Structurexplor: a platform for the exploration of structural features of RNA secondary structures

Jean-Pierre Séhi Glouzon, Jean-Pierre Perreault and Shengrui Wang

Bioinformatics, 33(19), 2017, 3117–3120
doi: 10.1093/bioinformatics/btx323
Advance Access Publication Date: 29 May 2017
Applications Note

Structural bioinformatics

Structurexplor: a platform for the exploration of structural features of RNA secondary structures

Jean-Pierre Séhi Glouzon[1,2], Jean-Pierre Perreault[2] and Shengrui Wang[1,*]

[1]Department of Computer Science, Faculty of Science, Université de Sherbrooke, Sherbrooke, QC, J1K 2R1 Canada and [2]RNA Group, Department of Biochemistry, Faculty of Medicine and Health Sciences, Applied Cancer Research Pavilion, Université de Sherbrooke, Sherbrooke, QC, J1K 2R1, Canada

*To whom correspondence should be addressed.

Associate Editor: Cenk Sahinalp

Received on September 30, 2016; revised on April 20, 2017; editorial decision on May 12, 2017; accepted on May 26, 2017

Abstract

Summary: Discovering function-related structural features, such as the cloverleaf shape of transfer RNA secondary structures, is essential to understand RNA function. With this aim, we have developed a platform, named Structurexplor, to facilitate the exploration of structural features in populations of RNA secondary structures. It has been designed and developed to help biologists interactively search for, evaluate and select interesting structural features that can potentially explain RNA functions.

Availability and implementation: Structurxplor is a web application available at http://structurexplor.dinf.usherbrooke.ca. The source code can be found at http://jpsglouzon.github.io/structurexplor/.

Contact: shengrui.wang@usherbrooke.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Because structure largely determines RNA function (Wan *et al.*, 2011), exploration of structural information in a population of RNA secondary structures to discover features related to specific functions of RNA is essential. A typical example of a function-related structural feature is the cloverleaf shape of transfer RNA secondary structures, which is known to play a crucial role in the translation mechanism.

Exploration of structural features is an iterative process in which a structural biologist, the key actor of the exploration process, searches for, evaluates and selects structural features potentially related to RNA functions, which will then be experimentally validated (Holzinger *et al.*, 2014; Shneiderman, 2002). Specifically, the exploration process involves searching for and evaluating structural features by first preprocessing, comparing and clustering the structures, and then computing and generating visualizations of structural features to facilitate their interpretation. Interesting structural features can then be selected for further experimental validation.

Exploring structural features is a complex and time-consuming task, especially for those who are not computer science specialists. In fact, a costly investment of effort and time is required to gain the necessary advanced knowledge of languages such as Bash, Python or R, and data analytics such as supervised or unsupervised learning methods (Holzinger *et al.*, 2014). Often, one has to use various tools in combination in a pipeline to enable effective exploration of structural features. Most existing tools were designed and developed to solve one particular challenge related to a specific task in the exploration process. For instance, the challenge of finding clusters of similar structures was addressed by Sfold (Chan *et al.*, 2005), GraphClust (Heyne *et al.*, 2012), NoFold (Middleton and Kim, 2014), while the challenge related to visual inspection and manual edition of structures was tackled by tools such as VARNA (Darty *et al.*, 2009), Forna (Kerpedjiev *et al.*, 2015) or 4SALE (Wolf *et al.*, 2014). Programs or methods such as RNAdistance (Hofacker *et al.*, 1989), the relaxed base pair measure (Agius *et al.*, 2010), ERA (Zhong and Zhang, 2013), RNAforester (Schirmer and Giegerich,

2013), LocARNA (Will *et al.*, 2007) or the super-n-motifs model (Glouzon *et al.*, 2017) have been designed to compare structures. The existing tools are unable to assist biologists in the multiple phases of the exploration process.

We propose a new platform, named 'Structurexplor', to facilitate the exploration process for a population of RNA secondary structures. The main contributions of this platform are as follows:

- It facilitates the whole exploration process by assisting the expert in preprocessing and comparing structures, and computing various features such as clusters of structures, representative and unusual structures, and many others. These features are useful since they provide insights into the data. For instance, the shape of the representative structures of clusters sheds light on the main structural shapes of a population of RNA.
- It assists in evaluation and interpretation of the computed features by providing interactive visualization functionalities to efficiently inspect those features. For instance, it provides a way to focus on a specific cluster by zooming in and interactively inspecting the shape of the member structures of this cluster.
- It is versatile, offering the capability to explore structural features of secondary structures from both linear and circular RNA, and to take pseudoknots and G-quadruplexes into account, through its use of the super-n-motifs model (Glouzon *et al.*, 2017).

Structurexplor combines a set of tools and models into a unified platform to accelerate the exploration process for RNA secondary structures. The following sections provide a description of the Structurexplor platform and a practical example of its use.

2 Materials and methods

Structurexplor is a web application mainly written in R (R Core Team, 2015) with the Shiny package (Chang *et al.*, 2016) providing a fast and responsive interface. It has been deployed using ShinyProxy (Verbeke and Michielssen, 2016) which is an alternative open-source program for Shiny server (Chang *et al.*, 2016). Structurexplor takes secondary structures in dot-bracket format as input and facilitates the exploration by a series of steps. First, it computes the structural dissimilarities using the *super-n-motifs* model. It clusters the structures according to their dissimilarity by employing one of the various clustering algorithms including unweighted pair group method using arithmetic mean (UPGMA), Ward or complete linkage (Pang-Ning *et al.*, 2006). Then, it computes various structural features such as the representative structure of each cluster. Finally, Structurexplor offers various interactive visualizations that can be used to explore structural features. For instance, it allows interactive visualization of the shape of a representative structure.

Among the structural features output, Structurexplor yields clusters of structures, i.e. groups of structures with similar shapes, indicating either possible functional groups or alternative foldings in the case of exploration of the structure ensemble. Clusters are computed using the stats package (R Core Team, 2015). Cluster quality is also assessed, based on the silhouette coefficient (Pang-Ning *et al.*, 2006) computed using the cluster package (Maechler *et al.*, 2015), which gives information about how compact and well-separated the clusters are. While other indexes such as Calinksi-Habaratz (Caliński and Harabasz, 2007) or Davies-Bouldin (Davies and Bouldin, 1979) can help to assess clustering quality, the silhouette coefficient has the advantage of being easier to interpret. In fact, the silhouette coefficient is bounded between -1 and 1. It offers a clear interpretation from -1 indicating a low quality clustering to 1 a very high quality of clustering. To facilitate linguistic expression of the clustering assessment, we use quality indices of 'Very high', 'High', 'Medium', 'Low' and 'Very low' corresponding to the silhouette coefficient intervals [1,0.7], [0.7, 0.5], [0.5, 0.3], [0.3, 0], [0, -1]. As an example, a very high clustering quality means that the clusters are far apart and members of a same cluster are very close to each other. The silhouette coefficient requires that the number of members in each cluster be at least 3.

Structurexplor provides cluster and structure hierarchies, which are useful for the study of structural phylogeny. Structurexplor also gives information about the most representative and unusual structural shapes of RNA by the identification of representative and unusual structures of clusters. The package assesses the structural variability of clusters, i.e. how structure shapes may vary within a cluster. Finally, it identifies the region that best describes the clusters. Details about representative and unusual structures, structural variability and the region best describing clusters are provided in the Supplementary Material.

Finally, Structurexplor provides an interactive visualization of the hierarchy, a two dimension representation of structures based on the *super-n-motifs* representations of structures (Glouzon *et al.*, 2017), visualization of the structure shapes and general information on structures and clusters in data table format. These are respectively based on phylotree.js (Pond *et al.*, 2015), rCharts (Vaidyanathan, 2013), Forna (Kerpedjiev *et al.*, 2015) and DT (Xie, 2015). Many interactive controls are available to facilitate exploration, such as the capacity to re-root the hierarchy on specific nodes or to focus on a specific cluster by zooming in on the corresponding region of the 2D representation of structures.

3 Example

When no prior information is available about a population of RNA secondary structures, being able to obtain a clear understanding of the most important structural shapes can be very useful. A typical example is provided to show how Structurexplor can facilitate the identification of these shapes. The example, consisting of 179 various-sized structures from transfer RNA (tRNA) (74-77nt), group-II-D1D4-1 (GII) (70-108nt), 5S ribosomal RNA (5S) (117-124nt), Ribonuclease P RNA (RNaseP) (300–330nt), Signal recognition particle RNA (SRP) (317–320nt) and transfer messenger RNA (tmRNA) (362-366nt) families, is available in the 'Prepare' menu. tRNA, 5S, RNaseP, SRP, tmRNA, originated from RNASTRAND database (Andronescu *et al.*, 2008). GII secondary structures coming from RFAM database (Nawrocki *et al.*, 2015) and representing a conserved structural region from the full structure of Group II Intron has been generated using the consensus structures as a constraint for folding, via the script refold.pl and RNAfold (Lorenz *et al.*, 2011). After running the example, Structurexplor switches to the 'Explore' menu, where users can interactively explore various clustering configurations, by changing the clustering algorithm, for instance. The impact of these modifications on the clustering quality can be instantly observed, which is useful for quickly assessing the quality of clustering configurations. After selecting the best clustering configuration, users can then investigate and visualize structures and cluster features.

As mentioned earlier, Structurexplor provides an effective way to visually exploring structural dissimilarity of RNAs. It can generate, in the panel 'Features visualization', a scatter plot in which each

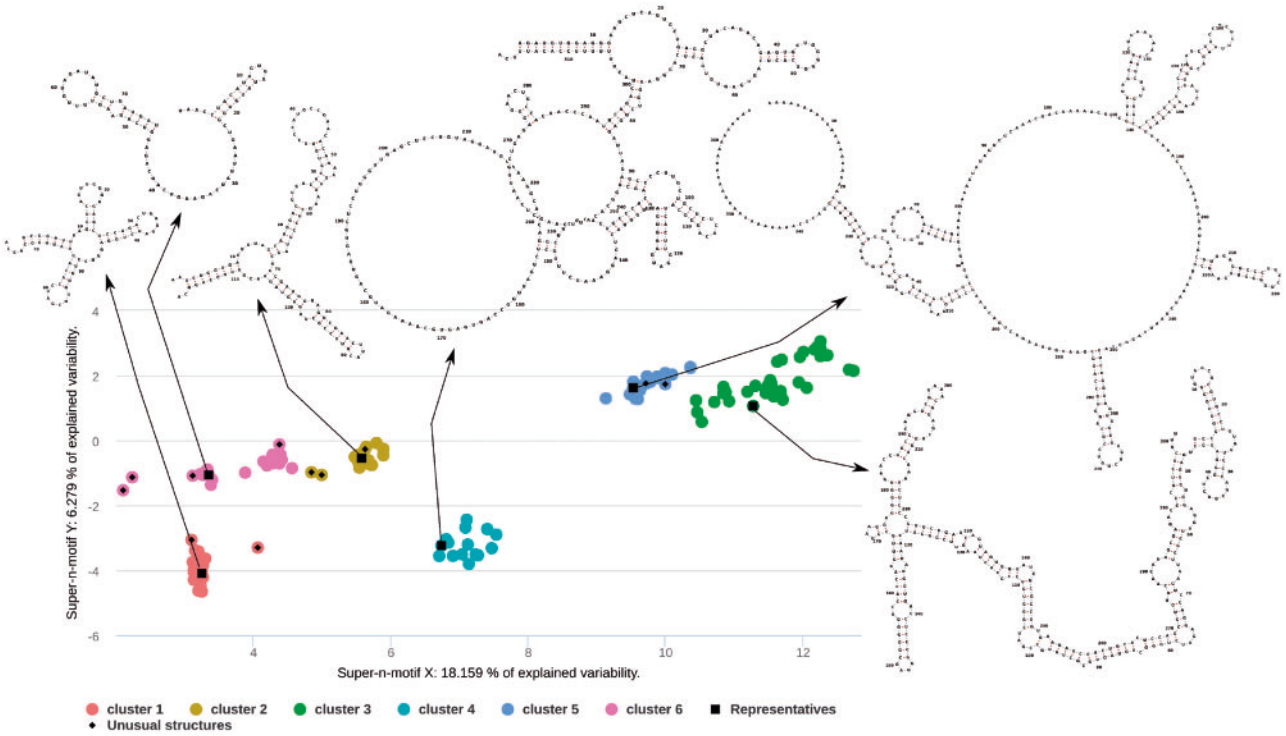

Fig. 1. Clusters of structures and visualization of representative structures for clusters 1, 2, 3, 4, 5 and 6, corresponding to the transfer RNA, Group II Intron, 5S ribosomal RNA, Ribonuclease P RNA, transfer messenger RNA and Signal recognition particle RNA families

structure is displayed as a point on a two-dimensional (sub)space (2D) and allows to visually inspect whether the structures are relatively close, i.e. either similar, or distant (Fig. 1). If two structures on the plot are selected, the structural dissimilarity computed on all the dimensions is also displayed to complement the dimensions-relevant dissimilarity information shown by the relative distance on the plot. Recall that each dimension in the super-n-motifs model represents specific combinations of structural features such as stems or hairpins. Each dimension on the plot is labeled by the amount of associated structural information representing the explained variability, i.e. the strength or the importance of a specific combination of structural features used to represent the structures. To get further information about the computation of the 2D visualization of structures, the structural dissimilarity and the structural information associated with each dimension, readers are referred to the super-n-motifs model (Glouzon *et al.*, 2017).

Structurexplor helps identify clusters. Figure 1 shows six clusters, the shapes of the representative structures of clusters 1 and 2, and the structures with unusual shapes. These features are computed taking into account all the dimensions used to represent the secondary structures i.e. the full set of super-n-motifs, and are reported for visualization in the scatter plot. We can see from the typical shapes of the representative structures that clusters 1, 2, 3, 4, 5 and 6, correspond to the functional families tRNA, GII, 5S, RNaseP, tmRNA and SRP. This example illustrates Structurexplor's suitability for discovering useful information, such as identification of functional groups, from RNA secondary structures.

Funding

This work has been supported by a joint grant from the Fonds de Recherche du Québec – Nature et Technologies (FRQ-NT) and the Université de Sherbrooke Research Chair on RNA Structure and Genomics to Prof. Perreault and the Natural Sciences and Engineering Research Council of Canada to Prof. Wang.

Conflict of Interest: none declared.

References

Agius,P. *et al.* (2010) Comparing RNA secondary structures using a relaxed base-pair score. *RNA*, **16**, 865–878.

Andronescu,M. *et al.* (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340.

Caliński,T. and Harabasz,J. (2007) A dendrite method for cluster analysis. *Commun. Stat. Methods*, **3**, 1–27.

Chan,C.Y. *et al.* (2005) Structure clustering features on the Sfold Web server. *Bioinformatics*, **21**, 3926–3928.

Chang,W. *et al.* (2016) shiny: web application framework for R. *Compr. R Arch. Netw.*

Darty,K. *et al.* (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

Davies,D.L. and Bouldin,D.W. (1979) A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **1**, 224–227.

Glouzon,J.-P.S. *et al.* (2017) The super-n-motifs model: a novel alignment-free approach for representing and comparing RNA secondary structures. *Bioinformatics*, btw773.10.1093/bioinformatics/btw773.

Heyne,S. *et al.* (2012) Graphclust: Alignment-free structural clustering of local RNA secondary structures. *Bioinformatics*, **28**, i224.

Hofacker,I.L. *et al.* (1989) Fast folding and comparison of RNA secondary structures. *Monatshefte Für Chemie Chem. Mon.*, **125**, 167–188.

Holzinger,A. *et al.* (2014) Knowledge Discovery and interactive Data Mining in Bioinformatics–State-of-the-Art, future challenges and research directions. *BMC Bioinformatics*, **15**, I1.

Kerpedjiev,P. *et al.* (2015) Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, **31**, 3377–3379.

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

Maechler,M. *et al.* (2015) cluster: cluster analysis basics and extensions. *Compr. R Arch. Netw.*

Middleton,S.A. and Kim,J. (2014) NoFold: RNA structure clustering without folding or alignment. *RNA*, **20**, 1671–1683.

Nawrocki,E.P. *et al*. (2015) Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res*., **43**, D130–D137.

Pang-Ning,T. *et al*. (2006) Introduction to data mining. *Libr. Congr*., **796**.

Pond,S. *et al*. (2015) phylotree.js: interactive viewer of phylogenetic trees. *GitHub Repos*.

R Core Team. (2015) R: a language and environment for statistical computing. *Compr. R Arch. Netw.*

Schirmer,S. and Giegerich,R. (2013) Forest alignment with affine gaps and anchors, applied in RNA structure comparison. *Theor. Comput. Sci*., **483**, 51–67.

Shneiderman,B. (2002) Inventing discovery tools: combining information visualization with data mining. *Inf. Vis*., **1**, 5–12.

Vaidyanathan,R. (2013) rCharts: interactive charts using javascript visualization libraries. *GitHub Repos*.

Verbeke,T. and Michielssen,F. (2016) ShinyProxy – open source enterprise deployment for shiny. *GitHub Repos*.

Wan,Y. *et al*. (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet*., **12**, 641–655.

Will,S. *et al*. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol*., **3**, 680–691.

Wolf,M. *et al*. (2014) ITS2, 18S, 16S or any other RNA – simply aligning sequences and their individual secondary structures simultaneously by an automatic approach. *Gene*, **546**, 145–149.

Xie,Y. (2015) DT: a wrapper of the javascript library 'DataTables'. *Compr. R. Arch. Netw*.

Zhong,C. and Zhang,S. (2013) Efficient alignment of RNA secondary structures using sparse dynamic programming. *BMC Bioinformatics*, **14**, 269.