OXFORD UNIVERSITY PRESS | Bioinformatics

APTANI^2: update of aptamer selection through sequence-structure analysis

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2019-1687.R1
Category:	Applications Note
Date Submitted by the Author:	n/a
Complete List of Authors:	Caroli, Jimmy; University of Modena and Reggio Emilia Department of Life Sciences, Department of Life Sciences Forcato, Mattia; University of Modena and Reggio Emilia Department of Life Sciences, Department of Life Sciences Bicciato, Silvio; University of Modena and Reggio Emilia Department of Life Sciences, Department of Life Sciences
Keywords:	Aptamers, Sequence analysis, Bioinformatics

RESPONSE TO REVIEWERS, CAROLI et al. BIOINF-2019-1687

We thank the reviewers for their careful consideration and valuable suggestions. We have revised the paper to fully address all their remarks. We include here the reviewers' comments *ad litteram*, as well as our detailed response to each of them (in blue). The changes introduced in the manuscript are in red in the main text.

Reviewer: 1

Comments to the Author
Caroli et al. present version 2 of their previous algorithmic work, APTANI, a computational pipeline for the selection of target-specific aptamers from HT-SELEX data. APTANI 2 extends the previous approach, consisting of aptamer frequency calculation, secondary structure prediction, and consecutive motif identification, with a mathematical framework for scoring motifs and aptamers according to their structural stability in order to identify aptamers of biological interest.
Furthermore, a series of post-analysis tools to inspect various properties of the data are presented. These aid the visualization of enrichment throughout the sequenced selection cycles, can extract aptamers containing a specific motif of interest from the data, as well as visualize the structural properties of motif bearing species.
The authors showcase the improvements of their new approach on a dataset previously analyzed with APTANI v1 and report significantly faster processing times, as well as more accurate motif reporting.

COMMENTS:

- The current downloadable package lacks the proper license files for the third-party applications included in the archive. The authors should investigate the exact requirements for VARNA and RNASubopt and include the licenses as per the original authors requirements.

  We thank the reviewer for pointing this out. The two third-party software embedded in APTANI^2 (i.e., VARNA and RNASubopt from the ViennaRNA package) are both released under GPL3.0 license. According to the guidelines defined by the original authors for the distribution of their codes, we included the files LICENSE and ViennaRNALicense.txt in the revised version of APTANI^2 package.

- The runtime comparison presented in the result section should state whether both tests were performed on identical hardware, and if wall time or CPU time is specified in the manuscript.

  All analyses and tests were performed on the same computer hardware, without limiting CPU usage or running time. According to the reviewer's suggestion, we specified in the main text that APTANI and APTANI^2 were run on the same computer, whose hardware characteristics are now reported in the section "Comparison with APTANI" of the Supplementary Information.

- It is highly recommended to update the GUI source code to match the python version used by APTANI2. Requiring the installation of two non-compatible python environments on a single system seems highly cumbersome and contradicts the "ease of use" claim the GUI is aimed at providing in the first place.

  We do agree with the reviewer that the installation of two Python environments on a single machine is in contrast with the spirit of this second version of APTANI. To resolve this dichotomy, we first verified that the code of the Graphic Use Interface (GUI) is fully compatible with all Python 3.x versions, once Python 3.x and all its required modules are installed (see the INSTALL guidelines for additional details). Then, we proved, through several tests, that the GUI is fully functional on Python versions 3.0, 3.4 and 3.6.

- Related to the point above, remove the references to the python versions in the IMPLEMENTATION section.

  As suggested, we modified both the "Implementation" section of the manuscript and the INSTALL file of the package specifying that the APTANI^2 package is implemented in Python 3.4.

Reviewer: 2

Comments to the Author
The authors present an updated version of their APTANI software. APTANI2 builds on previous methods used for aptamer identification by not just using the quantification of sequence enrichment or structural motifs but taking into account the stability and reliability of the structures predicted across the SELEX cycles.
They test their software on two datasets:
1) aptamers against IL4Rα, in which APTANI2 took less time to identify the lead motif than the original APTANI software;

2) aptamers against the prostate cancer cell line LN3, in which they identified the only two aptamers characterized as binders, which the original APTANI software was unable to identify.

This is a useful software package for aptamer scientists and the added features, such as visualization of secondary structures and a GUI are useful. The originality is lower as it is an update of pre-existing software, however the methodology is new, and many features are added.

I think it is a great paper, however I suggest that the software is tested on an additional dataset to demonstrate how robust the methodology is.

Specific comments for revision:

a. Major

- it seems that you have only included one new dataset for your analysis. Including at least one other dataset would show that your methodology is robust.

  We thank the reviewer for suggesting us to strengthen the reliability of APTANI^2 using additional data analyses. Given the space constraints of the Application Note format, we stated in the main text that additional analyses are now described in the completely revised section "Comparison with APTANI" of the Supplementary Information.

  Specifically, we analyzed the data sets deposited in the projects PRJNA321551 (Dao et al., Cell Syst. 2016 Jul;3(1):62-70) and PRJNA315881 (Levay et al., Nucleic Acids Res. 2015 Jul 13;43(12):e82) of the Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra).

  The PRJNA321551 data set consists of 10 different Cell-SELEX cycles (cycle 0 and 9 subsequent rounds of selection) performed using an RNA library of aptamers, with a variable region of 30 nucleotides, targeting CCR7. Dao and colleagues used AptaTRACE to select 5 most representative secondary structure motifs (i.e., CTGTG, TTATT, GTTTA, ATGTT and GTGTC). Here, we applied APTANI^2 and APTANI on the data from cycle 9 (using default settings and a frequency threshold of 10^{-8}) to verify if the two methods could identify the secondary structure motifs selected, as the most significant, by AptaTRACE. APTANI^2 ranked the secondary structure motifs of Dao et al. in the top 99th percentile of the MtfScore ranked motif list. Conversely, APTANI did not identified any of the 5 structural motifs.

  The PRJNA321551 data set comprises 5 subsequent rounds of selection against human IL-10RA using an RNA library of aptamers with fixed flanking regions surrounding a variable region of 40 nucleotides. Using an *ad-hoc* HT-SELEX experimental design and sequence clustering, Levay and coworkers identified 33 sequences as potential binders based on high values of prevalence and enrichment rate. In a binding assay, 7 of these sequences displayed high affinity to human IL-10RA (K_d<20 nM) and 4 of them were further proven to be target specific using human IgG and murine IL-10RA as control references. Here, we used APTANI^2 and APTANI on the data from cycle 5 to verify if our tools were able to indicate these 7 aptamers as potential binders. APTANI^2 ranked all 7 sequences in the top 99% of the ranked list, while APTANI could not analyze the 53 GB of cycle 5 data due to its intrinsic limitations in memory usage and computational time.

- you need to discuss any false positive results in more depth.

  In APTANI^2, each aptamer k is ranked according to a score that quantifies the normalized abundance of top scoring motifs in aptamer k given the total number of motifs retrieved in k. In this sense, the software does not select specific aptamers, but simply ranks all sequences (that pass the frequency filter) on a score quantifying the abundance of structural motifs. The final selection of the potentially relevant, high-affinity aptamers (positive results) is left to the user, who can select the sequences to validate considering the aptamer score, the aptamer frequency, or a combination of the two.

  The rationale behind assigning a score to all sequences, rather than selecting a subset of sequences, is to avoid that the software inflates the number of false positives inherent to the HT-SELEX technique. Indeed, in the SELEX protocols, selection pressure is not the only factor driving the enrichment of certain aptamers and the expansion of sequences based on factors other than target binding (e.g., "passenger" sequences due to PCR artifacts) is a common source of false positives (Levay et al., NAR 2015). In the experiment of Speransky et al. (2019), Aptamer 413 (Apt413) can be considered a clear example of a "passenger" false positive sequence generated by some protocol artifacts. Apt413 has an extremely high prevalence in the final aptamer pool (representing almost 60% of all sequences of cycle 11; see Table S1), reflecting a high affinity of this sequence to LN3 cells during the enrichment cycles. A high prevalence value is commonly considered one of the indicators of high affinity binders by most aptamer selection methods. However, despite its high prevalence, Apt413 did not show any positive staining when incubated with LN3 cells (see Online Resource 4 of Speransky et al. (2019)) suggesting that or

Apt413 was a PCR artifact or its binding to LN3 cells depended on the SELEX buffer conditions. Similar findings have been reported and are discussed e.g., by Levay and coworkers (Levay et al., NAR 2015) for aptamers against IL-10 and 4-1BB receptors.

In essence, we definitely agree with the reviewer that a thorough analysis of false positives (and false negatives) would be valuable per se and should account for artifacts generated by both the SELEX protocols and the computational tools. This would require either an extensive experimental validation (involving, at least, hundreds of molecules), ad-hoc experimental designs (as the co-incubation with reference targets proposed by Levay et al. to determine aptamer specificity) or simulated data (able to reproduce all potential artifacts arising from both the experimental protocols and the computational analysis). To our knowledge, these types of data are not currently available and their generation, although being a fascinating challenge, is beyond the scope of this Application Note.

Nonetheless, in accordance with the reviewer's comment and for sake of clarity, we revised the main and the supplementary texts highlighting that APTANI[2] does not select high-affinity aptamers (i.e., true/false positives), but rather ranks all aptamer sequences (that pass the prevalence filter) in terms of frequency and structural stability of sequence motifs.

b. Minor

- page 9 supplementary "APTANI did not associated any motifs" should be "APTANI did not associate any motifs "

We modified the text accordingly.

Application Note

APTANI²: update of aptamer selection through sequence-structure analysis

Jimmy Caroli, Mattia Forcato*, Silvio Bicciato*

Department of Life Sciences, University of Modena and Reggio Emilia – Modena, Italy

*To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Here we present APTANI², an expanded and optimized version of APTANI, a computational tool for selecting target-specific aptamers from HT-SELEX data through sequence-structure analysis. As compared to its original implementation, APTANI² ranks aptamers and identifies relevant structural motifs through the calculation of a score that combines frequency and structural stability of each secondary structure predicted in any aptamer sequence. In addition, APTANI² comprises modules for a deeper investigation of sequence motifs and secondary structures, a graphical user interface that enhances its usability, and coding solutions that improve performances.

Availability and implementation: Source code, documentation, and example command lines can be downloaded from http://aptani.unimore.it. APTANI² is implemented in Python 3.4, released under the GNU GPL3.0 License, and compatible with Linux, Mac OS and the MS Windows subsystem for Linux.

Contact: silvio.bicciato@unimore.it; mattia.forcato@unimore.it

Supplementary information: Supplementary information is available at *Bioinformatics* online.

1 INTRODUCTION

Aptamers are small, single-stranded DNA or RNA nucleotide sequences (<100 residues) that are selected in vitro through the Systematic Evolution of Ligands by Exponential Enrichment (SELEX) procedure (Kim *et al.*, 2011; Sherman and Contreras, 2018). SELEX and HT-SELEX (i.e., SELEX combined with high-throughput sequencing) are iterative processes that enrich DNA and RNA aptamer libraries with sequences characterized by high affinity and selectivity toward a target ligand. The ability of aptamers to recognize and bind their cognate partners depends on unique three-dimensional structures formed by the folding of the aptamer sequences into various secondary structures (as hairpin, intra-strand loop, bulge, and G-quadruplex) (O. Tucker *et al.*, 2012; Zhou and Rossi, 2017). Lead aptamers are commonly selected from SELEX data using computational methods that consider the relative sequence abundance (frequency) at the various cycles of the SELEX enrichment (Kinghorn *et al.*, 2017). Some of these methods contain modules to predict specific secondary structures of sequences in each selection round and to rank aptamers by motifs embedded in their predicted structures (Hoinka *et al.*, 2012; Caroli *et al.*, 2016; Dao *et al.*, 2016; Hoinka *et al.*, 2018). Despite their efficacy in aptamer selection, these computational tools focus only on the enrichment of sequences or structural motifs along the selection process, ignoring any quantification of the stability and reliability of the structures predicted across the SELEX cycles. To overcome this limitation, we developed APTANI², an updated version of the APTANI approach. APTANI² scores each aptamer on both the frequency and the structural stability of its predicted secondary structures. The scoring function is based on the minimum free energy (MFE) calculated for a variety of secondary structures including G-quadruplex, i.e. stable non-canonical secondary structures known to confer high specificity and drug delivery potential to aptamers (Park *et al.*, 2018). In APTANI² we also added a set of post-analysis modules to investigate aptamer evolution along the SELEX process and to efficiently retrieve and visualize sequence motifs and predicted secondary structures. Finally, APTANI² is accompanied by a user-friendly graphic interface (GUI) that enhances and expands the tool usability.

2 IMPLEMENTATION

APTANI² builds on the evidence that sequence motifs with a putative binding potential are i) enriched at a given SELEX cycle and ii) have an intrinsic high stability across the secondary structures of different aptamers. Therefore, in APTANI² aptamers are ranked using a scoring function that accounts for both the frequency and the stability of each motif retrieved in any aptamer sequence.

The core of APTANI² consists of three major steps: i) calculation of aptamer frequency; ii) identification of motifs associated to secondary structures; and iii) scoring of sequence motifs and aptamers. Starting from an input file in FASTQ format, the first step calculates the relative frequency of each aptamer sequence produced by the SELEX process, as described in the original APTANI (Caroli *et al.*, 2016). In the second step, aptamers that pass the frequency filter are processed with RNASubopt (Siebert and Backofen, 2007) to predict all possible secondary structures within a range of 3 Kcal/mol above the minimum free energy (MFE). This step has been entirely re-coded and optimized to use RNASubopt at an MFE threshold larger than 1 Kcal/mol (as required by the original APTANI) and to explore the complete pool of secondary structures. Each predicted secondary structure is then investigated to retrieve all the associated hairpin, intra-strand, bulge (left and right), and G-quadruplex sequence motifs. In the third step, first each motif is characterized in terms of frequency and structural stability of the related secondary structure through a motif score (*MtfScore*) calculated

as $MtfScore_i = f_i^{(1 - CRS_i)}$, where f_i and CRS_i are the frequency and the Coefficient of Relative Stability (CRS) of motif i, respectively. For each motif i, the Coefficient of Relative Stability (CRS) is defined as $CRS_i = \overline{E}_i/MFE_{min}$, i.e., the ratio between $\overline{E}_i$, the median energy value of all secondary structures containing motif i, and MFE_{min}, the lowest Minimum Free Energy of all predicted secondary structures. Then, the distribution of the *MtfScore* values is used to define the top scoring motifs as those motifs with an *MtfScore* greater than e.g., the 99th percentile of the *MtfScore* distribution. Finally, each aptamer k is ranked according to an aptamer score $AptScore_k = \#_k^{top\ scoring}/\#_k^{all}$ that quantifies the normalized abundance of top scoring motifs in sequence k ($\#_k^{top\ scoring}$) given the total number of motifs retrieved in k ($\#_k^{all}$).

Post analysis modules

APTANI^2 comprises three additional post-analysis modules to inspect the evolution of aptamer enrichment along the SELEX process (*Evolution Analyzer*), to retrieve motifs in aptamer sequences (*Motif Fetcher*), and to visualize their predicted folding and motif structure (*grAPhTANI*). In particular, *Evolution Analyzer* allows monitoring the evolutionary pressure on both aptamers and motifs through consecutive enrichment cycles. Given a specific aptamer (or a set of aptamers), *Motif Fetcher* retrieves all motifs contained in the input sequence and their associated motif scores. Finally, *grAPhTANI* exploits the drawing power of VARNA (Darty *et al.*, 2009) to generate a graphical representation of secondary structures and motifs predicted in a given aptamer.

Graphic User Interface (GUI)

To expand the tool usability also to non-expert users, APTANI^2 is available through a user-friendly graphic user interface (GUI). The graphic interface allows to execute both core analysis and post-analysis modules, simplifying the input selection and parameter setting through pre-compiled fields in a point-and-click environment.

The APTANI^2 package is implemented in Python 3.4. Further details on software implementation, installation, and usage are available in Supplementary Information.

3 RESULTS

To test APTANI^2 performances in indicating potentially high-affinity aptamers, we first re-analyzed the dataset of (Roth *et al.*, 2012) used with APTANI in (Caroli *et al.*, 2016). APTANI^2 correctly indicated the previously validated aptamer Cl.42 and intra-strand motif GGAAAAA||UCCAUGC as one of the lead sequences and motifs (top 0.07% of all ranked aptamers based on the *AptScore* and top 0.2% of all retrieved intra-strand motifs based on the *MtfScore*, respectively). On this dataset, APTANI^2 took 15 minutes to explore and analyze the entire pool of 4,194,000 secondary structures generated by RNASubopt at 3 Kcal/mol, as compared to APTANI that, using the same hardware configuration, required 45 minutes to explore only 268,500 structures, i.e., the maximum subset of all secondary structures constrained by the MFE threshold of 1 Kcal/mol (Caroli *et al.*, 2016). To further verify the efficacy of APTANI^2 to discover motifs in biologically relevant aptamers, we analyzed the most abundant aptamers from a SELEX experiment to identify sequences with differential binding to prostate cancer cell lines (Speransky *et al.*, 2019). APTANI^2 highlighted the presence of top scoring motifs in Apt63 and Apt41, the only two sequences validated in vitro for selective binding to the prostate cancer cell line LN3 (Speransky *et al.*, 2019). In particular, in Apt63 APTANI^2 identified the presence of one of the top five left bulge motifs and of several other left bulges and hairpins ranking in the top 1% of their respective rankings. In Apt41, APTANI^2 discovered three left bulge motifs ranking at the top 0.0005%, 0.001%, and 0.002% of the left bulge ranking, respectively. Conversely, APTANI did not associate any motif to Apt63 and Apt41, thus precluding the identification of these two sequences as potentially relevant.

Additional analyses of Cell-SELEX and HT-SELEX data from (Dao *et al.*, 2016) and (Levay *et al.*, 2015) are described in Supplementary Information.

4 CONCLUSIONS

Here we present APTANI^2, an evolution of our original tool to select target-specific aptamers through the combined analysis of sequence and structure data from SELEX and HT-SELEX experiments. APTANI^2 adopts an innovative strategy to rank aptamer sequences in terms of frequency and structural stability of sequence motifs. To our knowledge, APTANI^2 is the first tool that characterize aptamers based on the structural stability of the predicted secondary structures, thus expanding the former approach proposed by (Hoinka *et al.*, 2012) and adopted in APTANI (Caroli *et al.*, 2016), where the pool of structural motifs is reduced through alignment and definition of a consensus sequence. As compared to APTANI, APTANI^2 adopts some innovative coding solutions and post-analysis modules that allow to explore the entire pool of secondary structures with a major improvement in aptamer selection and computational efficiency. In addition to the command line scripts, the tool is easily accessible through a graphical user interface that improves its overall usability.

ACKNOWLEDGEMENTS

Funding: This work was supported by the Italian Ministry of Education, University and Research through EPIGEN – the Italian Flagship Project on Epigenomics.

Conflict of interest: none declared.

REFERENCES

Caroli,J. *et al.* (2016) APTANI: A computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data. *Bioinformatics*, **32**, 161–164.

Dao,P. *et al.* (2016) AptaTRACE Elucidates RNA Sequence-Structure Motifs from Selection Trends in HT-SELEX Experiments. *Cell Syst.*, **3**, 62–70.

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

Hoinka,J. *et al.* (2018) AptaSUITE: A Full-Featured Bioinformatics Framework for the Comprehensive Analysis of Aptamers from HT-SELEX Experiments. *Mol. Ther. - Nucleic Acids*, **11**, 515–517.

Hoinka,J. *et al.* (2012) Identification of sequence-structure RNA binding motifs for SELEX-derived aptamers. *Bioinformatics*, **28**, i215-23.

Kim,Y.H. *et al.* (2011) An RNA aptamer that specifically binds pancreatic adenocarcinoma up-regulated factor inhibits migration and growth of pancreatic cancer cells. *Cancer Lett.*, **313**, 76–83.

Kinghorn,A.B. *et al.* (2017) Aptamer Bioinformatics. *Int. J. Mol. Sci.*, **18**.

Levay,A. *et al.* (2015) Identifying high-affinity aptamer ligands with defined cross-reactivity using high-throughput guided systematic evolution of ligands by exponential enrichment. *Nucleic Acids Res.*, **43**, e82–e82.

O. Tucker,W. *et al.* (2012) G-quadruplex DNA Aptamers and their Ligands: Structure, Function and Application. *Curr. Pharm. Des.*, **18**, 2014–2026.

Park,J.Y. *et al.* (2018) Gemcitabine-Incorporated G-Quadruplex Aptamer for Targeted Drug Delivery into Pancreas Cancer. *Mol. Ther. - Nucleic Acids*, **12**, 543–553.

Roth,F. *et al.* (2012) Aptamer-mediated blockade of IL4Rα triggers apoptosis of MDSCs and limits tumor progression. *Cancer Res.*, **72**, 1373–83.

Sherman,M. and Contreras,L. (2018) Computational approaches in design of nucleic acid-based therapeutics. *Curr. Opin. Biotechnol.*, **53**, 232–239.

Siebert,S. and Backofen,R. (2007) A dynamic programming approach for finding common patterns in RNAs. *J. Comput. Biol.*, **14**, 33–44.

Speransky,S. *et al.* (2019) A novel RNA aptamer identifies plasma membrane ATP synthase beta subunit as an early marker and therapeutic target in aggressive cancer. *Breast Cancer Res. Treat.*, **176**, 271–289.

Zhou,J. and Rossi,J. (2017) Aptamers as targeted therapeutics: current potential and challenges. *Nat. Rev. Drug Discov.*, **16**, 181–202.