Bioinformatics, 32(2), 2016, 161–164
doi: 10.1093/bioinformatics/btv545
Advance Access Publication Date: 22 September 2015
Original Paper

Sequence analysis

APTANI: a computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data

J. Caroli[1], C. Taccioli[1], A. De La Fuente[2], P. Serafini[2] and S. Bicciato[1,*]

[1]Center for Genome Research, Department of Life Sciences, University of Modena and Reggio Emilia, Modena, Italy and [2]Department of Microbiology & Immunology, UM/Sylvester Comprehensive Cancer Center, Leonard M. Miller School of Medicine, University of Miami, Miami, FL 33136, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on May 26, 2015; revised on August 17, 2015; accepted on September 9, 2015

Abstract

Motivation: Aptamers are synthetic nucleic acid molecules that can bind biological targets in virtue of both their sequence and three-dimensional structure. Aptamers are selected using SELEX, Systematic Evolution of Ligands by EXponential enrichment, a technique that exploits aptamer-target binding affinity. The SELEX procedure, coupled with high-throughput sequencing (HT-SELEX), creates billions of random sequences capable of binding different epitopes on specific targets. Since this technique produces enormous amounts of data, computational analysis represents a critical step to screen and select the most biologically relevant sequences.

Results: Here, we present APTANI, a computational tool to identify target-specific aptamers from HT-SELEX data and secondary structure information. APTANI builds on AptaMotif algorithm, originally implemented to analyze SELEX data; extends the applicability of AptaMotif to HT-SELEX data and introduces new functionalities, as the possibility to identify binding motifs, to cluster aptamer families or to compare output results from different HT-SELEX cycles. Tabular and graphical representations facilitate the downstream biological interpretation of results.

Availability and implementation: APTANI is available at http://aptani.unimore.it.

Contact: silvio.bicciato@unimore.it

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

1 Introduction

DNA or RNA aptamers are small oligonucleotides (<100 nucleotides) that recognize their ligands with an affinity equal or superior to antibodies (Kim *et al.*, 2011). These small molecules are not immunogenic, have a high capacity to penetrate tissues and can be easily modified to regulate their half-life or conjugated with drugs (Wang *et al.*, 2011). Aptamers have entered the clinical pipeline with more than nine phase I and II ongoing clinical trials for diseases as macular degeneration, coronary artery bypass graft surgery and various types of cancer (Keefe *et al.*, 2010). Aptamers are selected *in vitro* by the use of an unsupervised iterative method called SELEX (Systematic Evolution of Ligands by Exponential Enrichment), in which aptamers specific for a ligand are selected from an initial pool of random oligonucleotides using counter-selection and selection procedures. Historically, a SELEX experiment required multiple (>15) rounds of selection with the undesirable effect of amplifying some intrinsic pitfalls, as e.g. the selection of polymerase chain reaction (PCR) artifacts. In recent years, the advent of high-throughput sequencing technologies revolutionized the SELEX technique, drastically reducing the number of cycles and, consequently, the possibility of generating artifacts. HT-SELEX (the SELEX procedure, coupled with high-throughput sequencing) boosted the use of SELEX technique to identify monoclonal aptamers against different epitopes of a ligand or of a whole cell.

However, the quantity of data derived from a single HT-SELEX experiment requires the development of new computational tools to effectively identify aptamers with elevate binding proprieties. Considering that the binding affinity, of a specific aptamer toward its target, depends on the aptamer secondary structure (Tucker *et al*., 2012), structural properties could be exploited to select those aptamers that more likely interact with a ligand. To date, even though interest in secondary structures is rising (Thiel *et al*., 2012), most of the available algorithms and software that analyze HT-SELEX data do not make use of any structural information but rely solely on the abundance (frequency) of the nucleotide sequences at the various cycles. To fulfill this gap, we developed APTANI a software package to select aptamers from HT-SELEX experiments using both sequence counts and structural motifs. APTANI builds on AptaMotif algorithm, originally implemented by Hoinka *et al.* (2012) to analyze SELEX data and extends its applicability to HT-SELEX data. Additionally, APTANI, through the integrative analysis of frequency and secondary structure, allows to identify not only those aptamers with the highest binding probability but also their putative binding motifs. Finally, APTANI comprises some additional functionalities as the possibility to cluster aptamer families or to compare output results from different HT-SELEX cycles.

2 Methods

APTANI workflow consists of four major steps: (i) frequency calculation; (ii) secondary structure and motif breakdown; (iii) extraction of structural motif consensus sequences and (iv) aptamer scoring and structural motif identification.

Given an input file in FASTQ format, the first step calculates the relative frequency of each individual aptamer sequence produced by the HT-SELEX process. Frequency counts can be quantified either using the whole aptamer sequence or only its variable region, i.e. the part of the aptamer comprised between right and left flanking sequences (tags). Low abundant sequences are then filtered, setting a threshold on the minimal frequency (the default value is 10^{-7}), to select only high frequent aptamers for the further steps.

In the second step, APTANI predicts, for each aptamer that passed the frequency filter, all secondary structures in a specific energy range and extracts the motifs represented in these structures (i.e. sub-structures and the correspondent sub-sequences). At this stage, APTANI adopts the same procedure of AptaMotif as described in Hoinka *et al.* (2012). Specifically, secondary structures are predicted using RNAsubopt (Hofacker *et al*., 1994; Wutchy *et al.* 1999), an algorithm, contained in the ViennaRNA package (Lorenz *et al.* 2011) and embedded in APTANI, that calculates all suboptimal secondary structures within a user defined energy range above a minimum free energy (MFE) threshold. For this type of calculation, the default MFE threshold is usually set to 3 Kcal/mol. However, given the massive number of analyzed sequences in SELEX/HT-SELEX experiments (from 1 to 3 million sequences), an MFE threshold of 3 Kcal/mol requires an exorbitant computation time. Thus, to identify an appropriate threshold for the MFE, we performed several rounds of secondary structure prediction testing different values of the MFE and identified in 1 Kcal/mol a reasonable trade-off between running time and thorough structure investigation. Nevertheless, this value can be easily modified setting the *energy* (-e) parameter from the APTANI command line (see the Supplementary Information for further details on the usage of APTANI parameters). As in AptaMotif (Hoinka *et al*., 2012), we consider four different types of secondary structure motifs, i.e. hairpin loops, bulge loops (either right or left) and intra-strand loops. Specifically, hairpin loops are closed continuous structures characterized by the pairing of two nucleotides that close the loop and confer the hairpin conformation; intra-strand loops are structures consisting of two different strands of variable length, ranging from three to a non-definite number of nucleotides and bulge loops (either right or left) are sub-structures of the intra-strand loop category in which one strand consists of two nucleotides, while the other has a non-defined length. Right and left bulge loops are defined depending on where the two-nucleotide strand lays. Since RNAsubopt outputs secondary structures as combinations of dots and brackets, we defined dedicated regular expressions to search each type of loop structure and retrieve the associated nucleotide sequence from the investigated aptamer. Secondary structures are identified on the whole aptamer sequence or, in case the frequencies have been quantified using the variable region only, on the variable region supplemented by right and left tags.

The third step extracts consensus representations for any of the four secondary structure motifs from the sub-sequences of all aptamers where a specific structural motif has been identified. To reduce the computational load of inspecting all sub-structures and sub-sequences of the whole aptamer pool, we assume that if a structural motif, with a given sub-sequence, is shared by a large fraction of aptamers, then it is highly probable that the motif will emerge even when considering only a subset of the entire pool (Bowser, 2005). Thus, a portion of all aptamers and their secondary structure motifs are iteratively randomly picked from the output of the second step, and their sub-sequences are aligned to obtain a consensus sequence for any of the four secondary structure motifs. This step results four different consensus sequences at any iteration, i.e. one for any of the four different types of secondary structure motifs (hairpin loops, left and right bulge loops and intra-strand loops). Consensus sequences, for any of the four secondary structure motifs, are constructed using the most frequent nucleotides of the aligned sequences. Gaps are introduced in the case nucleotides show a frequency lower than a background frequency or when two or more nucleotides have the same frequency. The background frequency is calculated counting the occurrences of each nucleotide in the filtered aptamers pool and dividing them by the total number of nucleotides investigated. The number of sub-samplings and the percentage of the whole aptamer pool selected at each iteration are specified through the *cycle* (-c) and the *percentage* (-p) parameters, respectively. Multiple sequence alignment is performed with Clustal Omega (Sievers *et al*., 2011) using default parameters. However, Clustal Omega parameters (as, e.g. the number of hidden Markov model iterations or the maximum number of examples in any cluster) can be modified from the APTANI command line. Clustal Omega alignments can be visualized as a clustering tree. The clustering plot is generated using FigTree, embedded in APTANI and freely available at http://tree.bio.ed.ac.uk/software/figtree/.

In the last step, any aptamer sub-sequence, correspondent to a specific secondary structural motif, is aligned to the consensus representation of that structural motif (i.e. the sequence of the hairpin loop in a given aptamer is aligned to the hairpin loop consensus sequence). Before alignment, the motif sub-sequence in each aptamer is trimmed to the length of the consensus sequence. The alignment score is then calculated using a match/mismatch-scoring scheme (match = +1; mismatch = −1) with a gap penalty of 0.5. The total score is finally normalized to account for the different consensus sequence lengths. Score values range from −1 to 1, with a score of −1 indicating that the aptamer secondary structure motif is completely different from its respective consensus sequence. Instead, a score of

1 indicates a complete match between the aptamer and the consensus secondary structure motifs. In essence, the score, quantifying the similarity of an aptamer secondary structure motif to the ideal motif of the most abundant secondary structures, gives an indication of the binding potential of aptamer sequences bearing the motif. The normalized alignment score is thus used to rank aptamers (that passed the frequency filter) in terms of matching of the motifs they contain to the corresponding secondary structure consensus motifs. As a result, APTANI returns, for any aptamer, its abundance, the alignment scores, the structural motifs and the consensus sequences (Supplementary Table S1).

2.1 Installation and usage

APTANI is written in Python 3.3, does not require any specific Python module and the package contains all necessary files for the execution of a complete analysis. However, the software requires the installation of Clustal Omega (Sievers *et al.*, 2011) and of a Java Virtual Machine while includes RNAsubopt 2.1.9 (Wuchty *et al.*, 1999), to calculate secondary structure and FigTree (http://tree.bio.ed.ac.uk/software/figtree/) to generate the clustering tree image. Further details on the software installation and usage are available in Supplementary Information.

3 Results

To test APTANI performances and validate its findings, we analyzed a sequence library corresponding to an HT-SELEX experiment designed to isolate aptamers specific for murine IL4Ra (Roth *et al.*, 2012). Briefly, epoxy beads conjugated with the extracellular domain of IL4Ra were used to screen a combinatorial random RNA library of approximately 10^{14} aptamer species. Libraries from the PCR reaction of SELEX cycles 0, 1, 3, 5 and 11 were tagged in 5' and 3' with DNA tails containing the primers and tags for hybridization and control. Libraries were quantified via real-time PCR and bio-analyzer, admixed equally and sequenced with an Illumina NGS. The derived FASTQ files were processed with Illumina software to separate the clones from each library and imputed to APTANI to select potentially binding aptamers directed to IL4Ra. The experiment comprised 11 different cycles of evolutionary selection of the randomly generated aptamers, leading to a final cycle containing approximately 2–3 million sequences.

Setting the threshold for the frequency cut-off at 10^{-7}, APTANI selected, from cycle 11 data, 410 842 different aptamer sequences of 99 base pair length. The majority of the sequences from the last cycle were extremely similar (95–99% similarity with Clustal Omega alignment) to the Cl.42 aptamer, previously demonstrated to be specific for mouse and human IL4Ra using conventional methods (i.e. cloning and sequencing; Roth *et al.*, 2012). Interestingly, while conventional methods were able to identify the Cl.42 clone starting from cycle 5, APTANI identified Cl.42 just from the data of the first cycle of selection (cycle 1), i.e. from a pool of sequences still in the process of being selected and composed mostly of the initial random oligonucleotides with only few aptamers displaying affinity for the target. The secondary structure analysis of these cycle 1 sequences (using 100 iterations and selecting, at each iteration, a number of random aptamers equal to 20% of the entire pool) allowed selecting 53 aptamers containing secondary structure motif sub-sequences with an alignment score (with loop consensus sequences) >0.25 and clustering in three major families dominated by Cl.42 aptamer (alignment score > 0.8; Supplementary Fig. S1). Frequency and score values of the intra-strand loop identified by APTANI in Cl.42 suggested that the CCAUGC secondary structure motif could be essential for the binding to IL4Ra. To test this hypothesis, we generated a mutant aptamer (mutCl.42) in which the CCAUGC motif was substituted by UUUCCC. The analysis of the putative secondary and tertiary structure of mutCl.42, using RNAfold (Denman, 1993) and RNAComposer (Popenda *et al.*, 2012), confirmed that this mutation completely disrupts the intra-strand loop of Cl.42 (Fig. 1). To experimentally verify the dependency of Cl.42 binding activity from the intra-strand loop, we evaluated Cl.42 and mutCl.42 affinities in a binding assay against epoxy beads loaded with recombinant IL4Ra. As shown in Figure 1, while Cl.42 correctly binds to beads loaded with IL4Ra, the mutant aptamer does not show any binding activity suggesting that the intra-strand loop identified by APTANI is indeed required for conferring functional activity to the aptamer sequence.

To access the impact of parameters on the final results, we run a second analysis with lowering the number of iterations to 50, while keeping constant the percentage of randomly picked aptamers at each cycle (i.e. 20% of all aptamers) and the frequency threshold. With this set of parameters, APTANI retrieved 30 aptamers, containing secondary structure motif sub-sequences with an alignment score >0.25 and clustering in three major families dominated by Cl.42 aptamer, whose alignment score remained >0.8. We finally performed the same analysis considering the aptamer variable region. With the same set of parameters used in the previous experiments, APTANI retrieved, from cycle 11, 416 616 different variable sequences, with a frequency spanning from 10^{-1} to 10^{-8} and a length ranging from 35 to 40 nucleotides. Of these variable sequences, 42 presented highly populated motifs and frequencies. As in the

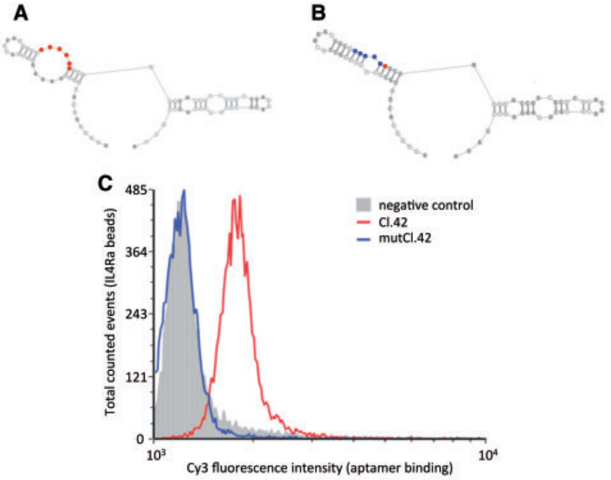

Fig. 1. The Cl.42 loop identified by APTANI is necessary for Cl.42 aptamer binding to the cognate receptor. To assess the functional relevance of the CCAUGC motif identified by APTANI, we first synthetized the Cl42 aptamer (**A**) and a mutant Cl.42 (mutCl.42, **B**), with the original CCAUG sequence substituted by UUUCC. Secondary structure analysis suggests that this five nucleotides substitution completely abrogates the original intra-strand loop (A and B). Then, to evaluate whether the CCAUGC motif was important for the aptamer binding to IL4Ra, Cl.42, mutCl.42 and an irrelevant aptamer were labelled with Cy3 and incubated with epoxybeads conjugated with recombinant IL4Ra. Binding was evaluated by FACS after washing out the unbound Cy3 labeled aptamer with PBS (**C**). As expected, the Cl.42 aptamer (red line) binds to the cognate receptor as determined by the higher fluorescence intensity detected on beads. On the contrary, the mutCl.42 aptamer (blue line) shows a binding similar to the irrelevant aptamer (gray), used as negative control

previous experiments, Cl.42 aptamer resulted the most abundant sequence and its intra-strand loop among the most represented motifs.

3.1 Comparison with other methods

To date, four different tools are available to analyze aptamers sequences obtained by SELEX and HT-SELEX experiments, i.e. AptaMotif (Hoinka *et al*., 2012), AptaCluster (Hoinka *et al*., 2014), MPBind (Jiang *et al*., 2014) and FASTAptamer (Alam *et al*., 2015). AptaMotif, the algorithm that inspired APTANI, performs a secondary structure motif analysis on SELEX data but cannot be applied to HT-SELEX experiments. AptaCluster is a robust method to cluster SELEX and HT-SELEX data, but although extremely efficient in clustering large aptamer pools, it does not take into account the secondary structure conformation of the aptamers during the investigation process. Moreover, the usability of AptaCluster is hampered by some software dependencies (i.e. MySQL, C++ libraries, etc.) that require a sound informatics expertise for installation and usage. MPBind scans aptamer sequences for conserved sub-sequence motifs and then applies a statistical analysis to define their relevance. MPBind cannot analyze HT-SELEX data and, although able to retrieve sub-sequence motifs, no secondary structure analysis is performed during this process. FASTAptamer handles both SELEX and HT-SELEX data and contains a dedicated tool (named FASTAptamer Search) that allows searching the aptamer pool for user-defined sequence motifs. However, FASTAptamer is designed neither to perform any secondary structure analysis nor to search for de-novo sub-sequences or sub-structures. Finally, also Galaxy (https://usegalaxy.org/) NGS-pipelines could, in principle, be applied for the analysis of HT-SELEX data. As evidenced in Supplementary Table S2, only AptaCluster and FASTAptamer can analyze HT-SELEX data; however, we could compared APTANI only to FASTAptamer, since a compilation error in the C++ code AptaCluster blocked installation on different Linux distributions.

When applied to the FASTAptamer reference dataset (Ditzler *et al*., 2013, APTANI identified as intra-strand loop the asymmetric loop structures of FASTAptamer original publication (Alam *et al*., 2015). Moreover, APTANI retrieved several different sub-sequences, related to this intra-strand loop, partially overlapping with the degenerate motifs ArCGUy and CArAr (r and y stand for any purine and any pyrimidine, respectively) identified by FASTAptamer (Supplementary Table S3). APTANI was also able to identify high-populated hairpin loops that FASTAptamer algorithm was unable to find (Supplementary Table S4). Although both algorithms used <8 GB of RAM, APTANI outperformed FASTAptamer in terms of computational speed, completing the analysis in about 30 min compared with the 90 min required by FASTAptamer. Finally, the FASTA format of FASTAptamer output resulted less intuitive to interpret compared with the tab delimited tables produced by APTANI.

The main characteristics of the different tools for the analysis of SELEX data are summarized and discussed in the Supplementary Information, along with a comparison of their performances.

Funding

This work was supported by the Italian Ministry of Education, University and Research (FIRB RBAP11T3WB and EPIGEN - Italian Flagship Project Epigenomics), by AIRC Special Program Molecular Clinical Oncology 5x1000 and by the JDRF foundation award 17-2013-326.

Conflict of Interest: none declared.

References

Alam,K.K. *et al*. (2015) FASTAptamer: a bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Mol. Ther. Nucleic Acids*, **4**, e230.

Bowser,M.T. (2005) SELEX: just another separation? *Analyst*, **130**, 128–130.

Denman,R.B. (1993) Using RNAFOLD to predict the activity of small catalytic RNAs. *BioTechniques*, **15**, 1090–1095.

Ditzler,M.A. *et al*. (2013) High-throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase. *Nucleic Acids Res*., **41**, 1873–1884.

Hofacker,I.L. *et al*. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, **125**, 167–188.

Hoinka,J. *et al*. (2012) Identification of sequence-structure RNA binding motifs for SELEX-derived aptamers. *Bioinformatics*, **28**, i215–i223.

Hoinka,J. *et al*. (2014) AptaCluster—a method to cluster HT-SELEX aptamer pools and lessons from its application. *Res. Comput. Mol. Biol*., **8394**, 115–128.

Keefe,A.D. *et al*. (2010) Aptamers as therapeutics. *Nat. Rev. Drug Discov*., **9**, 537–550.

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.

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

Popenda,M. *et al*. (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res*., **40**, e112.

Roth,F. *et al*. (2012) Aptamer-mediated blockade of IL4Rα triggers apoptosis of MDSCs and limits tumor progression. *Cancer Res*., **72**, 1373–1383.

Sievers,F. *et al*. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol*., **7**, 539.

Thiel,W.H. *et al*. (2012) Rapid identification of cell-specific, internalizing RNA aptamers with bioinformatics analyses of a cell-based aptamer selection. *PLoS One*, **7**, e43836.

Tucker,W.O. *et al*. (2012) G-quadruplex DNA aptamers and their ligands: structure, function and application. *Curr. Pharm. Des*., **18**, 2014–2026.

Wang,P. *et al*. (2011) Aptamers as therapeutics in cardiovascular diseases. *Curr. Med. Chem*., **18**, 4169–4174.

Wuchty,S. *et al*. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.