# AptamerRunner: An accessible aptamer structure prediction and clustering algorithm for visualization of selected aptamers

Dario Ruiz-Ciancio[1,2,3], Suresh Veeramani[4,5], Rahul Singh[6], Eric Embree[7], Chris Ortman[8], Kristina W. Thiel[5,9], and William H Thiel[4,*]

[1] Instituto de Ciencias Biomédicas (ICBM), Facultad de Ciencias Médicas, Universidad Católica de Cuyo, Av. José Ignacio de la Roza 1516, Rivadavia, 5400, San Juan, Argentina.

[2] National Council of Scientific and Technical Research (CONICET), Godoy Cruz 2290, C1425FQB Ciudad Autónoma de Buenos Aires Argentina.

[3] Cancer Genome Engineering Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona 08035, Spain.

[4] Department of Internal Medicine, University of Iowa, Iowa City, IA 52242, USA.

[5] Holden Comprehensive Cancer Center, University of Iowa, Iowa City, IA, 52242, USA.

[6]Department of Computer Sciences, University of Iowa, Iowa City, IA 52242, USA.

[7]Carver College of Medicine, University of Iowa, Iowa City, IA 52242, USA.

[8]Institute for Clinical and Translational Science, University of Iowa, Iowa City, IA 52242, USA.

[9]Department of Obstetrics and Gynecology, University of Iowa, Iowa City, IA 52242, USA.

*Corresponding author. Email: william-thiel@uiowa.edu

## Abstract

Aptamers are short single-stranded DNA or RNA molecules with high affinity and specificity for targets and are generated using the iterative Systematic Evolution of Ligands by EXponential enrichment (SELEX) process. Next-generation sequencing (NGS) revolutionized aptamer selections by allowing a more comprehensive analysis of SELEX-enriched aptamers as compared to Sanger sequencing. The current challenge with aptamer NGS datasets is identifying a diverse cohort of candidate aptamers with the highest likelihood of successful experimental validation. Herein we present AptamerRunner, an aptamer sequence and/or structure clustering algorithm that synergistically integrates computational analysis with visualization and expertise-directed decision making. The visual integration of networked aptamers with ranking data, such as fold enrichment or scoring algorithm results, represents a significant advancement over existing clustering tools by providing a natural context to depict groups of aptamers from which ranked or scored candidates

can be chosen for experimental validation. The inherent flexibility, user-friendly design, and prospects for future enhancements with AptamerRunner has broad-reaching implications for aptamer researchers across a wide range of disciplines.

## Introduction

Aptamers are short synthetic RNA or DNA oligonucleotides that recognize target epitopes with specificity and affinity analogous to antibody-antigen interactions[1]. Applications of aptamers are broad-reaching and include biosensors[2], research reagents[3, 4], tools for mechanistic discovery[3], diagnostics[5], delivery platform[6] and therapeutics[5]. Recently the aptamer avacincaptad pegol, that target complement C5, was FDA-approved for the treatment of geographic atrophy[7]. Aptamers are generated using an *in vitro* process known as Systematic Evolution of Ligands by EXponential enrichment (SELEX)[8, 9]. The SELEX process now includes numerous variations with new SELEX strategies constantly being developed[1, 10]. At the completion of SELEX, selected aptamers are identified by sequencing, with the field shifting from Sanger sequencing towards next-generation sequencing (NGS) platforms. NGS yields hundreds of millions of reads, with each read containing the entirety of an aptamer sequence. Thus, the aptamers enriched during SELEX can now be interrogated to a degree not achievable with Sanger sequencing[11, 12]. However, these large NGS datasets have created new challenges in terms of how parse the millions of reads to identify the best candidate aptamers to then validate experimentally. Testing thousands or even hundreds of aptamer candidates is still unattainable for most aptamer researchers; therefore, identification of the top candidates is paramount. To address this need, several bioinformatics approaches specific for aptamer NGS datasets have emerged[13, 14]. The analysis of an aptamer NGS dataset includes processing the FASTQ data and application of strategies to identify candidate aptamers using various motif identification, scoring, and clustering algorithms[13, 14]. Bioinformatics tools to identify candidate aptamers are frequently applied in concert, for example clustering is used to identify separate groups of aptamers[15-17], followed by ranking aptamers within these groups using fold enrichment or scoring algorithms[18, 19].

The central theory behind aptamer clustering is that aptamers that are closely related based on their sequence and predicted structures are likely to target the same epitope[13, 14]. The earliest efforts to cluster aptamers used sequence alignment algorithms such as ClustalW[20-22], but these data can be

difficult to interpret and this method does not take into account the predicted structures of the aptamers. Our group introduced the concept of clustering aptamers by either sequence relatedness using Levenshtein edit distance or by predicted secondary structure relatedness using tree distance[17]. We applied a clustering strategy, termed the *all-vs-all* approach, wherein all aptamers within a dataset are compared to each other. Clusters were defined as the aptamers that interconnect within a threshold distance measure (e.g., edit distance of 3). However, this early clustering algorithm was not easily accessible and thus has not been widely adopted. The current prevailing clustering algorithms include AptaCluster[23, 24], FASTAptamer[25], and FASTAptameR 2.0[26]. These tools generate networks of related aptamers using either Hamming or Levenshtein edit distance and introduced a new concept for clustering aptamers termed the *seed approach*. The seed approach generates networks of aptamers centered around a *seed sequence*, which is defined as the sequence with the most reads within an aptamer NGS dataset (i.e., most abundant sequence). All aptamers that connect to the seed sequence within a threshold edit distance measure are designated as a cluster. This process is iterated by removing the initial seed sequence and all connected aptamer sequences from further analysis, and the next most abundant sequence becomes the seed for the next cluster. These seed-based aptamer clustering algorithms are significantly more accessible than our initial algorithm, and the seed approach has significant computational advantages over the all-vs-all approach. A limitation of the seed approach is that it is a greedy process that can potentially miss important inter-aptamer relationships identified by the all-vs-all clustering approach. In addition, the available seed-based algorithms do not consider structure when generating networks of related aptamers. The output from these algorithms is text-based rather than graphically represented, which severely limits interpretation of the clustering results and prevents integration of ranking data that is necessary to identify candidates within the groups of interconnected aptamers.

Herein, we introduce *AptamerRunner*, an accessible aptamer structure prediction and clustering algorithm for the visualization of networked aptamers. AptamerRunner was designed based on the principles of *experiential computing*[27, 28], which is founded on the idea that an understanding of complex biological data comes from integrating user expertise with algorithmic processing via data visualization and user-data interactions. Using the paradigm of experiential computing, we designed AptamerRunner so that it has the flexibility to ensure that aptamer researchers can apply different clustering strategies to suit their needs, with customizable visualization support enabling

user-specific data interpretation. AptamerRunner includes several novel clustering features not previously available: 1) the option to use either the all-vs-all approach or the seed approach. From the perspective of computational complexity, if, we have $n$ aptamer sequences, with $m$ being the length of the longest aptamer, the time complexity of the all-vs-all approach is $O(n^2m^2)$, due to the complexity of computing the Levenshtein distance. The seed approach, on the other hand, involves identifying the most abundant aptamers, which requires determining the frequency of each unique aptamer sequence present in the dataset. Using hashing, this can be obtained in $O(nm)$ time. If $k$ different seeds are considered, then the complexity of the approach is $O(nmk)$; 2) the option to interrogate both sequence and structure relatedness simultaneously by applying logical operators (AND, OR) with edit and tree distance thresholds; and 3) inclusion of distance measure data as metadata within the edges of the interconnected aptamer sequences. To provide easier access and functionality, AptamerRunner and all dependencies have been packaged into a Docker image[29] that is operated by command line using a platform-independent .NET script. The AptamerRunner clustering algorithm outputs results as an eXtensible Graph Markup and Modeling Language (XGMML) file that permits graphical representation of the clustering results using network analysis programs such as the open-source Cytoscape program[30]. Importantly, through the graphical representation of the clustering data, additional information such as ranking data can be overlayed onto the networked aptamers to facilitate an integrated analysis whereby clusters of aptamers and ranking data can be interpreted at the same time. This integration of clustering with ranking data presents a novel analysis of selected aptamers that aids in the identification of the best candidates.

## Results

### AptamerRunner Overview

AptamerRunner is a .NET program coded in C# that generates the Docker commands to adapt to various operating system constraints (**Figure 1A**). AptamerRunner will check the Docker repository and download the most recent AptamerRunner Docker image (for detailed instructions to use AptamerRunner refer to supplemental methods). The AptamerRunner Docker image is comprised of two independent python algorithm components (**Figure 1B and 1C**). The first component predicts the secondary structure of RNA or DNA aptamer sequences (**Figure 1B**). The second component calculates aptamer relatedness to generate networks of related aptamer sequences (**Figure 1C**). This segmentation permits each component to be implemented independently and provides additional flexibility to the user. For more technically proficient users, both python algorithm components can be operated by command line independent of Docker. Once the AptamerRunner Docker container has completed running the structure predication algorithm or clustering algorithm, AptamerRunner will close the AptamerRunner Docker container.

To demonstrate the utility of AptamerRunner and compare to other aptamer clustering tools , we used a published aptamer NGS dataset from a selection against B-cell leukemia cells[16]. With these data, we applied the various AptamerRunner clustering options with either edit distance of 1 or tree distance of 0, when applicable. When comparing AptamerRunner against FASTAptamer, FASTAptameR 2.0 and AptaCluster, we applied comparable clustering parameters using an edit distance of 1 with the seed approach.

### Visualization of AptamerRunner clustering results

A central principle of experiential computing is combining algorithms with data visualization and user-data interactions to facilitate the expertise of a user to interpret complex data. We enabled the visualization of AptamerRunner clustering results by exporting the clustering data into the XGMML format, which can be imported into the network visualization software Cytoscape (see supplemental methods for specific details). Once the clustering results have been imported into Cytoscape, they can be visualized using the multitude of Cytoscape's built-in network layout functions. This approach not only provides a clear and organized representation of the data but also facilitates deeper analysis through the various customization and analytical tools available within Cytoscape. AptamerRunner and Cytoscape's visual representation of networked aptamers

provides a natural context to interpret clustering data that is significantly easier to interpret than text-based results. For example, when the seed option is used to generate the networks of related aptamers, the seed sequence is clearly identifiable as the central node and the number of sequences connecting to each seed sequences are easily discerned (**Figure 2A and 2B**). As compared to the seed approach, the all-vs-all option produces fewer but more complex clusters of inter-related aptamer sequences (**Figure 2C and 2D**). The all-vs-all approach clustering results can include large, complex hairball networks (**Figure 2C**, largest cluster) and smaller clusters with naturally occurring central nodes. Importantly, clustering conducted using either the seed approach or the all-vs-all approach can provide different perspectives of the same data.

A second important improvement within AptamerRunner, as compared to other clustering algorithms, is the introduction of logical operators AND and OR with edit distance and tree distance. The AND function and the OR function can be applied to gain insight into potential functional aspects of an aptamer. The AND function can reveal areas of nucleotide substitutions that are well-tolerated or that impart beneficial properties. Conversely, the OR function can reveal nucleotide changes that have a significant impact on the predicted structure of an aptamer. The use of the AND logical operator places a higher stringency, and the use of the OR logical operator less stringency, onto the networks of related aptamer generated. For example, clustering the data presented in **Figure 2** using the AND logical operator with an edit distance 1 and tree distance 0 yields smaller, more constrained networks of related aptamers when using the seed (**Figure 3A**) and all-vs-all approaches (**Figure 3B**). Conversely, the OR logical operator yields larger less constrained networks of related aptamers (**Figure 3C and 3D**).

Comparison of AptamerRunner to other aptamer clustering algorithms: FASTAptamer, FASTAptameR 2.0 and AptaCluster

We next compared the capabilities and output of AptamerRunner to the capabilities and output of FASTAptamer, FASTAptameR 2.0 and AptaCluster[24-26, 31]. Details and features of FASTAptamer, FASTAptameR 2.0, AptaCluster and AptamerRunner are summarized in **Table 1**. The FASTAptamer clustering algorithm (*FASTAptamer-Cluster*) and FASTAptameR 2.0 clustering algorithm (*cluster module*) determine sequence families of aptamers by Levenshtein edit distance using the seed approach. FASTAptameR 2.0, an update to FASTAptamer, operates through an offline web-browser that accesses a Docker container. FASTAptameR 2.0, like FASTAptamer, applies a seed approach using edit distance only, but has a computationally faster clustering

algorithm and a cluster visualization function (*cluster diversity module*) that generates a principal component plot (PCA) using a k-mer matrix of the clustered aptamer sequences. AptaCluster is provided as a component of the AptaSuite package with a GUI that is accessible as a Java application[32]. AptaCluster filters input data using a local sensitivity hash function and clusters aptamers by Hamming edit distance using the seed approach.

Of note, FASTAptamer, FASTAptameR 2.0 and AptaCluster are incapable of generating clusters based upon predicted structures, nor do they compare all aptamers to each other regardless of their enrichment during SELEX (e.g., all-vs-all approach). AptaSuite, of which AptaCluster is a component, includes an algorithm AptaTrace that identifies sequence-structure motifs as sequence logos with secondary structure probability profiles[33]. The AptaTrace algorithm does not perform clustering of these predicted secondary structure and thus is not included in the comparison to AptamerRunner. To compare AptamerRunner to FASTAptamer, FASTAptameR 2.0 and AptaCluster, we applied the seed approach using only edit distance with no logical operators, and we used aptamer NGS data from a selection against B-cell leukemia cells[16].

The FASTAptamer and FASTAptameR 2.0 clustering algorithms produced identical modified FASTA formatted text files (**Figure 4A**). Each sequence identifier within the FASTA file (denoted with the ">" annotation) contains the following data features for each aptamer sequence: sequence ID, raw read count, normalized read count, cluster ID, rank within the cluster, and edit distance from the seed sequence. For example, in **Figure 4A**, the first listed sequence has an identifier of ">1-3290398-899773.77-1-1-0" which denotes a sequence ID of 1, 3290398 raw read count, 899773.77 normalized read count, cluster ID of 1, rank within cluster of 1, and edit distance of 0 from seed. Within a given cluster ID, the seed sequence is the listed first, is the highest ranked aptamer sequence and will have an edit distance from the seed of 0. Subsequent clusters can be identified by the cluster ID. For example, in **Figure 4A,** the second cluster begins with the following identifier ">2-41052-11225.85-2-1-0". FASTAptameR 2.0 includes the option to output a CSV data table (**Figure 4B**), which makes the clustering results sortable and easier to parse separate clusters.

AptaCluster outputs two data files: a modified FASTA file and a data table of seed sequences (**Figure 4C, D**). The AptaCluster modified FASTA file (**Figure 4C**) includes a sequence identifier line (denoted by ">>") with the cluster ID followed by the aptamer sequences within that cluster

(denoted by ">"), starting with the seed sequence. Read counts are included on the sequence line following the aptamer variable region sequence. For example, in **Figure 4C**, the first cluster is identified as Cluster 103475, with a total of 640,120 read counts among all sequences within that cluster. The first listed sequence, which is the seed of Cluster 103475, is named Aptamer_2 and has a read count of 638,988. The AptaCluster seed sequence data table includes the cluster ID, seed sequence, and information about the clusters from each selection round; example data are provided for Round 9 in **Figure 4D**. For each selection round, AptaCluster seed table provides the proportion of a cluster relative to other identified clusters ("R9 Size"), the number of sequences within that cluster ("R9 diversity") and the total number of normalized read counts per million ("R9 CPM").

By comparison, AptamerRunner yields a visual output of clusters as depicted in **Figure 2A**. In this example using the same data that were analyzed by FASTAptamer, FASTAptameR 2.0 and AptaCluster, the seed sequence was the central node in the top left cluster. Note that all programs identified the same sequence as the seed, but the visual output by AptamerRunner significantly improves data interpretation and candidate aptamer selection because all clusters can be easily viewed simultaneously. Users can determine aptamer sequence, predicted secondary structure, and any other imported metadata (e.g., fold enrichment) by simply clicking on a node (diagrammed in **Supplemental Figure S1**). This interactive visualization is not possible with text-based results. FASTAptameR 2.0 does include functions to visualize an analysis of the clustering data within the *diversity* module. The diversity module provides a series of graphs (cluster metaplots) that depict information about the population of clustered aptamers (**Supplemental Figure S2A**, sequence count, read count and average LED) and a PCA plot of a k-mer matrix (**Supplemental Figure S2B and S2C**). While these PCA plots can provide insight into relative diversity of the different clusters, this function is limited to plotting no more than 15 clusters concurrently, and users must cross-reference text-based results to determine the identify of specific aptamer sequences within the PCA plot. Whereas results from AptamerRunner within Cytoscape are interactive. The limitations of FASTAptamer-Cluster, FASTAptameR 2.0 and AptaCluster text-based outputs highlight the importance of visualizing clustering data to provide a natural context for representing the different clusters of aptamers.

<u>Improving candidate aptamer selection by integrating fold enrichment data and scoring algorithm results onto AptamerRunner clustered aptamers</u>

A limitation of both the text-based and visualized clustering results is that they do not provide insight into which clusters are most likely to yield the best aptamers. Clustering separates aptamers into groups that likely target the same epitope, but additional data are necessary to score and rank the aptamers to identify ideal candidates from within each cluster of aptamers. We hypothesized that integration of fold enrichment or data from scoring algorithms significantly enhances candidate selection when integrated with clustering results. Unfortunately, the text-based outputs of FASTAptamer and FASTAptameR 2.0 do not allow for easy integration of fold enrichment or scoring data. The AptaCluster seed sequence data table provides selection round normalized read counts that can be used to calculate the overall round-to-round enrichment of each cluster (see **Figure 4D**), but data from this table must be manually cross-referenced with the clustering results to determine the enrichment of all sequences within the cluster.

AptamerRunner overcomes the limitations of text-based results through visual integration of fold enrichment data and results from scoring algorithms. Data tables were imported into Cytoscape; these tables contained log10 normalized read counts of round 9 and the log2 fold enrichment between selection rounds (e.g., round 2 to round 5) of aptamers that were clustered using the seed approach with an edit distance of 1. As shown in **Figure 5**, such visual overlays of node size and color provide easy to interpret visual cues of individual aptamer abundance and enrichment during the SELEX process. Groups of aptamers showing positive enrichment (green) from negative enrichment (yellow) are easily discernable. Furthermore, the log2 fold enrichment can be easily evaluated between different selection rounds. The B-cell SELEX process sequence enrichment exhibited a sigmoidal curve, with rounds 0 to 2 representing the initial exponential phase, rounds 2 to 5 representing the linear phase and rounds 5 to 9 representing the asymptotic phase (**Figure 5A**). Interestingly, with this visual comparison we observed positive enrichment from rounds 0-2 of the selection (**Figure 5B**), but the most the most interesting changes of log2 fold enrichment seem to occur between selection rounds 2 to 5 (**Figure 5C**) and rounds 5 to 9 (**Figure 5D**). For example, when comparing round 5 to 9, the largest group of clustered aptamers (top left cluster) has a significant diversity of log2 fold enrichments that ranges from -9.95 to 5.1, and the seed sequence remained close to 0 (**Figure 5D**). These data suggest that the seed sequence may not be the ideal candidate from this cluster, but rather one of the other aptamer sequences within 1 edit distance of the seed sequence with a positive log2 fold enrichment is preferred for subsequent testing. By contrast, the seed sequence in the top right cluster is a desirable candidate based on its

positive log2 fold enrichment from round 5 to round 9. This type of analysis of AptamerRunner clustering results mapped with log2 fold enrichment data was used with the Ruiz-Ciancio *et. al.* study to identify 38 candidate aptamers from 38 separate sequence clusters (all-vs-all approach with edit distance 1) or structure clusters (all-vs-all approach with tree distance 3). Within each cluster, candidates were defined as the aptamer sequences with the highest log2 fold enrichment from rounds 2 to 5 or rounds 5 to 9. We favor the all-vs-all approach based on the supposition that the seed approach may miss interesting inter-aptamer relationships; however, we present data in this head-to-head comparison by using the seed approach since other aptamer clustering tools cannot accommodate the all-vs-all approach.

Bioinformatics tools such as MPbind[19] and RaptRanker[18] score and rank aptamers enriched during SELEX, and these data can be overlayed onto AptamerRunner clustering data in Cytoscape. The MPbind scoring algorithm generates a combined meta z-score of aptamer sequences based on relative motif enrichment and abundance of the final aptamer selection round. RaptRanker evaluates aptamer sequence motif and structure of subsequence groups to generate an average motif enrichment score (AME). Integrating data from scoring algorithms enables us to examine the predicted relative aptamer affinities, which can be used to select candidates within each cluster. Interestingly, MPbind and RaptRanker demonstrates significant variation in the predicted affinities of aptamers from the B-cell SELEX dataset (**Figure 6**). Specifically, MPBind predicted that most of the clusters had high affinity for their targets: the majority of the nodes were visualized as red (meta z-scores of 40-60, **Figure 6A**), whereas RaptRanker predicted fewer high affinity aptamers, with most nodes visualized in the blue to yellow spectrum (average motif enrichment of 0-5, **Figure 6B**). Since MPBind and RaptRanker use different principles for scoring, we asked if the highest scoring aptamers were similar between the two algorithms by plotting the scores as interactive scatter plots in Cytoscape. Clusters of aptamers that were scored high by both the algorithms could be identified, but, consistent with the visual representation of the clusters, many more aptamers scored highly by MPBind vs. RaptRanker (**Figure 6C**).

Application of AptamerRunner metadata to re-organize networked aptamers

One rationale for using the seed approach rather than the all-vs-all approach to cluster aptamers is that the all-vs-all approach will frequently generate large, disorganized hairball clusters of aptamers as highlighted in **Figure 7A** (grey nodes and red edges), which are difficult to interpret. To address this limitation with all-vs-all data, the metadata generated by AptamerRunner during

clustering can be used to deconstruct and re-organize the hairballs. For example, the hairball cluster within **Figure 7A** was isolated and the aptamer sequences re-organized using the metadata property of structure relatedness. This approach identified groups of aptamer sequences that are related by edit distance 1 and have identical structures (tree distance = 0), as shown by edges colored blue (**Figure 7B**). Several groups of the re-organized aptamer sequences also exhibited positive log2 fold enrichment from round 5 to round 9 (green nodes). We also asked if this de-convolution approach could be used for other metadata features. In the example shown in **Supplemental Figure S3**, we tested whether enriched aptamers within this hairball are structurally similar. First, we re-organized the hairball in **Supplemental Figure S3A** using the log2 fold change enrichment observed between round 5 to round 9 (Log2 R5:R9, **Supplemental Figure S3B**). Next, we excluded any aptamers with a log2 R5:R9 less than ≤ 0, and the remaining subset of aptamers were reorganized for identical predicted structures (tree distance = 0, **Supplemental S3C**). This approach identified multiple sub-groups of aptamer sequences within the hairball network that were positively enriched during SELEX and are closely related by both sequence and structure. Taken together, these capabilities of AptamerRunner to incorporate metadata for clusters to be visualized in Cytoscape allow for a more in-depth analysis and identification of candidate aptamers. Use of different parameters provides greater resolution of the enriched sub-groups of aptamers, which is a major improvement by AptamerRunner over text-based clustering algorithms.

Analysis of the Interleukin(IL)-10 aptamer dataset by AptamerRunner:

To evaluate the broader utility of AptamerRunner, we analyzed a publicly available aptamer NGS dataset from the NIH Sequence Read Archive (SRA). This database was used to describe AptaCluster and AptaMut[24], as well as to assess potentially beneficial mutations of the IL-10 aptamers[34]. This database contained sequencing data across five selection rounds with approximately 16.7 million reads representing approximately 11.7 million aptamer sequences. Using AptaCluster within AptaSuite, we generated clustering results based on the reported AptaCluster constraints (see supplemental methods).

By comparison, we generated a non-redundant database from the NCBI SRA data, compiling read counts of the ~11.5 million aptamer sequences across the five selection rounds. From this dataset, we identified 2,140 unique aptamer sequences for clustering, representing ~4.5 million.

Using AptamerRunner, we generated clusters of the IL-10 aptamers employing an edit distance of 5, utilizing both the seed approach and the all-vs-all approach (**Figure 8A and 8B**). We then overlaid the published dissociation constant ($K_D$) values onto the networked aptamers to ascertain which sequences had been experimentally evaluated for their ability to bind IL-10. Our results indicate that one larger cluster and seed sequence identified by AptamerRunner appears to have remained untested. This seed sequence was the third most abundant aptamer from the fifth selection round and is evident in the AptaSuite aptamer pool data. These results highlight AptamerRunner's ability to cluster and visualize seed sequences that may be important aptamers for experimental evaluation.

Using AptamerRunner, we next evaluated how the IL-10 aptamers cluster based on predicted secondary structure. We first established an appropriate tree distance measure by examining the distribution of tree distances found within the edit distance 5 seed edges (**Supplemental Figure S4A**). These data indicate that most aptamer sequences within an edit distance of 5 are structurally closely related, with tree distances clustering around 10. Beyond this point, the histogram begins to level off, suggesting a diminishing return in structural similarity. Building on these insights, we evaluated the IL-10 aptamers using AptamerRunner with a tree distance of 10 and the all-vs-all approach (**Figure 8C**). The tree distance of 10 generated numerous clusters that overlapped with several identified clusters at edit distance 5. However, a notable difference emerged; specifically, two clusters, D and K, which are greater than edit distance 5, exhibited similar predicted secondary structures. These results indicate that the IL-10 aptamers may cluster most effectively when both edit distance and tree distance are considered. Consequently, using AptamerRunner, we clustered the IL-10 aptamers within an edit distance of 5 and a tree distance of 10 employing the all-vs-all strategy with the AND logical operator (**Figure 8D**). This methodology provided significant granularity in the clustering outcomes, revealing that clusters with similar $K_D$ values contained multiple aptamers. Importantly, the analysis using AptamerRunner, which considers both edit distance and tree distance, indicates the presence of several clusters lacking representatives tested for IL-10 binding that show substantial enrichment from rounds four to five (**Supplemental Figure S4B**).

## Discussion

The rationale for clustering aptamer sequences is to discern which aptamers likely target the same epitope and conversely, which aptamers likely target different epitopes. By understanding how aptamer sequences are related, and not related, by sequence and structure, a diverse cohort of aptamers can be identified for experimental validation. The AptamerRunner clustering algorithm applies the principles of experiential computing to support expertise-driven decision making for clustering and identification of candidate aptamer sequences. Novel features of AptamerRunner that facilitate expertise-driven decision making include retention of distance measures as metadata and the incorporation of logical operators (AND, OR) for clustering. Maintaining distance measures as metadata provides additional opportunities to analyze, re-evaluate and interpret clustering results. Within Cytoscape, additional data such as the log2 fold enrichment of aptamer sequence across selection rounds and data from other aptamer bioinformatics scoring algorithms can be mapped onto the networks generated by AptamerRunner to aid in the interpretation of the clustering results, further supporting the experiential computing goal. Having the AptamerRunner Docker container controlled by a .NET program presents a simple method to enable users to access AptamerRunner. Users only need to use a command line interface to initiate the .NET script, which then dynamically generates all Docker commands necessary to run AptamerRunner's secondary structure prediction or clustering algorithms. Taken together, the innovative features of AptamerRunner enable a more in-depth analysis of aptamer NGS data and allows for better identification of candidate aptamers. In addition, AptamerRunner can be used to ask new questions about how sequence and structure relatedness contribute to library convergence during the SELEX process.

The superiority and flexibility of AptamerRunner is highlighted by two recent publications that made use of AptamerRunner clustering capabilities in fundamentally different ways[15, 16]. Ruiz-Ciancio *et. al.*[16], using the same NGS dataset as in the present study, applied the AptamerRunner clustering algorithm to identify 38 candidate aptamers from unique clusters related by sequence or by structure. These 38 candidates were then ranked for their potential to bind the CD22 protein through a molecular docking and molecular dynamic approach[16]. The aptamer that exhibited specificity for CD22 (B-ALL1 aptamer) was identified within a unique group of aptamers that were related by structure. Within this cluster of structurally related aptamers, the B-ALL1 aptamer exhibited the greatest positive fold enrichment and was therefore identified as a potential

candidate. An edit distance clustering analysis did not identify B-ALL1 as a candidate. Also, because B-ALL1 was the 573rd most abundant aptamer, it would likely have been missed using traditional candidate selection approaches. This highlights the importance of clustering by related structures, which is not an available capability in FASTAptamer, FASTApameR 2.0 or AptaCluster.

A second study, by Santana-Viera *et. al.*[15], applied AptamerRunner to examine the relatedness of aptamer libraries enriched in two independent SELEX processes: a protein-based SELEX using recombinant human EphA2 as target and a Cell-Internalization SELEX using EphA2-expressing MDA231 cells as targets[15, 35]. The AptamerRunner clustering algorithm identified a group of aptamers within these two different SELEX processes that shared structure and sequence relatedness. From this group of aptamers, the candidate aptamer ATOP was observed to target hEphA2 and exhibited antitumorigenic effects *in vitro* and *in vivo*. The flexibility of AptamerRunner permitted the researchers to develop a novel clustering strategy that made use of both edit distance and tree distance clustering using the all-vs-all approach. A seed approach with two different SELEX processes would have been challenging due to the complication in defining which SELEX would provide the aptamer sequences to serve as seeds. Importantly, without the in-depth analysis of aptamer relatedness enabled by AptamerRunner, the aptamers described by these two studies would have been prohibitively challenging to identify with the other available clustering tools.

Our analysis of the IL-10 aptamer database, which was previously used to demonstrate the utility of AptaCluster, employed AptamerRunner to focus on clustering based on both edit distance and tree distance. Initial observations indicated that aptamers with an edit distance of 5 are structurally related, corresponding to a tree distance of 10. The application of a tree distance of 10 revealed overlapping clusters and underscored the necessity of considering both measures for optimal clustering outcomes. This comprehensive clustering approach, which combines an edit distance of 5 with a tree distance of 10, provides detailed granularity of distinct groups of aptamers, uncovering clusters that had not been previously tested for IL-10 binding. The analysis of the IL-10 aptamer database illustrates how the AND function of AptamerRunner offers a more nuanced understanding of aptamer networks.

The aforementioned examples of AptamerRunner identifying candidate aptamers made use of the all-vs-all clustering approach rather than the seed approach. The seed approach, introduced by AptaCluster and FASTAptamer, is founded on the idea that certain aptamer sequences, called the seed, serve as the basis from which mutations during SELEX accumulate to yield more specific or higher affinity aptamers. However, the seed approach for clustering aptamers is a greedy process whereby sequences connected to the seed are removed from the pool of aptamers available for clustering. Therefore, the seed approach will miss inter-aptamer connections identified by the all-vs-all approach. However, the all-vs-all approach is a significantly more computationally intensive process than the seed approach and can yield hairball networks that are more complex to interpret. The concept of the seed sequence is potentially more important for aptamer libraries with longer variable regions. Longer variable region libraries (e.g. >30 nucleotides) have a large starting complexity that cannot be sampled at the start of SELEX and PCR-generated mutations are more likely to introduce beneficial aptamer sequences not present during the initial selection rounds. FASTAptameR 2.0 includes a function (distance module) that can specifically evaluate edit distance from a seed sequence, or other sequences, to evaluate accumulation of mutations during SELEX. The FASTAptameR 2.0 distance module plots the edit distance distribution from the seed sequence as a histogram. Given that the accumulation of mutations is more likely to occur with more abundant aptamers and less likely with aptamers of lower abundance, AptamerRunner could interrogate larger more complex networks of related aptamers identified by the all-vs-all approach by re-evaluating them using the seed approach. While AptamerRunner did not aim to settle this debate, it does provide experiential computing that allows aptamer researchers to investigate their data independently based on their goals and expertise as to how the clustering data should be visualized.

Future versions of AptamerRunner could include additional aptamer bioinformatics tools to process, compile, and analyze raw aptamer NGS data with a GUI interface like the integrated pipelines offered by FASTAptameR 2.0 and AptaCluster. The structure prediction algorithm could incorporate additional structure prediction algorithms such as Mfold [36] or be modified to permit multiple structures for each aptamer sequence including suboptimal structures. Furthermore, AptamerRunner does not include any option to limit what aptamers are clustered. AptaCluster applies a hashing function that filters the dataset by identifying pairs of aptamers that are likely to be dissimilar and FASTAptameR 2.0 includes a filter function to cluster only aptamer sequences

of a minimum abundance or produce a set number of clusters. With AptamerRunner we filtered the aptamer NGS dataset prior to clustering using a separate aptamer abundance and persistence analysis [16, 37]. Ideally, algorithms that can filter aptamer NGS datasets, such as the AptaCluster hashing function, could be applied independently prior to clustering to investigate the effectiveness of different filtering strategies. Additional future directions include determining the range between separate groups of networked aptamers with the idea that more distantly related aptamers are more likely to target different epitopes. Analysis options could include ranking different groups of clusters aptamers and ranking individual aptamers within each cluster by integrating scoring or application of molecular docking to predict which groups of aptamer bind to same regions of a target protein[18, 19, 33, 38-40].

In summary, AptamerRunner seeks to facilitate human-computer synergy[27] for clustering aptamer NGS data with innovative approaches, enabling diverse sequence and structure relatedness, introducing logical operators, and offering seamless integration with Cytoscape for visualization and interpretation. The inherent flexibility, user-friendly design and prospects for future enhancements collectively position AptamerRunner at the forefront of advancing aptamer research.

## Materials and Methods

RNA or DNA aptamer secondary structure prediction algorithm

The structure prediction component of AptamerRunner requires either full-length aptamers or only the variable region of the aptamers in a FASTA-formatted file. The FASTA file should contain collapsed aptamer NGS data, in which all unique aptamer sequences are represented once and are ranked in descending order based on number of duplicate reads (**Supplemental Figure S5A**). If the FASTA file contains only the variable region sequences, AptamerRunner provides an option to automatically append the constant regions per user input. Secondary structures are predicted using the RNAfold structure prediction algorithm from the Vienna Package v2.0 [41-43], with the lowest minimum free energy structure being retained. Pass-through commands specific for RNAfold can be included during AptamerRunner execution. The predicted aptamer secondary structures are appended to a modified FASTA-formatted file (FASTA.struct) using dot-bracket annotation (**Supplemental Figure S5B**). Properties associated with the predicted structures (e.g,. minimum free energy) that are output by the RNAfold are compiled into a separate tab-delimited file with the aptamer FASTA header information as a key. These structural properties of each aptamer sequence can be imported, if needed, for overlaying when visualizing the clustering data.

Aptamer clustering algorithm

The aptamer clustering component of AptamerRunner generates networks of related aptamers from aptamer sequences and predicted secondary structures. Networks of related aptamers are constructed using either the all-vs-all approach or the seed approach. User options include selecting the 1) *Edge Type* and the 2) *Maximum Distance Measure*.

*Edge Type*: The Edge Type defines the relatedness distance metric applied by the clustering algorithm in order to decide if two aptamers sequences should be connected when building networks of aptamers. Sequence similarity is determined using Levenshtein edit distance [44] and structure similarity is determined using tree distance [45]. These distance measurements are calculated by the clustering algorithm using RNAdistance from the Vienna Package v2.0 [41, 46]. Options for clustering include 1) *edit* to use only edit distance data, 2) *tree* for tree distance data only, 3) *both* to apply the logical operator AND with edit distance and tree distance data, and 4) *or* to apply the logical operator OR with edit and tree distance data. Regardless of Edge Type applied, AptamerRunner will calculate both the edit and tree distances between aptamer sequences and include these data as metadata for the edges connecting two aptamer sequences

(**Supplemental Figure S1B**). This enables users to perform additional analyses using the edit and tree distance values within Cytoscape. For example, if *edit* is designated as the Edge Type option, AptamerRunner will only apply edit distance data when constructing the networks of related aptamers, but it will also calculate tree distance values and include these data as edge metadata.

*Maximum Distance Measure*: The Maximum Distance Measure determines the threshold value of a maximum edit and/or tree distance value for a given Edge Type from which networks of related aptamers will be constructed. A separate Maximum Distance Measure can be set for edit distance and for tree distance. The clustering algorithm requires that the Maximum Distance Measure match the Edge Type and, if using a logical operator, that a Maximum Distance Measure be set for both edit distance and tree distance.

*Output Files:* Three files are outputted by the clustering program: 1) an eXtensible Graph Markup and Modeling Language (XGMML) file containing the aptamer clustering data, which can be imported into Cytoscape; 2) the input FASTA.struct file; and 3) a log file of the commands used to initiate AptamerRunner.

Cytoscape visualization of AptamerRunner clustering results

The XGMML file of clustering results from AptamerRunner can be visualized through Cytoscape [30], an open-source network analysis program (see supplemental methods for specific details). Networks of clustered aptamer sequences are visualized using *nodes*, which represent unique aptamer sequences, and *edges* that connect related aptamer sequence nodes (**Supplemental Figure S6**). Nodes and edges include metadata associated to each aptamer sequence (e.g., name, sequence, structure) and between interconnected aptamers (e.g., edit distance and tree distance values). Additional metadata, such as normalized read counts, fold enrichment or data from scoring algorithms, can be imported into Cytoscape's node data tables from comma or tab delimited text files. These metadata can be used to dictate the visual properties of nodes and edges within Cytoscape to facilitate interpretation of the clustering results. Cystoscope's interactive interface permits easy selection of nodes to isolate potential candidate aptamer sequences. Nodes selected from networks are compiled into Cytoscape's data table along with corresponding node data (**Supplemental Figure S1A**). The Cytoscape data table may be copied directly or exported as a text file. In addition to selecting individual aptamer sequences, groups of aptamers may be readily identified and examined in more detail using network analysis tools such as the clusterMaker algorithm [47]. The clusterMaker algorithm assigns an identification to each group of related

aptamers and appends these identifications to the Cytoscape data table (**Supplemental Figure S1B**).

**NGS datasets:** Cibiel *et. al.* 2014 PLoSOne dataset and Hoinka *et. al.* 2015 NAR were imported into Galaxy from the NCBI SRA database (ERR2121976-91 for Cibiel selection round 0 – 15; SRR3279660 for Hoinka selection rounds 1 – 4 and SRR3279661 for Hoinka selection round 5). For each dataset, Galaxy was used to generate a NrD, conduct a persistence and abundance analysis and filter the NrD. The compiled NrDs for all datasets are available through supplemental. The Cibiel *et. al.* 2014 PLoSOne NrD was filtered for aptamer sequences observed in at least four of the sixteen sequenced selection rounds and with 14 reads, resulting in 6,639 unique aptamer sequences. The published aptamer sequence ACE23 did not meet with criteria being observed in only three selection rounds with maximum read count of one. The ACE23 aptamer sequence was added back into the filtered NrD. The Hoinka *et. al.* 2015 NAR NrD was filtered for aptamer sequences observed in at least one of the four sequence selection rounds with at least 50 reads, resulting in 2,140 unique aptamer sequences.

**Keywords**

Aptamer; Bioinformatics; AptamerRunner; Sequence analysis; Structure analysis; Cytoscape.

**Data availability**

The AptamerRunner .NET script Windows and Linux versions are available in supplemental and on GitHub (https://github.com/ui-icts/aptamer-runner/releases/tag/v0.0.3).

NGS source data[16] used as an example are provided with this paper in supplemental. Further data supporting the findings of this study are available from the corresponding author upon reasonable request.

**Acknowledgments**

Dario Ruiz-Ciancio: Investigation, Formal Analysis, Visualization, Writing – original draft. Suresh Veeramani: Investigation, Formal Analysis, Writing – review & editing. Rahul Singh: Formal Analysis, Writing – original draft. Eric Embree: Software, Resources, Data curation. Chris Ortman: Software, Resources, Supervision. Kristina W. Thiel: Formal Analysis, Writing – original draft. William H. Thiel: Conceptualization, Methodology, Formal Analysis, Project administration, Writing – original draft.

This work was supported by an American Heart Association Scientist Development Grant (14SDG18850071 to WHT), the National Institutes of Health (R01HL139581 and R01HL157956 to WHT; K22CA263783 to KWT), the National Science Foundation (IIS-1817239 to RS), the Department of Defense (DOD CDMRP-PRCRP CA220729 to KWT), the Bunge and Born Fund (FBB-20170609 to DRC) and Fulbright-Argentinian Ministry of Education (ME-FLB-2022-2023 to DRC).

**Conflict of Interest**

Authors declare no conflict of interest.

## References


1. Zhu, C, Feng, Z, Qin, H, Chen, L, Yan, M, Li, L, and Qu, F (2023). Recent progress of SELEX methods for screening nucleic acid aptamers. *Talanta* **266**: 124998.
2. Chauhan, N, Saxena, K, and Jain, U (2022). Single molecule detection; from microscopy to sensors. *Int J Biol Macromol* **209**: 1389-1401.
3. Xie, S, Sun, W, Fu, T, Liu, X, Chen, P, Qiu, L, Qu, F, and Tan, W (2023). Aptamer-Based Targeted Delivery of Functional Nucleic Acids. *J Am Chem Soc* **145**: 7677-7691.
4. Fan, D, Wang, J, Wang, E, and Dong, S (2020). Propelling DNA Computing with Materials' Power: Recent Advancements in Innovative DNA Logic Computing Systems and Smart Bio-Applications. *Adv Sci (Weinh)* **7**: 2001766.
5. Li, L, Xu, S, Yan, H, Li, X, Yazd, HS, Li, X, Huang, T, Cui, C, Jiang, J, and Tan, W (2021). Nucleic Acid Aptamers for Molecular Diagnostics and Therapeutics: Advances and Perspectives. *Angew Chem Int Ed Engl* **60**: 2221-2231.
6. Esposito, CL, Catuogno, S, Condorelli, G, Ungaro, P, and de Franciscis, V (2018). Aptamer Chimeras for Therapeutic Delivery: The Challenging Perspectives. *Genes (Basel)* **9**.
7. Mullard, A (2023). FDA approves second RNA aptamer. *Nat Rev Drug Discov* **22**: 774.
8. Tuerk, C, and Gold, L (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**: 505-510.
9. Ellington, AD, and Szostak, JW (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**: 818-822.
10. DeRosa, MC, Lin, A, Mallikaratchy, P, McConnell, EM, McKeague, M, Patel, R, and Shigdar, S (2023). In vitro selection of aptamers and their applications. *Nature Reviews Methods Primers* **3**: 55.
11. Metzker, ML (2010). Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31-46.
12. Quang, N, Perret, G, and Duconge, F (2016). Applications of High-Throughput Sequencing for In Vitro Selection and Characterization of Aptamers. *Pharmaceuticals (Basel)* **9**.
13. Sun, D, Sun, M, Zhang, J, Lin, X, Zhang, Y, Lin, F, Zhang, P, Yang, C, and Song, J (2022). Computational tools for aptamer identification and optimization. *TrAC Trends in Analytical Chemistry* **157**: 116767.
14. Komarova, N, Barkova, D, and Kuznetsov, A (2020). Implementation of High-Throughput Sequencing (HTS) in Aptamer Selection Technology. *Int J Mol Sci* **21**.
15. Santana-Viera, L, Dassie, JP, Rosas-Lapena, M, Garcia-Monclus, S, Chicon-Bosch, M, Perez-Capo, M, Pozo, LD, Sanchez-Serra, S, Almacellas-Rabaiget, O, Maqueda-Marcos, S*, et al.* (2023). Combination of protein and cell internalization SELEX identifies a potential RNA therapeutic and delivery platform to treat EphA2-expressing tumors. *Mol Ther Nucleic Acids* **32**: 758-772.
16. Ruiz-Ciancio, D, Lin, L-H, Veeramani, S, Barros, MN, Sanchez, D, Di Bartolo, AL, Masone, D, Giangrande, PH, Mestre, MB, and Thiel, WH (2023). Selection of novel cell-internalizing RNA aptamer specific for CD22 antigen in B- Acute Lymphoblastic Leukemia. *Molecular Therapy - Nucleic Acids*.
17. Thiel, WH, Bair, T, Peek, AS, Liu, X, Dassie, J, Stockdale, KR, Behlke, MA, Miller, FJ, Jr., and Giangrande, PH (2012). Rapid identification of cell-specific, internalizing RNA

aptamers with bioinformatics analyses of a cell-based aptamer selection. *PLoS One* **7**: e43836.

18. Ishida, R, Adachi, T, Yokota, A, Yoshihara, H, Aoki, K, Nakamura, Y, and Hamada, M (2020). RaptRanker: in silico RNA aptamer selection from HT-SELEX experiment based on local sequence and structure information. *Nucleic Acids Research* **48**: e82-e82.
19. Jiang, P, Meyer, S, Hou, Z, Propson, NE, Soh, HT, Thomson, JA, and Stewart, R (2014). MPBind: a Meta-motif-based statistical framework and pipeline to Predict Binding potential of SELEX-derived aptamers. *Bioinformatics* **30**: 2665-2667.
20. Thompson, JD, Higgins, DG, and Gibson, TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-4680.
21. Edgar, RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.
22. Bayrac, AT, Sefah, K, Parekh, P, Bayrac, C, Gulbakan, B, Oktem, HA, and Tan, W (2011). In vitro Selection of DNA Aptamers to Glioblastoma Multiforme. *ACS Chem Neurosci* **2**: 175-181.
23. Hoinka, J, Berezhnoy, A, Sauna, ZE, Gilboa, E, and Przytycka, TM (2014). AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application. *Res Comput Mol Biol* **8394**: 115-128.
24. Hoinka, J, Berezhnoy, A, Dao, P, Sauna, ZE, Gilboa, E, and Przytycka, TM (2015). Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Res*.
25. Alam, KK, Chang, JL, and Burke, DH (2015). FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. *Mol Ther Nucleic Acids* **4**: e230.
26. Kramer, ST, Gruenke, PR, Alam, KK, Xu, D, and Burke, DH (2022). FASTAptameR 2.0: A web tool for combinatorial sequence selections. *Mol Ther Nucleic Acids* **29**: 862-870.
27. Singh, R, Yang, H, Dalziel, B, Asarnow, D, Murad, W, Foote, D, Gormley, M, Stillman, J, and Fisher, S (2013). Towards human-computer synergetic analysis of large-scale biological data. *BMC Bioinformatics* **14 Suppl 14**: S10.
28. Singh, R, and Jain, R (2006). From Information-Centric to Experiential Environments. In: Goldin, D, Smolka, SA and Wegner, P (eds). *Interactive Computation: The New Paradigm*. Springer Berlin Heidelberg: Berlin, Heidelberg. pp 323-351.
29. Boettiger, C (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review* **49**: 71-79.
30. Shannon, P, Markiel, A, Ozier, O, Baliga, NS, Wang, JT, Ramage, D, Amin, N, Schwikowski, B, and Ideker, T (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498-2504.
31. Hoinka, J, Berezhnoy, A, Sauna, ZE, Gilboa, E, and Przytycka, TM (2014). AptaCluster – A Method to Cluster HT-SELEX Aptamer Pools and Lessons from Its Application. In: Sharan, R (ed). *Research in Computational Molecular Biology*, vol. 8394. Springer International Publishing: Pittsburgh PA. pp 115-128.

32. Hoinka, J, Backofen, R, and Przytycka, TM (2018). AptaSUITE: A Full-Featured Bioinformatics Framework for the Comprehensive Analysis of Aptamers from HT-SELEX Experiments. *Mol Ther Nucleic Acids* **11**: 515-517.
33. Dao, P, Hoinka, J, Takahashi, M, Zhou, J, Ho, M, Wang, Y, Costa, F, Rossi, JJ, Backofen, R, Burnett, J*, et al.* (2016). AptaTRACE Elucidates RNA Sequence-Structure Motifs from Selection Trends in HT-SELEX Experiments. *Cell Syst* **3**: 62-70.
34. Levay, A, Brenneman, R, Hoinka, J, Sant, D, Cardone, M, Trinchieri, G, Przytycka, TM, and Berezhnoy, A (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.
35. Ducrot, C, and Piffoux, M (2023). Combining independent protein and cellular SELEX with bioinformatic analysis may allow high affinity aptamer hit discovery. *Molecular Therapy - Nucleic Acids* **33**: 254-256.
36. Zuker, M (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406-3415.
37. Thiel, WH (2016). Galaxy Workflows for Web-based Bioinformatics Analysis of Aptamer High-throughput Sequencing Data. *Mol Ther Nucleic Acids* **5**: e345.
38. Caroli, J, Forcato, M, and Bicciato, S (2020). APTANI2: update of aptamer selection through sequence-structure analysis. *Bioinformatics* **36**: 2266-2268.
39. Hoinka, J, and Przytycka, T (2016). AptaPLEX - A dedicated, multithreaded demultiplexer for HT-SELEX data. *Methods* **106**: 82-85.
40. Shieh, KR, Kratschmer, C, Maier, KE, Greally, JM, Levy, M, and Golden, A (2020). AptCompare: optimized de novo motif discovery of RNA aptamers via HTS-SELEX. *Bioinformatics* **36**: 2905-2906.
41. Lorenz, R, Bernhart, SH, Honer Zu Siederdissen, C, Tafer, H, Flamm, C, Stadler, PF, and Hofacker, IL (2011). ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.
42. Mathews, DH, Sabina, J, Zuker, M, and Turner, DH (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**: 911-940.
43. Walter, AE, Turner, DH, Kim, J, Lyttle, MH, Müller, P, Mathews, DH, and Zuker, M (1994). Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad USA* **91**: 9218-9222.
44. VI, L (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov Phys Dokl* **10**: 707-710.
45. Fontana, W, Konings, DA, Stadler, PF, and Schuster, P (1993). Statistics of RNA secondary structures. *Biopolymers* **33**: 1389-1404.
46. Hofacker, IL, Fontana, W, Stadler, PF, Bonhoeffer, LS, Tacker, M, and Schuster, P (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte fr Chemie Chemical Monthly* **125**: 167-188.
47. Morris, JH, Apeltsin, L, Newman, AM, Baumbach, J, Wittkop, T, Su, G, Bader, GD, and Ferrin, TE (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics* **12**: 436.

List of Figures Captions

**Figure 1. AptamerRunner, a structure prediction and clustering algorithm a to visualize selected aptamers.** AptamerRunner consist of two independent algorithms, a secondary structure prediction algorithm and an aptamer clustering algorithm. (**A**) AptamerRunner is a .NET bash script coded in C#. The AptamerRunner .NET script communicates with the Docker repository to download the latest version of the AptamerRunner Docker image and then initiates the AptamerRunner Docker image within the Docker platform as a Docker container. The Docker container includes the AptamerRunner structure prediction algorithm and the clustering algorithm with all dependencies. Once either algorithm has finished processing any user commands and output results, the AptamerRunner .NET script shuts down the AptamerRunner Docker container. (**B**) The secondary structure prediction algorithm utilizes collapsed aptamer NGS data in FASTA format to predict the secondary structure of a full-length aptamer using RNAfold. The secondary structure prediction algorithm has the option to append the constant region sequence if needed. Output includes a modified FASTA-formatted file with a third line containing the dot-bracket annotation of each predicted structure and a properties file that includes information about the predicted structures (e.g., minimum free energy). (**C**) The AptamerRunner clustering algorithm uses the FASTA-formatted file with the predicted structures to generate networks of related aptamers using options selected by the user for the *Clustering Method*, the *Edge Type,* and the *Display Threshold*. Output files include the clustering results compiled into a XGMML file, which is visualized using Cytoscape, a log file, and the input file.

**Figure 2: AptamerRunner all-vs-all and seed clustering approaches using either edit distance or tree distance.**

AptamerRunner clustering aptamer NGS data using different clustering methods with either edit distance 1 or tree distance 0. (**A**) Seed approach with edit distance 1 or (**B**) tree distance 0. (**C**) All-vs-all approach with edit distance 1 or (**D**) tree distance 0. Ruiz-Ciancio *et. al.* [16] data was used as example data for clustering and clustering results were visualized using Cytoscape (v 2.8.1).

**Figure 3: Use of logical operators (AND, OR) with AptamerRunner clustering.**

Clustering using logical operator AND with edit distance 1 and tree distance 0 with the (**A**) seed approach and the (**B**) all-vs-all approach. Clustering using the logical operator OR with edit

distance 1 and tree distance 0 with the (**C**) seed approach and the (**D**) all-vs-all approach. Ruiz-Ciancio *et. al.* [16] data was used as example data for aptamer classification and selection, and data are visualized using Cytoscape (v 2.8.1).

**Figure 4: Clustering results from FASTAptamer, FASTAptameR 2.0 and AptaCluster.**
(**A**) The modified FASTA file clustering results from FASTAptamer and FASTAptameR 2.0. The FASTA header information follows as rank, reads, reads per million (RPM), cluster id, rank within the cluster, and edit distance from the seed sequence. (**B**) FASTAptameR 2.0 clustering results outputted as a data table. (**C**) AptaCluser exported clustering results from round 9 and (**D**) exported cluster table. Data shown are only a portion of larger files. Gaps in data are denoted by a "…".

**Figure 5: Visual integration of round-to-round log2 fold enrichment data mapped onto clustered aptamers.**
The log2 fold enrichment data between selection rounds from a SELEX process imported into the Cytoscape data table can be applied to determine the visual properties of aptamers clustered by AptamerRunner. The node size was set to the log10 normalized read count of round 9 and node color was determined by the log2 fold enrichment between different selection rounds based on (**A**) the phases of the sequence enrichment % sigmoidal curve fit; exponential (R0:R2), linear (R2:R5) and asymptotic (R5:R9). (**B**) The log2 fold change enrichment of round 0 to round 2 shows early linear phase enrichment of aptamer sequences, (**C**) round 2 to round 5 shows changes in aptamer sequences during the linear phase of sequence enrichment when the aptamer library experienced the greatest changes in convergence and (**D**) round 5 to round 9 show changes in aptamer sequences during the asymptotic phase when the aptamer library had reached maximum convergence.

**Figure 6: Aptamer scoring algorithm mapped to clustered aptamers.**
Data from aptamer scoring algorithms (**A**) MPBind and (**B**) RaptRanker were mapped to the networks generated by AptamerRunner. (**C**) Cytoscape can produce scatter plots that can compared the scoring data from MPBind and RaptRanker to identify aptamers and clusters of aptamers that were scored highly by both algorithms.

**Figure 7: Re-clustering aptamer networks within Cytoscape**

Groups of aptamers can be re-clustered using tools within Cytoscape. (**A**) Nodes within a large hairball network of aptamers clustered by edit distance 1 using the all-vs-all approach were selected and (**B**) re-organized within Cytoscape by identical structures.

**Figure 8: AptamerRunner analysis of IL-10 aptamers**

(**A**) AptamerRunner edit distance 5 using seed approach and (**B**) all-vs-all approach. (**C**) Tree distance 10 using the all-vs-all approach. (**D**) Edit distance 5 and tree distance 5. Clusters are labeled as reported in Levay *et. al.*[34] * indicates cluster found by AptamerRunner using edit distance 5 with the seed approach that was not identified by AptaCluster.

770 Table 1

| | AptamerRunner | FASTAptamer | FASTAptameR 2.0 | AptaCluster |
|---|---|---|---|---|
| **Interface** | Command line C# .NET script operating a Docker container | Command line Perl-based programs | Offline web browser accessing a Docker container | AptaSuite Java application |
| **Components** (*clustering) | • Secondary structure prediction<br>• Clustering algorithm* | • Count<br>• Compare<br>• Cluster*<br>• Enrich<br>• Search | • FASTAptamer components<br>• Cluster module*<br>• Edit distance module<br>• Motif discovery module<br>• Position enrich module | • AptaGUI<br>• AptaPlex<br>• AptaSim<br>• AptaCluster*<br>• AptaTrace<br>• AptaMut |
| **Third party software** | Docker, Cytoscape | none | Docker, web browser | Java |
| **Clustering approach** | seed, all-vs-all | seed | seed | seed |
| **Distance measures** | • Edit distance<br>• Tree distance<br>• Logical operators (AND, OR) | • Edit distance | • Edit distance | • Edit distance |
| **Visualization** | Cytoscape display of networked aptamers (nodes) connected by distance measures (edges) | N/A | • Cluster metaplots<br>• k-mer PCA | N/A |
| **Metadata** | Nodes:<br>• Aptamer name<br>• Predicted structure<br>• User-imported<br>Edges:<br>• Edit distance values<br>• Tree distance values<br>• User-imported | N/A | N/A | N/A |

771