BIOINFORMATICS

Vol. 28 ISMB 2012, pages i215–i223
doi:10.1093/bioinformatics/bts210

Identification of sequence–structure RNA binding motifs for SELEX-derived aptamers

Jan Hoinka[1,†,‡], Elena Zotenko[2,†,‡], Adam Friedman[3], Zuben E. Sauna[3,*] and Teresa M. Przytycka[1,*]

[1]National Center for Biotechnology Information, NLM, NIH, 8600 Rockville Pike, Bethesda, MD 20894, USA, [2]Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010, Australia and [3]Laboratory of Hemostasis, Division of Hematology, Center for Biologics Evaluation and Research, US Food and Drug Administration, Bethesda, MD 20892, USA

ABSTRACT

Motivation: Systematic Evolution of Ligands by EXponential Enrichment (SELEX) represents a state-of-the-art technology to isolate single-stranded (ribo)nucleic acid fragments, named aptamers, which bind to a molecule (or molecules) of interest via specific structural regions induced by their sequence-dependent fold. This powerful method has applications in designing protein inhibitors, molecular detection systems, therapeutic drugs and antibody replacement among others. However, full understanding and consequently optimal utilization of the process has lagged behind its wide application due to the lack of dedicated computational approaches. At the same time, the combination of SELEX with novel sequencing technologies is beginning to provide the data that will allow the examination of a variety of properties of the selection process.

Results: To close this gap we developed, Aptamotif, a computational method for the identification of sequence–structure motifs in SELEX-derived aptamers. To increase the chances of identifying functional motifs, Aptamotif uses an ensemble-based approach. We validated the method using two published aptamer datasets containing experimentally determined motifs of increasing complexity. We were able to recreate the author's findings to a high degree, thus proving the capability of our approach to identify binding motifs in SELEX data. Additionally, using our new experimental dataset, we illustrate the application of Aptamotif to elucidate several properties of the selection process.

Contact: przytyck@ncbi.nlm.nih.gov, Zuben.Sauna@fda.hhs.gov

1 INTRODUCTION

Aptamers are synthetic but biologically active single-stranded (ribo)nucleic molecules, typically ranging between 15 and 120 nt (James, 2000). These short sequences can be designed to bind, with high affinity and specificity, a vast spectrum of molecular targets (apatopes), spanning from small organic molecules (Barrick and Breaker, 2007; Lozupone *et al.*, 2003) over macromolecules such as proteins (Dobbelstein and Shenk, 1995; Kim *et al.*, 2011) to entire organisms (Li *et al.*, 2011). Their high structural stability over a wide range of pH and temperatures and their diverse functionality make aptamers ideal candidates for a broad spectrum of *in vitro* assays and *in vivo* tools. Applications for aptamers include protein inhibition and purification (Esposito *et al.*, 2011; Walker *et al.*, 2011), molecular detection systems (Nielsen *et al.*, 2010; Zelada-Guilln *et al.*, 2010), therapeutic drugs (Wang *et al.*, 2011) and antibody replacement (Bunka *et al.*, 2010; Ni *et al.*, 2011); the latter being of high interest in the pharmaceutical industry due to substantially lower production costs, shelf lives of years and, in many cases, higher target specificity (Kupakuwana *et al.*, 2011). Moreover, since aptamers are chemically synthesized, they provide a more consistent source of material than antibodies that are secreted by cells.

Aptamers targeting a specific apatope are experimentally identified through the Systematic Evolution of Ligands by EXponential Enrichment (SELEX) protocol (Tuerk and Gold, 1990). The experimental design of SELEX is based on the assumption that a large enough pool of candidate sequences is likely to contain nucleotide strands capable of binding to any target molecule. The traditional SELEX procedure iterates over five basic steps defining one selection cycle: incubation, binding, partitioning and washing, target-bound elution, and amplification. Starting with a single-stranded (ribo)nucleic acid sequence library of, typically, 10^{15} random species flanked by primer sites to aid amplification, at each cycle a sequence pool is incubated with target molecules. The species in the pool potentially bind the target with specificity, depending on their sequence and structure. At the end of each cycle, low-affinity binders are removed from the solution whereas bound species are eluted and amplified, forming the input for the next round. Eventually, only molecules that bind with high affinity remain.

However, detailed understanding and studies regarding precise aptamer–apatope interaction and consequently the evolutionary processes induced by this interplay lag behind the wide application of SELEX products. Aptamer binding properties are a function of both, sequence and structure; however, current analyses are predominately focused on sequence. Until now, a common practice was to sequence a sample of aptamers from the last cycle of SELEX and examine them for possible sequence motifs using *ad hoc* approaches or motif finding programs, such as MEME (Bailey and Elkan, 1994) or GLAM2 (Frith *et al.*, 2008). These motifs are then validated in expensive and time-consuming wet lab experiments, frequently by introducing a series of point mutations into the identified sequence regions with the goal of inducing conformational changes in the structures. Using adequate binding affinity assays then allows for the quantification of this mutation with respect to the binding strength, considering the region as functionally active,

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[‡]The study performed in part when these authors were at Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Campus E1.4, 66123 Saarbrücken, Germany

*To whom correspondence should be addressed.

i.e. part of the binding motif, if a significant decrease is measured (Dubey *et al.*, 2005; Lozupone *et al.*, 2003; Lvesque *et al.*, 2007; Yamamoto *et al.*, 2000). Although, in addition to the sequence-based analysis, secondary structures predicted by programs such as MFold (Zuker, 1989) are manually compared (e.g. Lozupone *et al.*, 2003) when the structural properties are neglected during motif identification. However, for aptamers targeting proteins, the binding properties are largely determined by how well these molecules fit into the cavities of their apatope counterparts, and are hence strongly dependent on their shape. Thus, understanding of the structural restrictions imposed on binders is crucial in aptamer studies. In addition, comparing properties of motifs originating from subsequent cycles might then represent the first step toward an enhanced understanding of the SELEX process and its more full utilization. Finally, new sequencing technologies have opened the doors to sequencing of complete pools of aptamers. This provides an unprecedented opportunity to address many selection-related questions but demands appropriate computational tools.

Can we adopt solutions, which have been developed with other applications in mind, to uncover and study aptamer motifs? One promising tool is MEMERIS (Hiller *et al.*, 2006) (MEME for RNAs including Secondary Structures), which is specifically designed for searching sequence motifs in a set of RNA sequences and simultaneously integrating information about secondary structures. MEMERIS uses secondary structure information to bias its search toward substrings that tend to reside in single-stranded regions (Hiller *et al.*, 2006). A similar approach is used in the recently published RNAcontext program developed to predict binding affinities of RNA molecules to a given RNA binding protein (Kazan *et al.*, 2010). RNAcontext fits a statistical model that incorporates both sequence and secondary structure information to the input pool of sequences and their experimentally measured binding affinity values. The fitted models are then interpreted to extract binding motifs and their preferred secondary structure. RNAprofile (Pavesi *et al.*, 2004) on the other hand, first identifies candidate regions containing a prescribed number of hairpins, and then performs all-by-all pairwise alignment of candidate regions that are later clustered to uncover common motifs. Finally, several methods exist for pairwise alignment of RNA molecules that utilize both, sequence and secondary structure information. Methods, such as RNAForester (Höchsmann *et al.*, 2003) and ExpaRNA (Heyne *et al.*, 2009) assume that the secondary structure is known and is provided as input. Other methods, such as LocARNA (Will *et al.*, 2007) and its extension LocARNATE (Otto *et al.*, 2008) do not require secondary structure to be known in advance and compute the alignment and corresponding consensus secondary structure using a variation of the Sankoff fold-align algorithm (Sankoff, 1985).

Unfortunately, these approaches are not fully compatible with the requirements imposed on aptamer motif identification procedures. First, aptamer motifs, while connected in secondary structure, are often discontinuous in sequence, which limits the applicability of MEMERIS. Next, we lack any prior knowledge about possible conformations of the aptamers, rendering RNAprofile-type preprocessing useless in this context. Additionally, it has been repeatedly realized that the minimum free energy (MFE) structure is not necessarily the only structure assumed by RNA molecules. This is even more important in the context of aptamers, where the MFE structure would have to be computed independently of its apatope counterpart. Last but not the least, any approach that requires all-by-all alignment of the input sequences (local or global) cannot scale up to deal with the number of species resulting from sequencing entire aptamer pools.

Here, we report Aptamotif, an ensemble-based method for effective extraction of sequence–structure motifs from SELEX-derived aptamers. Building on the broad success of ensemble-based approaches, Aptamotif considers optimal and suboptimal structures within the proximity of the MFE structure. We represent each aptamer by its functional space, i.e. the set of substructures that might potentially undergo binding interaction with the target molecule. These substructure ensembles are then concurrently compared to identify candidates with strong common features in terms of sequence, by aligning the underlying primary structure of the substructures, and structure, by restricting these alignments to compatible substructures with similar shape. Our approach hence takes into account both, sequence variability and indels as well as structural fluctuations with respect to small loop size alterations.

Aptamotif is well suited for identifying functional sites in their native configuration and in capturing relationships between binders selected in consecutive SELEX cycles. While currently implemented with the aim of providing the proof of principle in the context of traditional, small-sized, sampled SELEX data, Aptamotif can be adopted to deal with large datasets that will result from sequencing whole aptamer pools using next-generation sequencing technology, which we are currently exploring.

2 RESULTS AND DISCUSSION

In this section, we present Aptamotif, a novel computational approach for aptamer motif identification. We begin with a high-level description of the algorithm. We then test Aptamotif on two publicly available SELEX datasets of different complexity for which binding sites where determined experimentally. We demonstrate the benefits of our ensemble approach and show that motifs identified by Aptamotif are in better agreement with published binding sites than those identified by two widely used RNA motif extraction programs. Next, we apply our approach to study selection properties of the SELEX protocol. For this purpose, we experimentally generated a new dataset sampling two consecutive SELEX cycles. Our results show variability of sequence–structure motifs in the SELEX cycles and the utility of our approach for studying the evolution process in the SELEX procedure.

2.1 Algorithm outline

Secondary structure contributes to the stability and biochemical properties of the RNA molecule such as its affinity toward its interacting partners. RNA secondary structure is defined by a specific pairing pattern between nucleotides or base pairing (Zuker and Sankoff, 1984). For the purpose of the discussion below, it is important to mention that every secondary structure can be uniquely decomposed into a set of basic non overlapping substructures as shown in Figure 1. Stacking pairs contain nucleotides that participate in base pairing, whereas loops consist of closing and interior base pairs together with nucleotides that do not participate in such contacts and therefore comprise single-stranded regions of the structure. Loops are further classified into four different types: (i) hairpin loops; (ii) bulge loops; (iii) interior loops; and (iv) multibranch loops.

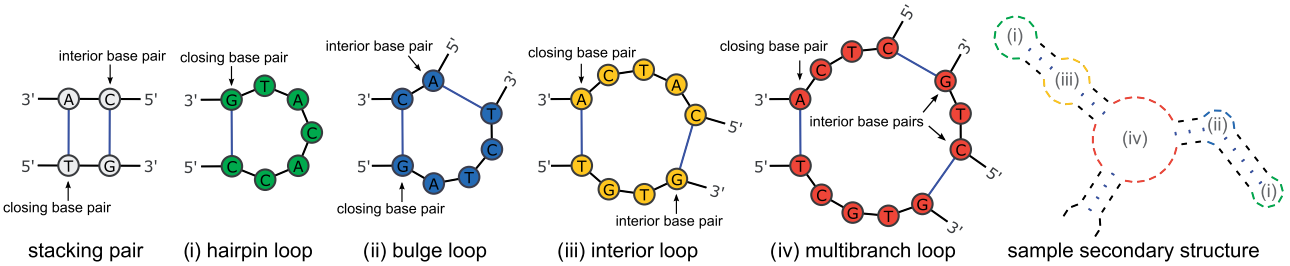

Fig. 1. Components of RNA secondary structure. In this study, bulge loops and interior loops are treated individually. Despite the fact that one is a degenerate version of the other, they represent substructures of different flexibility. Multibranch loops are limited to three components. Due to generally short length of the aptamers, loops with more than three branches are highly unlikely. At present, substructures of known aptamers fall into one of the categories (i)–(iv), but we might consider extending our current loop treatment as additional data becomes available and such need emerges. Secondary structure representations were in part created using VARNA (Darty *et al.*, 2009)

We define an aptamer motif as a loop substructure present in a large fraction of aptamers and showing statistically significant sequence similarity in its single-stranded regions. There is growing evidence that the biological function of many RNA molecules crucially depends on single-stranded regions (Schudoma *et al.*, 2010). This seems to be especially true for RNA molecules evolved by the SELEX protocol. In fact, the majority of binding sites reported in manually curated databases of SELEX experiments (Lee *et al.*, 2004) reside in loops.

Given a pool of aptamer specie, our algorithm uses a three-step procedure to identify sequence–structure motifs. For each aptamer, optimal and suboptimal secondary structures are computed and loops with their sequences are extracted in the *structural processing step*. The *seed identification step* then proceeds in iterations, where at each iteration K (a user defined parameter), aptamers are sampled uniformly at random from the pool. The loops of matching type extracted from these aptamers are then aligned and scored. A small number of iterations are performed and the best-scoring loop alignments are retained and searched against the entire pool of aptamers in the *seed extension step*. In what follows we elaborate on and justify important aspects of these three steps.

Structural processing: we use the *RNAsubopt* algorithm (Backofen and Siebert, 2007; Wuchty *et al.*, 1999) to enumerate all secondary structures within a user-defined energy range from the MFE structure, the base pair configuration with the smallest attainable folding free energy, for every aptamer sequence in the pool. Primer sites are included in the prediction as these tend to influence the folding properties of the species *in vitro* and must therefore also be considered *in silico*. The generated secondary structures are then processed to extract the loops and their sequences. Thus each aptamer is associated with a substructure ensemble, a collection of unique substructures that are divided into four non overlapping subsets one for each loop type (hairpin, bulge, interior and multibranch). Here, all loops partially or entirely located in primer sites are excluded from the substructures ensemble and hence not further considered.

It has been repeatedly shown that biologically active conformations of non coding RNAs do frequently not correlate with the MFE structure but are rather present as structures with slightly higher free energy values (Ding *et al.*, 2005; Zuker, 1989). We expect this discrepancy to be even more prominent for RNA molecules resulting from SELEX experiments as biologically active conformations of aptamers are thought to be stabilized through interactions with their binding partners. It is important, therefore, to set δ, a user-defined parameter that defines energy range for secondary structure enumeration, high enough to guarantee inclusion of biologically active conformations in the analysis. Extremely high values of δ, however, will substantially increase the size of the substructure ensemble and thus increase computational time. We found that for aptamer datasets analyzed in this article, the value of 3 kcal/mol provides a good tradeoff between sensitivity and computational complexity.

Seed identification: the brute-force approach for identifying common sequence–structure motifs involves an exhaustive comparison of substructure ensembles. This, however, would be prohibitively expensive even for small problem sizes both in terms of average aptamer size and the number of aptamers in the pool. Assuming R aptamer sequences and denoting by N_k, the size of the substructure ensemble associated with aptamer k, the number of comparisons involved is on the order of $O\left(\prod_{i=1}^{R} N_i\right)$.

In order to reduce the problem to a manageable size, we rely on the following key observation: if the motif is present in a large enough fraction of aptamers, then K aptamers selected uniformly at random from the pool will contain the motif with a non negligible probability. In practice, final SELEX cycles are expected to contain at least 50% of highly target-specific binders (Bowser, 2005), in most of the cases however, these percentages, denoted as f in the following, are much larger . Accordingly, our algorithm performs a small number of sampling iterations where at each iteration K aptamers are sampled from the pool and their substructure ensembles are exhaustively aligned and scored. The probability of success of one iteration to draw a sample containing K species with a common motif is therefore f^K, meaning that on average every $\frac{1}{f}^K$ sampling step will retrieve at least one motif.

Hence, given K aptamers, our method performs all possible alignments of K substructures such that each substructure is selected from a different substructure ensemble. Only substructures of the same type are aligned i.e. hairpins are aligned with hairpins, internal loops with internal loops, etc. Substructure alignment involves multiple sequence alignment (MSA) of the single-stranded regions of the loops as shown in Figure 2 in the case of internal loops. In order to substantially reduce the number of MSA computations, we exclude all combinations with a length difference of more

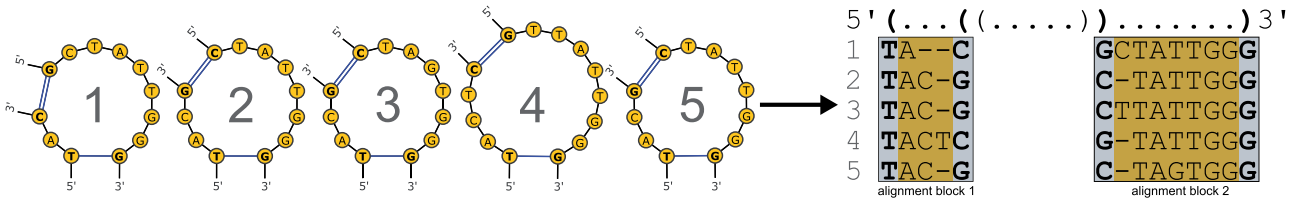

Fig. 2. Example of one substructure alignment step comprising a combination of five interior loops from different substructure ensembles, which are processed by individually aligning the single-stranded regions in 5′–3′ direction. Note that closing and interior base pairs (indicated in bold) are excluded from the alignment since these are assumed to contribute structural stability

than G bases between the aligned regions. In our test scenarios setting, $G=2$ has proven to eliminate up to 70% of all substructure combinations without noticeable impact onto the quality of the resulting motifs. Other, more sophisticated procedures for reducing the number of MSAs are possible but not implemented in the current version of the algorithm.

Each MSA is scored by its degree of conservation. By treating the individual alignment components as a single block of K sequences and L columns over the alphabet Ω, the conservation is described as the sum I over the column wise information content $I(i)$

$$I=\sum_{i=1}^{L} I(i), \qquad I(i)=\sum_{j=1}^{|\Omega|} n_{ij}\log\left(\frac{n_{ij}/K}{b_j}\right) \tag{1}$$

where n_{ij} denotes the number of occurrences of the j-th letter in the i-th column and b_j refers to the background frequency of the j-th letter (Hertz and Stormo, 1999). While this measure allows ranking alignments of equal dimension, it is inapplicable for comparisons between alignments of varying K and L. In order to overcome this problem, we assign a P-value to each MSA. For any alignment with score I_0, the P-value indicates the probability $p(I_x \geq I_0)$ of observing a score I_x equal to or better than I_0, under the assumption that each of the alignment columns have been sampled independently according to a certain background distribution. Naively solving this problem requires traversing all $I \geq I_0$, which is prohibitively expensive. We therefore use the currently fastest and most accurate algorithm for P-value approximation, by Nagarajan and Keich, 2008; Nagarajan *et al.*, 2005.

Special attention must be paid to the role of gaps. We introduce an additional letter '-' into the alphabet and modify the background frequencies to reflect high abundance of this artificial letter. Consequently, columns with high gap ratios contain less information as compared to columns with highly conserved letters resulting in a decreased score.

Sorting the combinations by increasing P-values establishes a ranking order in which top scoring candidates correspond to substructures with high sequence and structure similarity. We denote these potential motifs as seeds since at this point their frequency of occurrence in the aptamer pool remains unknown, and retain the set of P (where P is a user defined parameter) best scoring alignments at each iteration.

Seed extension: in order to assess the degree of abundance of the seeds in the aptamer pool, each seed is 'extended' by searching for approximate sequence–structure matches against the remaining aptamer ensembles. To account for nucleotide variations, we opted for a profile matching-oriented approach in which each of the seed's alignment block is first converted into a position-specific scoring matrix (PSSM). These profiles are then matched against the corresponding single-stranded regions of all loops of equal type in the pool. We use local profile sequence alignments in order to account for possible gaps in the motif instances within individual aptamers. Consequently, spurious seeds not shared by a large fraction of aptamers are excluded, whereas those still persisting after this procedure are considered motifs and reported to the user.

2.2 Recreating experimentally determined motifs

We applied our method onto two SELEX datasets in an attempt to recreate the experimentally determined motifs *in silico*. In traditional aptamer research, only top binders or their refined versions are commonly published, leaving only a limited number of publications additionally containing the set of raw aptamer sequences. However, we have been able to identify two reasonably large datasets from this limited repertoire as representatives for aptamer classes of increasing complexity regarding sequence continuity and shape of the motif. In each experiment we generated all suboptimal structures within a region of 3 kcal/mol from the MFE structure and performed 50 iterations of substructure ensemble alignments, choosing 5 ensembles in each round. The maximal length difference between substructure components was set to $G=2$.

Our first validation dataset was published by Dobbelstein and Shenk in which aptamers were selected against the ribosomal protein L22 (Dobbelstein and Shenk, 1995). The RNA library consists of a pool of 16 aptamers containing a 30 nt, initially randomized region flanked by primer binding sites of length 24 nt (5′) and 23 nt (3′), respectively. Dobbelstein and Shenk identified a highly conserved stem–loop structure as the functional site of the L22 binding aptamers. The hairpin consists of 6–9 nts, with the most 3′ nt being thymine in most of the species and a (5′)G-C(3′) configuration forming the closing base pair. This scenario represents a motif of intermediate complexity since the motif is continuous in primary structure but it contains a large number of insertions, deletions and variations in sequence.

The first five top scoring motifs reported by Aptamotif together with the *in vitro* derived gold standard are listed in Figure 3A and are in good agreement with the author's findings. All motifs are correctly predicted as hairpins and show high consistency between each other, indicating that these likely represent instances of the same functional site. Furthermore, each hit matches the

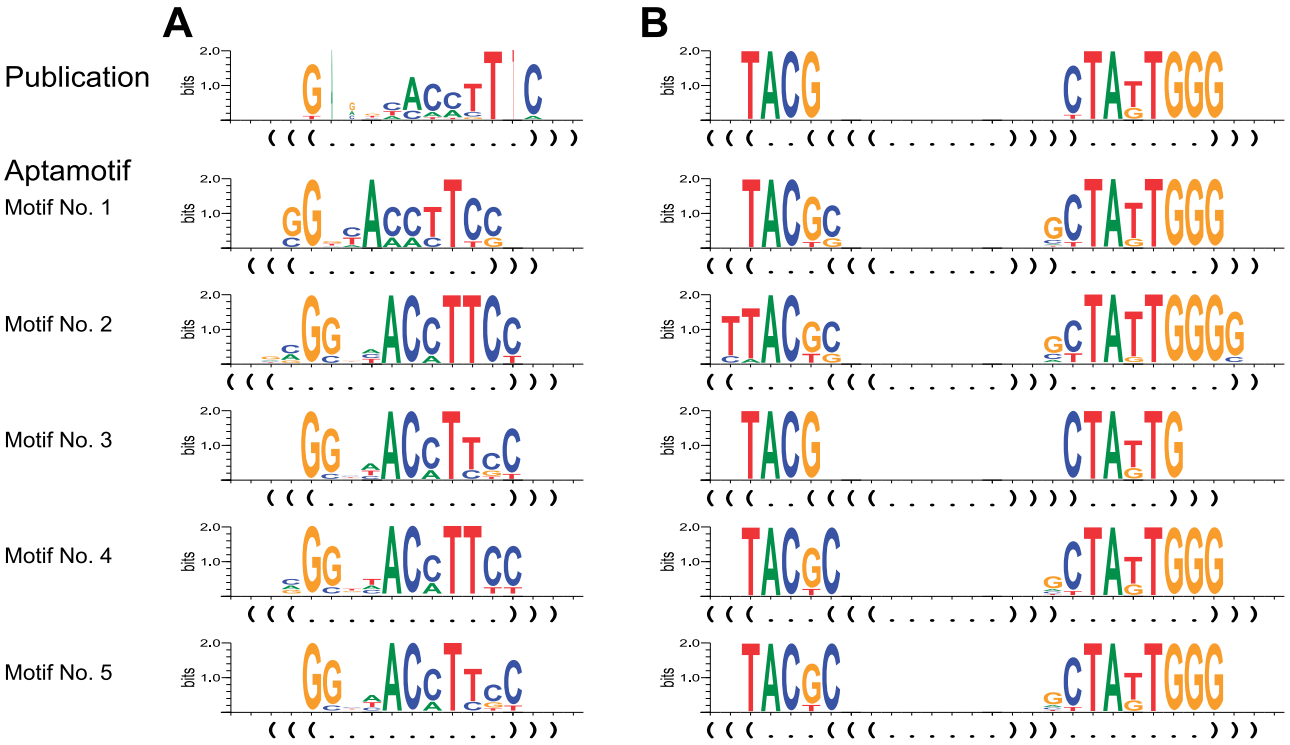

Fig. 3. First five top scoring motifs for the datasets of (**A**) Dobbelstein and Shenk (1995) and (**B**) Lozupone *et al*. (2003) identified by Aptamotif. Motifs are drawn as sequence logos with corresponding secondary structure in dot-bracket notation below. Secondary structures without sequence logos correspond to the consensus structures of adjacent regions to the motifs. The thinness of a letter denotes the likelihood for gap insertions. First row: experimentally determined motifs as described in corresponding publications

experimentally determined motif to a large extent in terms of motif size and location on the individual aptamers. Although our results suggest a slightly larger loop size that includes the reported (5′)G-C(3′) base pair as single-stranded nucleotides, we also found the closing base pair to have a strong tendency toward a combination of guanine and cytosine, indicating that this constellation might play an important role in the structural stability of the hairpin. Further inspection revealed a total of two conserved thymines at the 3′-end of the loop structure, the latter corresponding to the thymine identified by Dobbelstein and Shenk. Our method additionally reported a highly conserved adenine near the hairpins center, not mentioned in the publication, which might contribute to the biological activity of the aptamers.

Summarizing, the reported motifs are in good agreement with the published structures, hence demonstrating the flexibility of our approach to account for nucleotide modifications, while maintaining the capability of extracting important structural features responsible for the biological fitness of the aptamers.

In order to test the full potential of our method, as our second test case, we selected a set of aptamers binding the amino acid isoleucine, published by Lozupone *et al*. (2003). In this case, the biologically active site was identified in an asymmetric interior loop, thus representing a scenario with non continuous primary structure. The motif comprises two sequence modules with `TACG` and `CTATTGGGG` as the consensus sequences for modules 1 and 2, respectively. Within the internal loop, all nucleotides show absolute conservation except for the second thymine in module 2. Furthermore, the base pairs closing the loop structure are also highly conserved.

Figure 3B demonstrates the first five top scoring motifs identified by Aptamotif as well as the structural properties and sequence conservation of the published active site. All results are correctly identified as interior loop structures and associated with the motif reported by the authors, however, presenting slight variations in the module sizes. The less conserved second thymine in module 2 is also captured by each motif of our approach. The variability in the number of nucleotides in each module might be explained by the fact that our method relies on the comparison of entire structure ensembles and is therefore more sensitive to statistical variations. We were hence able to recreate the results published by Lozupone *et al*. in the majority of the points, suggesting that our algorithm is capable of extracting the biologically relevant motif from a pool of aptamer sequences.

2.3 Assessing the significance of ensemble utilization

In order to provide evidence for the necessity of ensemble-based motif identification, we repeated the experiments described in Section 2.2 predicting only the MFE structure for each aptamer. Figure 4 depicts the first five top scoring motifs for the datasets of Dobbelstein and Shenk (Fig. 4A) and Lozupone *et al*. (Fig. 4B), respectively.

In comparison to the ensemble-based results shown in Figure 3, the motifs identified in the first dataset (A) describe the native

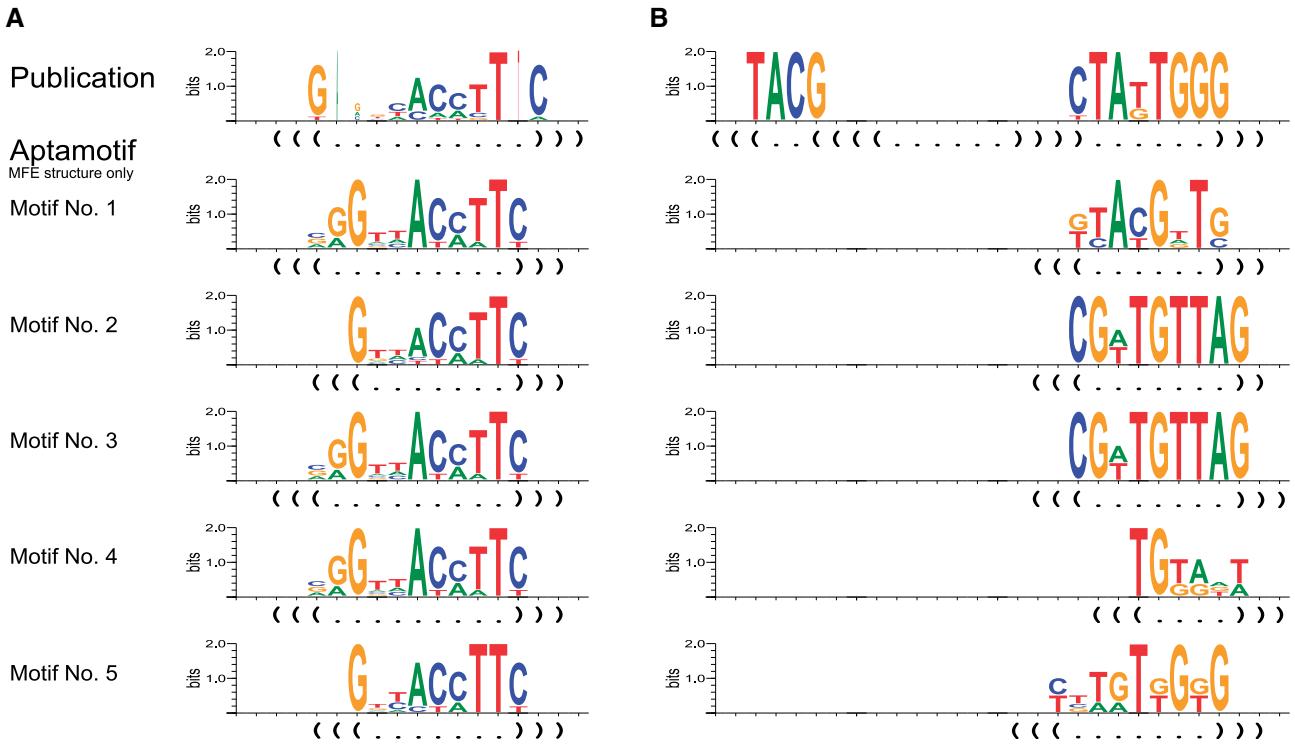

Fig. 4. First five top scoring motifs for the datasets of (**A**) Dobbelstein and Shenk (1995) and (**B**) Lozupone *et al*. (2003) identified by Aptamotif using only MFE structures. Notations as in Figure 3

binding site (hairpin) with only relative accuracy whereas the results from the binding site residing in the inner loop (B) show no resemblance to the native motif and were instead predicted as hairpins. Furthermore, only a small fraction of sequences (∼1–3) match the seeds after the extension step, hence indicating low reliability of the results.

These tests show that motifs residing in energetically favorable substructures, such as small hairpins, could still be identified using conventional MFE-based approaches. However, more complex and energetically less favorable active sites, such as larger hairpins or inner loops, require the prediction of suboptimal structures for successful motif extraction since their MFE configurations lack the corresponding base pair formations.

2.4 Comparison to alternative approaches

In order to demonstrate the true capabilities of our program, we applied two motif prediction programs, MEMERIS and GLAM2, to the datasets presented above and compared the resulting motifs to the outcome of Aptamotif. MEMERIS was chosen since it makes use of secondary structure information to guide the motif search toward single-stranded regions whereas GLAM2, unlike many other methods, allows for gaps in the subsequences constituting the identified area. To guarantee a fair comparison, each program was allowed to generate 10 solutions out of which the one that best fits the description of the published motif was chosen. Since, in contrast to Aptamotif, MEMERIS and GLAM2 require the expected width w of the motif as a parameter, we generated 10 results for each possible value of w and selected the best matches from this result set.

As shown in Figure 5A, for the dataset of Dobbelstein and Shenk, all three methods perform similarly well in capturing primary structure features of the published motif, demonstrating that Aptamotif is capable of competing with state of the art sequence based approaches in scenarios with continuous primary structures. Furthermore, the additional information regarding the secondary structure provided by our program outperforms the predictive capabilities of MEMERIS and GLAM2, offering higher level insight into possible binding interaction mechanisms between the aptamer and its apatope.

The advantage of ensemble-based secondary structure information is even more evident when inspecting the comparison of the second dataset, summarized in Figure 5B. While MEMERIS and GLAM2 are capable of pinpointing the individual modules of the interior loop, determining a functional dependency between these is purely based on the user's expertise and is considered a non trivial task. GLAM2 additionally tends to overestimate the true size of the modules due to its gap extension policy. In fact, many motifs reported by this program spanned the entire sequence region between modules 1 and 2, a region forming a non conserved hairpin structure of variable size (data not shown). These observations, combined with the knowledge about the role of suboptimal structures demonstrated in the previous section, illustrate that these discordancies cannot be solved by a simple post-processing analysis. Aptamotif, however, is capable of not only exactly identifying the involved primary structure but also provides the crucial information that the functional site constitutes two modules in order to form a biologically active entity.

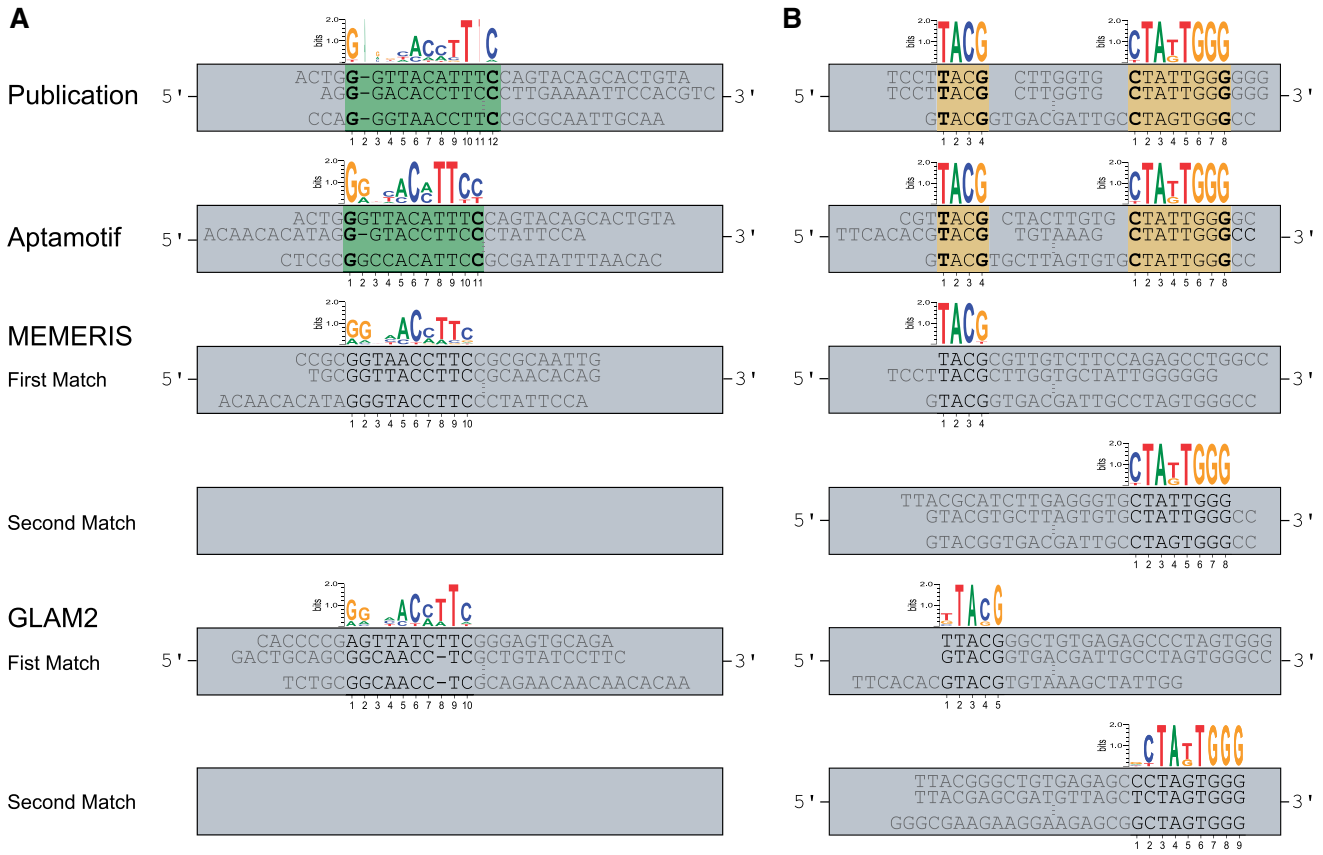

Fig. 5. Performance comparison of Aptamotif against MEMERIS and GLAM2 on the dataset of (**A**) Dobbelstein and Shenk (30 nt region) and (**B**) Lozupone *et al.* (26 nt region). Each program was allowed to generate 10 solutions out of which the best fitting the description of the published motif was chosen. MEMERIS and GLAM2 were allowed to generate two motifs to give each method equal chance of identifying the interior loop components. Sequences in the boxed areas represent a small subset of aptamers with an instance of the identified motif. Nucleotides participating in the motifs are colored in black, bold letters stand for nucleotides forming a base pair and sequence similarity is indicated by the sequence logo above. Green and yellow shaded areas show secondary structure features in terms of hairpins and interior loops, respectively, if provided by the method

We also applied RNAcontext to these two datasets. Even though the authors claim that RNAcontext can be used for motif extraction in the absence of detailed affinity data, we could not confirm this using our datasets. There are several potential reasons for this poor performance. First, RNAcontext was primarily developed for prediction of binding affinity values by fitting a statistical sequence–structure context model to the input pool of sequences and their experimentally measured affinity values, which is then interpreted to extract the binding motifs. Another confounding factor may be in the way the structural information is incorporated in the statistical model. Each nucleotide in the sequence is associated with its propensity to be in one of several secondary structure contexts. Thus the structures in the ensemble are 'averaged', which may cause the program to miss on motifs residing in secondary structure conformations not heavily represented in the ensemble. Our method on the other hand explicitly considers each structure in the ensemble by performing exhaustive search.

These examples clearly demonstrate the advantages and potentials of ensemble-based motif elucidation for sequence–structure motifs in aptamers. Our approach is capable of accurately identifying biologically active regions, even if conservation is shifted toward the secondary structure layer and can cope with motifs, whose native shape does not coincide with the MFE structure.

2.5 Application of aptamotif to analyzing ensemble properties across SELEX cycles

The selection process and evolutionary pressure exerted on the aptamer pool during SELEX are currently not well understood. We therefore generated and analyzed two new experimental sets of aptamers from consecutive SELEX cycles containing aptamers designed against the coagulation Factor VIII, a crucial blood clotting protein in humans. Factor VIII deficiency is known to cause the disease hemophilia A. The initial RNA pool contained a randomized region of 60 nt, flanked by 20 and 21 bases long primer sites in 5′ and 3′ direction, respectively. In each SELEX cycle, the strongest binders were selected and sequenced, leading to a dataset of 41 aptamers from cycle 3, and 45 representatives from cycle 5. Preliminary sequence analyses showed a low degree

of nucleotide similarity indicating that target specificity might be strongly shifted toward structural features. We applied Aptamotif to each of these sets and compared the reported motifs.

We then used these results to elucidate possible selection patterns and evolutionary properties of the species emerging from one cycle to the other. Our analysis revealed a multibranch loop structure comprising three components and containing the absolutely conserved sequence pattern `TTA` in the longest, central block. This multibranch loop was identified in both cycles, but with less frequency in cycle 5 (29 times in cycle 3 against 21 times in cycle 5). Interestingly, in the latter cycle, the same sequence pattern was additionally identified in an abundant hairpin motif not detected in round 3. In some cases, these two configurations were even found to occur together in the same aptamer, possibly enhancing the binding affinity of these species. These findings suggest that species of successive cycles represent a non random subset of previous rounds hence supporting the original theory of selection in SELEX experiments. Furthermore, they indicate that hairpins seem to outcompete multibranch loop binders, suggesting a selective pressure toward energetically more favorable substructures. We point that this type of insight would be impossible to obtain by analyzing sequence motifs alone, therefore, demonstrating the power of Aptamotif to gain insight into the selective mechanisms governing the SELEX protocol.

2.6 Runtime and implementation details

A middle sized dataset of approximately 50 sequences and 100 nt in length (including primers) requires $\sim$1.5–2 h of computation time on a 3 GHz dual-core CPU using the standard parameters defined in Section 2.2. The choice of K, G and δ has substantially higher impact on the runtime as compared to the dataset size, and is ranging between 2 min ($K=3, G=2, \delta=3$) up to 94 h ($K=6, G=4, \delta=3$). We did not find any improvement in motif quality with values $K>5$, $G>2$ and $\delta>3$. Values for $\delta\leq 2$ where insufficient for the generation of the inner loop motif.

MSA are computed using the MUSCLE alignment software (Edgar, 2004). Aptamotif generates an HTML file as output, containing detailed information of each motif in a tree-like, collapsible and well-organized structure, displaying loop type, P-value and information content, frequency of occurrence in the ensemble, and the seed sequences and substructure alignments. Furthermore, sequence logos of the single-stranded motif components as well as the secondary structure representation of each match in the structure ensemble allow for a graphical exploration of the motifs properties. Aptamotif is currently implemented as a modular and extendible python module and available upon request.

3 CONCLUSION AND OUTLOOK

We have presented Aptamotif, the first ensemble-based method for the identification of sequence–structure motifs in SELEX-derived aptamers, provided proof of principle validation on the example of two aptamer datasets containing experimentally determined motifs of increasing complexity, and highlighted its advantages over traditional motif finding programs. Additionally, we illustrated its application in elucidating several properties regarding the selection process between consecutive SELEX cycles.

In this work, we focused on the analysis of traditional SELEX data, i.e. relatively small samples of species from the sequence pool. Our results, and especially the comparison between consecutive cycles, suggest the existence of certain mechanisms governing the selection process that are still to be fully uncovered. It will be interesting to study such selection with the complete set of aptamers in a SELEX cycle, allowing for a more accurate motif analysis and providing a more detailed insight into the evolutionary concepts acting during this process. With the emergence of next-generation, high-throughput sequencing technologies, these data are now increasingly becoming available. In fact, we are currently generating such data for every pool of entire SELEX experiments, yielding millions of aptamers per cycle. This new dimension of information should allow for addressing numerous open questions that cannot be answered by traditional datasets alone. These include the correlation between indels and target affinity in binding regions, and elucidating the importance of motif adjacent stem–loop structures and their stabilizing contribution to the active site as well as their overall importance in the selection process. This can be addressed by an expansion of the current treatment of the motifs to broader local structures as, for example, considered in Backofen and Will (2004) for pairwise local sequence–structure alignments. Tracking the evolution of motifs from the initial pool to the last cycle is likely to additionally reveal specific selection properties in SELEX experiments, such as whether in early selection stages inner loops or multiple loops in which only one single-stranded region binds the target are always competed out of the pool in favor for energetically more stable hairpin or bulge structures. We are currently developing novel approaches that will allow the principles of Aptamotif to effectively scale with these emerging data masses in terms of sequence–structure processing and substructure ensemble-based motif examination.

Aptamotif addresses the urgent need of developing more advanced computational tools to support aptamer–apatope interaction research. We believe that our method not only provides a tool for efficient SELEX data analysis, but might also prove to be useful in the study of other naturally occurring non coding RNAs that might present conserved sequence–structure patterns among different organisms.

ACKNOWLEDGMENTS

Part of this work was performed when Jan Hoinka and Elena Zotenko were at the Department of Computational Biology and Applied Algorithmics at Max Planck Institute for Informatics, Saarbrücken, Germany. The authors thank Prof. Thomas Lengauer for his input and support. The authors also thank Raheleh Salari, NCBI, NIH for her helpful suggestions and careful survey of this manuscript.

Funding: Intramural Research Program of the National Institutes of Health, National Library of Medicine (partial); and Center for Biologics Evaluation and Research, Food and Drug Administration's Modernization of Science program (ZES) (partial). The findings and conclusions in this article have not been formally disseminated by the Food and Drug Administration and should not be construed to represent any Agency determination or policy.

Conflict of Interest: none declared.

REFERENCES

Backofen,R. and Siebert,S. (2007) Fast detection of common sequence structure patterns in RNAs. *J. Discrete Algorithms*, **5**, 212–228.

Backofen,R. and Will,S. (2004) Local sequence–structure motifs in RNA. *J. Bioinform. Comput. Biol.*, **2**, 681–698.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, **2**, 28–36.

Barrick,J.E. and Breaker,R.R. (2007) The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol.*, **8**, R239.

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

Bunka,D.H.J. *et al.* (2010) Development of aptamer therapeutics. *Curr. Opin. Pharmacol.*, **10**, 557–562.

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

Ding,Y. *et al.* (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11**, 1157–1166.

Dobbelstein,M. and Shenk,T. (1995) *In vitro* selection of RNA ligands for the ribosomal L22 protein associated with Epstein-Barr virus-expressed RNA by using randomized and cDNA-derived RNA libraries. *J. Virol.*, **69**, 8027–8034.

Dubey,A.K. *et al.* (2005) RNA sequence and secondary structure participate in high-affinity CsrA-RNA interaction. *RNA*, **11**, 1579–1587.

Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Esposito,C.L. *et al.* (2011) A neutralizing RNA aptamer against EGFR causes selective apoptotic cell death. *PloS One*, **6**, e24071.

Frith,M.C. *et al.* (2008) Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.*, **4**, e1000071.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Heyne,S. *et al.* (2009) Lightweight comparison of RNAs based on exact sequence–structure matches. *Bioinformatics*, **25**, 2095–2102.

Hiller,M. *et al.* (2006) Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, **34**, e117.

Höchsmann,M. *et al.* (2003) Local similarity in RNA secondary structures. *Proc. IEEE Comput. Soc. Conf. Bioinform.*, **2**, 159–168.

James,W. (2000) Aptamers. In: Meyers RA, ed. *Encyclopedia of Analytical Chemistry*. John Wiley and Sons Ltd., Chichester, pp. 4848–4871.

Kazan,H. *et al.* (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.*, **6**, e1000832.

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.

Kupakuwana,G.V. *et al.* (2011) Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing. *PLoS ONE*, **6**.

Lee,J.F. *et al.* (2004) Aptamer Database. *Nucleic Acids Res.*, **32**, D95–D100.

Li,H. *et al.* (2011) Aptamer selection for the detection of Escherichia coli K88. *Can. J. Microbiol.*, **57**, 453–459.

Lozupone,C. *et al.* (2003) Selection of the simplest RNA that binds isoleucine. *RNA*, **9**, 1315–1322.

Lvesque,D. *et al.* (2007) *In vitro* selection and characterization of RNA aptamers binding thyroxine hormone. *Biochem. J.*, **403**, 129–138.

Nagarajan,N. and Keich,U. (2008) FAST: Fourier transform based algorithms for significance testing of ungapped multiple alignments. *Bioinformatics*, **24**, 577–578.

Nagarajan,N. *et al.* (2005) Computing the *P*-value of the information content from an alignment of multiple sequences. *Bioinformatics*, **21** (Suppl 1), i311–i318.

Ni,X. *et al.* (2011) Nucleic acid aptamers: clinical applications and promising new horizons. *Curr. Med. Chem.*, **18**, 4206–4214.

Nielsen,L.J. *et al.* (2010) Aptamers embedded in polyacrylamide nanoparticles: a tool for *in vivo* metabolite sensing. *ACS Nano*, **4**, 4361–4370.

Otto,W. *et al.* (2008) Structure local multiple alignment of RNA. In *Proceedings of German Conference on Bioinformatics (GCB'2008)*, volume P-136 of *Lecture Notes in Informatics (LNI)*, pp. 178–188. Gesellschaft für Informatik (GI).

Pavesi,G. *et al.* (2004) RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. *Nucleic Acids Res.*, **32**, 3258–3269.

Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Applied Math.*, **45**, 810–825.

Schudoma,C. *et al.* (2010) Sequence–structure relationships in RNA loops: establishing the basis for loop homology modeling. *Nucleic Acids Res.*, **38**, 970–980.

Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.

Walker,S.C. *et al.* (2011) The dual use of RNA aptamer sequences for affinity purification and localization studies of RNAs and RNA-protein complexes. *Method Mol. Biol.*, **714**, 423–444.

Wang,P. *et al.* (2011) Aptamers as Therapeutics in Cardiovascular Diseases. *Curr. Med. Chem.*, **18**, 4169–4174.

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

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

Yamamoto,R. *et al.* (2000) A novel RNA motif that binds efficiently and specifically to the Ttat protein of HIV and inhibits the trans-activation by Tat of transcription *in vitro* and *in vivo*. *Genes Cells*, **5**, 371–388.

Zelada-Guilln,G.A. *et al.* (2010) Real-time potentiometric detection of bacteria in complex samples. *Anal. Chem.*, **82**, 9254–9260.

Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.

Zuker,M. and Sankoff,D. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol.*, **46**, 591–621.