket notation.

The key idea of MSA is to align sequences by inserting gaps (denoted by the "-" symbol) into the sequences in order to make the aligned sequences as similar as possible in the gap-free regions. Note that which alignment is chosen as the best depends on the scoring function used. For pairwise sequence alignment, the optimal alignment for a given cost function can be found in $O(n^2)$ time using dynamic programming [9]. However, this problem is intractable for an arbitrary number of input sequences and therefore heuristic algorithms are commonly used [10]. Our implementation uses the well-known multiple sequence alignment utility Clustalw2 [11], [12]; our method simply takes the secondary structures of the input molecules in dot-bracket notation and passes them to the Clustalw2 tool. The output of this tool is the optimal multiple sequence alignment expressed as a matrix with aligned input positions.

Now, we need to calculate conservancy level for each individual position. We define conservancy level as the percentage of similarity at given position and its close neighborhood. This is done with the help of the sliding window method (as illustrated in Fig. 2). Sliding window is a general technique for iterating over a list of elements where instead of looking at one element at a time we include also its neighborhood – the window – into consideration. The window is slid one position at a time and the number of elements processed at each step is specified by a parameter called *window size*. The conservancy at position i is based on the ratio of common characters at each position in the window centered at position i.

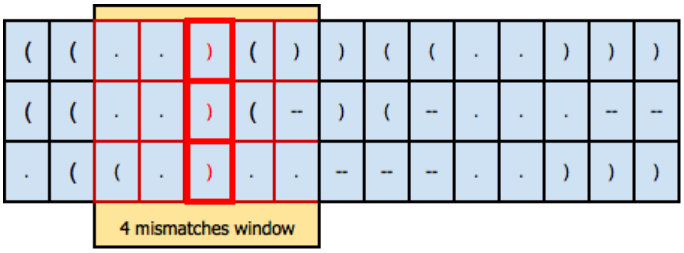

Fig. 2. Illustration of the sliding window method.

An important question is how to weigh and calculate the matches and mismatches in the MSA. For this purpose, the *consensus character* is defined at each position. The consensus character is the most common character at a given position, with the exception of gap which can never be a consensus character. The mismatch ratio at the i-th position is then the number of input structures not containing the consensus character at given position, and the conservancy measure is the aggregation of mismatch ratios over the sliding window.

More formally, let A be the MSA matrix of length n containing s structures. Let w be the sliding window size and C_i consensus character at the i-th position of the alignment. First, we define an auxiliary function f' as:

$$f'(i) = \frac{s - \sum_{j=0}^{s} e\,(i,j)}{s}$$

where $e(i,j)$ is equal to 1 if j-th structure contains C_i at i-th position, 0 otherwise. The conservancy function for i-th position is then defined as follows:

$$m(i) = \frac{\sum_{j=i-\left\lfloor \frac{w}{2} \right\rfloor}^{j+\left\lceil \frac{w}{2} \right\rceil} f'(j)}{w+1}$$

Having a conservancy function and an MSA based on the secondary structures enables us to slide a window along the MSA and assign conservancy to each position. Finally, we project conservancy back on the input sequences, which gives us conservancy level for each nucleotide of every input molecule.

The advantage of the introduced approach lies in its simplicity and ability to quickly produce levels of conservancy for each position for even the largest RNA structures. However, the method also brings several disadvantages which a potential user should be aware of. Due to the use of the sliding window, the method is not capable of precise determination of the conserved regions borders (the borders are rather fuzzy). Also, the quality of the conservation assignment is clearly heavily dependent on the quality of the MSA. Moreover, the method is not very suitable for small molecules since the window should have sufficient size to take a reasonable neighborhood into account; the bigger the neighborhood, the more is the method oblivious to small local changes in structure and is able to correctly cover common local structural motifs such as hairpins. However, it is not possible to use a big window with small structures, where a window of size 30 covers half or third of the structure. For example, in our experiments carried out over ribosomal RNA molecules, the window size was empirically set to 40.

3. Implementation

The algorithm is implemented in Python using the Biopython library [13]. The main class of the algorithm implementation is called *StructureComparator* which handles the conservation measurement process. The input for this class is the list of sequences and their corresponding secondary structures in FASTA format and the result is a list of SVG files with visualizations including conservancy levels.

The class first calls the Clustalw2 utility to produce alignment of the structures. Then, it executes the *ComparationTagger* on the alignment which calculates the conservation level for each column of the alignment (and therefore each base of every input structure). In the next step, RNAplot [14] is run for all of the input structures. RNAplot is a tool from the ViennaRNA package that produces a visualization of a structure in a dot-bracket notation. *StructureComparator* runs this tool with the input FASTA files, resulting

in an SVG visualization of the secondary structures. Finally, the SVG files are modified in order to add the conservancy level of each base. This is realized using an attribute of the XML node representing the base color. The freely available implementation can be downloaded from https://github.com/siret/rna-ss-conservation-annotator.

4. Evaluation

Our algorithm is aimed at annotation of bigger molecules and its comparison with other tools is thus difficult since existing evaluations are focused on smaller RNA structures. Therefore, we decided to evaluate our tool in two tests: (i) we visually inspected annotation of small ribosomal subunits from three organisms and (ii) we tested how well the conservancy annotation correlates with evolutionary distance. All the sequences were downloaded from the SILVA database [15] (release 119) and their structures were predicted using the cppredict algorithm [2].

Fig. 3. Conservation annotation of Homo sapiens (AC139250 – upper left), Western lowland gorilla (CABD02136541 – upper right) and *Rubrobacter* sp. LYG58 (JQ087461 – bottom).

To visually validate the conservation annotation of our tool and to show how to utilize conservation annotation at the individual nucleotide level for visualization, we selected small ribosomal subunits from three organisms: Homo sapiens (AC139250), Western lowland gorilla (CABD02136541) and *Rubrobacter* sp. LYG58 (JQ087461). The structures of the organisms were used as input, their sequences annotated, converted into SVG and the resulting images are displayed in Fig. 3. Positions with high level of conservancy are shown in green while unconserved positions are shown in red. We can see that the structure of Homo sapiens (upper left) differs quite substantially from the other two structures, which can be the consequence of large regions where structures are not resolved well (the big regions of unpaired bases). We also highlighted two motifs which correspond to two regions in the MSA where the conservation maps quite well in the visual layout of the secondary structures. Obviously, if we input several very distant molecules, the result will not be very representative – as can be already seen in the example, where the mapped regions are conserved in terms of the base pairing, but some of the secondary structure motifs do not correspond very well. Also different values of the sliding window can influence quite significantly the sensitivity of the method.

Next, we validated the algorithm outputs with a test based on the premise that structures that are close in the phylogenetic tree should be annotated as relatively conserved. We built ten sets, each consisting of four different structures with various intra-set phylogenetic distances. Then, we annotated conservation for each of these sets. Conservation of the set was measured as the average conservancy of all positions in the corresponding MSA. The expected result was that the conservancy should be lower as the evolutionary distance gets larger.

The test started on comparison of four different Homo sapiens records of large rRNA subunits and ended with comparison of four records of extremely distant organisms. In order for the sets to have a fixed point and not to select each set completely randomly each time, one sequence was selected to be present in every set (Homo sapiens record with accession number HI519109). This fixed sequence can thus be viewed as an indicator of how conserved are the other sequences relatively to this one. To asses the evolutionary distance within each set, we computed the distance from the HI519109 to the lowest common ancestor of the remaining three records. The lowest common ancestor (LCA) for a pair of nodes is defined as the lowest (i.e. deepest) node in the phylogenetic tree that has both of these nodes as descendants. This method of computing evolutionary distance for a group requires that all chosen sequences have the same LCA. This means that the LCA has to have at least four children and each of the molecules has to exist in different subtree rooted in the LCA. When this requirement cannot be fulfilled (e.g. the root of the phylogenetic tree has only three categories: Archaea, Bacteria and Eukaryota), two of the molecules which come from the same child branch need to split at the next level. Ten sets were chosen randomly using this scheme (record HI519109 is present in each of them):

- Distance 1: AC148612 (Homo sapiens), JB867405 (Homo sapiens), GN334774 (Homo sapiens)
- Distance 6: AAPY01089594 (Tupaia belangeri (northern tree shrew)), AAGW02077981 (Oryctolagus cuniculus (rabbit)), AAYZ01272771 (Ochotona princeps (American pika))
- Distance 6: AAGW02082887 (Oryctolagus cuniculus (rabbit)), AFSB01135779 (Heterocephalus glaber (naked mole-rat)), V01270 (Rattus norvegicus (Norway rat))
- Distance 9: AB099628 (Pelophylax nigromaculatus (dark-spotted frog)), AAWZ02036944 (Anolis carolinensis (green anole)), AY859626 (Chrysemys sp. JM-2004)
- Distance 10: AF061798 (Petromyzon marinus (sea lamprey)), AF278683 (Okamejei schmidti (browneye skate)), AY049812 (Callorhinchus milii (elephant shark))
- Distance 10: AEFG01063065 (Petromyzon marinus (sea lamprey)), AY859641 (Narcine brasiliensis), AY049812 (Callorhinchus milii (elephant shark))
- Distance 13: AY026379 (Acanthascus dawsoni (boot sponge)), AY026378 (Pleurobrachia bachei), HQ856867 (Paradrepanophorus crassus)
- Distance 14: DQ343684 (Acrosymphyton caribaeum), GU001164 (Breviata anathema), AJ238659 (Leptolegnia caudata)
- Distance 15: AICP01000064 (Streptococcus anginosus subsp. whileyi CCUG 39159), CAGS01000539 (Nitrolancea hollandica Lb), AP011784 (uncultured crenarchaeote)
- Distance 15: BD250982 (Cenarchaeum symbiosum), ADLS01000036 (Collinsella tanakaei YIT 12063), ASPK01000003 (Thaumarchaeota archaeon SCGC AB-179-E04)

The results of this test are shown by the graph in Fig. 4. We can see that as evolutionary distance among molecules in the sets grows, the average conservancy as measured by our algorithm does indeed decrease. The decrease is not steady, but it can happen that some of the incorrectly aligned regions can show higher conservancy simply by chance. The trend, which can be seen in this graph, holds in general. A set of 10 other test sets with additional details can be found at https://github.com/siret/rna-ss-conservation-annotator/evolution-conservation.zip.

5. Conclusion

We have introduced a new method for conservation annotation of a set of RNA molecules using their secondary structures. The simplicity of the method which is based on MSA of secondary structure and a sliding window procedure makes it suitable for use with even the largest RNA molecules such as ribosomal RNAs. The method is available as a software tool implemented in Python built over the Biopython library

and using Clustalw2 for building MSA. The tool returns information about conservation level for every nucleotide of the input structures with a visualization of each structure where the nucleotides are colored based on the computed conservancy level. However, the visualization is clearly one of the weakest parts of the tool since when the structures are not similar enough, the inspection of aligned parts becomes very difficult. Therefore, in the future we would like to focus on development of visualization which would allow simpler investigation of the conservation results.

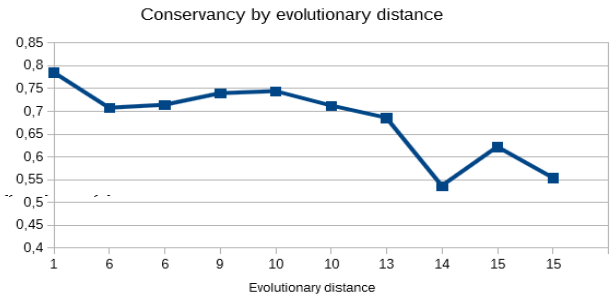

Fig. 1. Visualization of average conservancy measured in the sets of 4 different sequences with different evolutionary distance.

Acknowledgment

This work has been supported by grant no. 550214 of the Charles University Grant Agency and by Charles University projects P46 and SVV-2015-260222.

References

[1] Charpentier, E., & Hess, W. R. (2015). Editorial: RNA in bacteria: biogenesis, regulatory mechanisms and functions. *FEMS Microbiol Rev*, *39(3)*, 277-279.

[2] Pánek, J., Hajic, J., & Hoksza, D. (2014). Template-based prediction of ribosomal RNA secondary structure, in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, IEEE (pp. 18-20).

[3] Held, C. (2000). Phylogeny and biogeography of serolid isopods (Crustacea, Isopoda, Serolidae) and the use of ribosomal expansion segments in molecular systematics. *Mol Phylogenet Evol*, *15(2)*, 165-178.

[4] Gruber, A. R., *et al*. (2008). Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, *9*, 122.

[5] Gardner, P. P., Wilm, A., & Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, *33(8)*, 2433-2439.

[6] Will, S., *et al*. (2012). LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA*, *18(5)*, 900-914.

[7] Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29(22)*, 2933-2935.

[8] Edgar, R. C., & Batzoglou, S. (2006). Multiple sequence alignment. *Curr Opin Struct Biol*, *16(3)*, 368-373.

[9] Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, *48(3)*, 443-453.

[10] Lipman, D. J., Altschul, S. F., & Kececioglu, J. D. (1989). A tool for multiple sequence alignment. *Proc Natl Acad Sci U S A*, *86(12)*, 4412-4415.

[11] Goujon, M., *et al*. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res*, *38*(Web Server issue), 695-699.
[12] Larkin, M. A., *et al*. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23(21)*, 2947-2948.
[13] Cock, P. J., *et al*. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25(11)*, 1422-1423.
[14] Lorenz, R., *et al*. (2011). ViennaRNA Package 2.0. *Algorithms Mol Biol*, *6*, 26.
[15] Quast, C., *et al*. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res*, *41*(Database issue), 590-596.

Jan Pešek obtained his M.Sc. degree from the Department of Software Engineering at the Faculty of Mathematics and Physics, Charles University in Prague. His research focus is on the RNA structure bioinformatics and on both academic and commercial development of web-based solutions.

David Hoksza received the PhD degree in 2010 from the Department of Software Engineering, Charles University in Prague, Prague, Czech Republic. Since 2011, he has been an associate professor of software engineering in the Department of Software Engineering at the Charles University in Prague. And between years 2011 and 2015, he was an assistant professor in the Laboratory of Informatics and Chemistry at the Institute of Chemical Technology, Prague. His current research interests include structural bioinformatics, chemical informatics, data engineering and similarity searching.