Biomedical Data Mining

Extended Abstract

A pipeline for ncRNA sequence reconstruction and structure characterization of potential homologs from BLAST output

Schwarz M and Panek J

Abstract

The BLAST calculation is utilized by numerous investigates as an exploratory RNA arrangement search apparatus. It is amazingly valuable, yet its yield incorporates essentially grouping data just, which isn't adequate for portrayal of arrangement sections. Subsequently we have fostered a pipeline to recognize total groupings of the parts, foresee auxiliary designs of the subject arrangements and gather their homology to the question RNA. The pipeline incorporates a few phases: 1) reconstitution of BLAST hits with secured Locarna calculation, 2) surmising of homology to the question RNA with RSEARCH calculation, 3) forecast of an auxiliary design with Centroid-homfold calculation. Our pipeline can be utilized for portrayal of ncRNAs overall by broadening data remembered for the BLAST yield. Additionally, it tends to be very helpful when homologs of uncharacterized, for example recently recognized ncRNAs should be found and for which more complex strategies for homology search can't be utilized as they require more data of the RNA in their information that isn't accessible.

Searching for similar sequences in a database via BLAST or a similar tool is one of the most common bioinformatics tasks applied in general, and to non-coding RNAs. However, the results of the search might be difficult to interpret due to the presence of partial matches to the database subject sequences. Here, we present rboAnalyzer – a tool that helps with interpreting sequence search result by (1) extending partial matches into plausible full-length subject sequences, (2) predicting homology of RNAs represented by full-length subject sequences to the query RNA, (3) pooling information across homologous RNAs found in the search results and public databases such as Rfam to predict more reliable secondary structures for all matches, and (4) contextualizing the matches by providing the prediction results and other relevant information in a rich graphical output. Using predicted full-length matches improves secondary structure prediction and makes rboAnalyzer robust with regards to identification of homology. The output of the tool should help the user to reliably characterize non-coding RNAs in BLAST output. The usefulness of the rboAnalyzer and its ability to correctly extend partial matches to full-length is demonstrated on known homologous RNAs. To allow the user to use custom databases and search options, rboAnalyzer accepts any search results as a text file in the BLAST format. The main output is an interactive HTML page displaying the computed characteristics and other context of the matches. The output can also be exported in an appropriate sequence and/or secondary structure formats.

Keywords: RNA, sequence, database, search, secondary structure, RNA homology

INTRODUCTION

The output of a BLAST (Camacho et al., 2009) search is a list of hits of the query sequence in the search database that are called high-scoring pairs (HSPs). They are characterized by their statistically estimated quality and position within the sequences in the search database. A HSP contains the sequences of the matched RNA and the query RNA that are similar to each other. These sequences can be either full sequences or fragments of the full sequences, so called partial matches.

Schwarz M and Panek J
Czech Academy of Sciences, Czech Republic, E-mail: marek.schwarz@biomed.cas.cz

Since it is frequently impossible to reliably determine secondary structure, homology and function from a fragment of a non-coding RNA, the interpretation of results of a sequence search for non-coding RNAs requires full-length sequences of the matched RNAs. The full-length sequences of the partial matches are usually identified manually using external bioinformatics tools for individual RNAs which can be laborious and inefficient. The aim of the presented tool, rboAnalyzer, is to replace the manual work by an automated workflow and thereby to make the interpretation of the database sequence search results easier.

The rboAnalyzer pipeline extends partial or otherwise imperfect matches to the length of the query sequence and computes its secondary structure and a homology to the query RNA. These tasks are handled with available bioinformatics algorithms integrated into a framework that combines the information contained in the BLAST output with external sources such as Rfam (Nawrocki et al., 2015). rboAnalyzer runs from command line. Its input consist of a BLAST output text file, a FASTA file with the query RNA sequence and the database used in the search. The output is a HTML page that integrates the computed characteristics of the subject RNAs together with the subject RNAs data. Results are presented in a clear, interactive and exportable form.

While the presented version of rboAnalyzer takes BLAST results as an input, the algorithm is general and can be easily extended to accept matches obtained with other database sequence search tools.

METHODS

To evaluate the performance of the methods for extension of partial matches we prepared a dataset with known RNA sequences located at known positions in the database sequence. The database sequence was constructed artificially using sequences of RNAs families in CompaRNA dataset (Puton et al., 2013). CompaRNA contains those Rfam families that have at least one homolog with experimentally identified structure. Of the families in CompaRNA, we used only those whose Rfam seed alignments included at least 20 homologs. The homolog with experimentally identified structure was used as a template for evaluation of the accuracy of our secondary structure prediction.

The above-mentioned criteria were fulfilled by the following RNA families: RF00001, RF00002, RF00005, RF00008, RF00015, RF00017, RF00020, RF00095, RF00100, RF00162, RF00167, RF00169, RF00175, RF00209, RF00230, RF00250, RF00374, RF00379, RF00380, RF00480, RF01051, RF01725, RF01739, RF01807, RF01831, RF01852, RF02095, RF02253, and RF02348. For each family, three RNA sequences were chosen randomly and set aside to be used as query sequences. The remaining sequences were used to construct an artificial subject sequence, in which they were placed one after another, separated from each other by their 1000 nucleotides long 5′ and 3′ flanking regions, forming a long single sequence. The sequences of the flanking regions were obtained from NCBI using Rfam accession numbers of appropriate RNAs. When flanking regions with 1000 nucleotides were not available, a random sequence was used to fill the missing section.

Furthermore, to create decoys in the artificial subject sequence, the same RNA sequences were shuffled 10 times each, and together with their flanking regions included into the artificial subject sequence in the same way as for the original RNAs. For this artificial subject sequence, a BLAST database was build using makeblastdb program (Camacho et al., 2009).

Identification of homology of subject RNAs

In this step, the homology of the subject RNAs to the query RNA was identified using sequences of extended matches of the subject RNAs. First, a covariance model of the query RNA was computed with RSEARCH, as described for the “meta” extension method in Step 1. Then, rboAnalyzer scored each of the extended matches by comparing it to the covariance model using cmalign (Nawrocki and Eddy, 2013). The score is a measure of the homology of the subject RNAs to the query RNA in terms of similarity of their sequences represented by their extended matches and their potential secondary structure.

Prediction of secondary structure of subject RNAs

The subject RNAs are further characterized by a secondary structure predicted using their extended matches.

rboAnalyzer offers 15 methods for the secondary structure prediction implemented using available algorithms and their combinations in order to efficiently exploit the information in the BLAST output.

The sequences are used in two ways depending on the prediction methods:

(a) to build a reference consensus secondary structure using RNAalifold (Bernhart et al., 2008) with the multiple sequence alignment made by Clustal Omega (Sievers et al., 2011) or muscle (Edgar, 2004), followed by refold.pl (Tafer et al., 2011)/RNAfold –C (Lorenz et al., 2016), or UNAfold (Markham and Zuker, 2008);

(b) to serve as reference sequences by the methods based on TurboFold (Tan et al., 2017) or CentroidHomfold (Hamada et al., 2009).

The methods belonging to the category 2 include RNAfold, which is a de novo prediction method, then prediction methods that use covariance models identified in Rfam, and finally, a prediction method that uses RNAfold to predict secondary structure of the query RNA, which is then used as a structural template for finding a best matching structure among suboptimal structures of subject RNA predicted by UNAfold.

RESULT

Default Values of rboAnalyzer Parameters

The default set-up for rboAnalyzer includes "locarna" method for extension of partial matches and three methods for the secondary structure prediction, RNAfold, TurboFold, and rfam-Rc (which is a shortcut for RNAfold –C with a Rfam consensus structure as constraint). TurboFold was chosen as it performed best of all the prediction methods. RNAfold was chosen as a standard with minimum input providing an output under any conditions. RNAfold –C with a Rfam consensus structure as constraint was chosen as a representative of the methods using information from an external source. The three selected methods were chosen as they are based on different prediction principles and guarantee that the user gets most accurate prediction available.

rboAnalyzer has numerous parameters that belong to the algorithms used for its construction. Their default values were optimized using RNAs with experimentally identified secondary structures. The list of the parameters and the algorithms used in rboAnalyzer and the details about the parameter optimization are described in the Supplementary Material "Optimization of rboAnalyzer parameters."

Test of rboAnalyzer Performance

Here, the performance of rboAnalyzer with respect to the quality of HSPs in a BLAST output was tested.

We first created a synthetic sequence database from the sequences of the RNA families selected using the criteria described in the "Evaluating partial matches extension methods" section above.

For each RNA family we downloaded sequences from CompaRNA dataset which were then used as queries for BLAST. Then we extracted sequences from Rfam seed alignment for each RNA family that were used as known homologs. For each of them we downloaded its parent sequence from NCBI that were placed between sequences of its randomly shuffled 500 nucleotide long 5′ and 3′ flanking regions. These sequences were used to construct a BLAST database, in which the query RNAs with experimentally identified structures were searched with blastn program (parameters: -gapopen 2 -gapextend 1 -penalty -1 -reward 1 -word_size 7) obtaining BLAST outputs. For the families that included more than one RNA with experimentally identified secondary structure, more than one BLAST output were obtained and of them the one with the largest number of matches to the sequences from Rfam seed alignments was used for further analysis.

Matches in each BLAST output were sorted into three groups according to their quality. The quality was defined as high, moderate and low and the thresholds were set so that each of the three groups contained equal number of matches to the sequences from Rfam seed

Schwarz M and Panek J
Czech Academy of Sciences, Czech Republic, E-mail: marek.schwarz@biomed.cas.cz

alignments. Finally, the groups were used to form new synthetic BLAST outputs resulting into three synthetic outputs for each of the original BLAST outputs. These synthetic outputs were analyzed by rboAnalyzer to find out how its performance depends on the quality of BLAST outputs.

The performance was measured by structural similarity between the secondary structures of subject RNAs predicted using sequences of their extended partial matches and the experimentally identified structures of the query RNAs. In this test, the subject RNAs and the query RNAs were known to be homologous as they came from the same families, and therefore the predicted secondary structures and experimentally identified structures should be similar if the rboAnalyzer characterizing pipeline was correct and accurate. The higher the similarity, the higher the overall performance of rboAnalyzer, because the secondary structure prediction is the final step of rboAnalyzer and therefore depends on the performance of the previous steps.

The rather flat curves showed steady rboAnalyzer performance regardless of the quality of HSPs in input BLAST outputs. This indicated that rboAnalyzer is robust and capable to produce accurate secondary structures even for short partial matches of subject RNAs.

DISCUSSION AND CONCLUSION

We present rboAnalyzer, a tool for interpreting RNA sequence search outputs. It characterizes the hits in the outputs by prediction of their full-length sequences, homology to the query molecule and secondary structures. The tool is primarily aimed at non-coding RNA molecules but can also be used with other RNAs that have defined structure (e.g., riboswitches).

In our opinion, the tool is needed, because only full-length sequences allow for effective analysis of RNAs in general. The prediction and analysis of secondary structure, homology and function identification is also essential, as RNAs function only with their full-length sequences. The partial and/or gapped matches in the output of sequence search usually come without any other information than their quality. The next step naturally is the identification of full-length sequences of the matched RNAs. So far, up to our knowledge, there is no other choice than to do it manually The presented tools facilitates this task by integrating appropriate tools to one framework with rich user control and results output.

By testing rboAnalyzer with BLAST outputs of varying quality we demonstrated that rboAnalyzer was able to give accurate secondary structure predictions even for HSPs that corresponded to short fragments and with low-quality HSPs.

Since running rboAnalyzer on typical BLAST results takes from several minutes to about an hour, depending on the number of HSPs, the length of the query RNA and chosen prediction methods, it is suitable for analyzing small number of BLAST outputs on a personal workstation, but requires cluster-scale computational resources for larger analyses.

rboAnalyzer is not suitable in situations, when subject RNAs include intronic RNA while the query RNA does not. It is because the difference between the length of the query sequence and the size of the genome locus containing the subject RNA with an intron makes it impossible to correctly extend the partial match of the subject RNA.

Currently, the rboAnalyzer webserver is being developed. Also a minimal version of rboAnalyzer fast enough to be able to analyze individual HSPs in real time is under preparation and will be included in the webserver.

rboAnalyzer is available under GPL 2.0 license.

Biomedical Data Mining

Extended Abstract

This work is partly presented at 9th International Conference and Expo on Proteomics and Molecular Medicine, November 13-15, 2017, Paris, France.

Schwarz M and Panek J
Czech Academy of Sciences, Czech Republic, E-mail: marek.schwarz@biomed.cas.cz