On the Design of RNA Sequences for Realizing Extended Shapes

Marc Hellmuth*, Daniel Merkle†, and Martin Middendorf‡

*Interdisciplinary Centre for Bioinformatics
Department of Computer Science
University of Leipzig
Haertelstr. 16-18, D-04107 Leipzig, Germany
Email: marc@bioinf.uni-leipzig.de

†Department of Mathematics and Computer Science
University of Southern Denmark
Campusvej 55, DK-5230 Odense M, Denmark
Email: daniel@imada.sdu.dk

‡Parallel Computing and Complex Systems Group
Department of Computer Science
University of Leipzig
Johannisgasse 26, D-04103 Leipzig, Germany
Email: middendorf@informatik.uni-leipzig.de

Abstract—It is known that for two given secondary structures (defined by position of base pairings) an RNA string can easily be found that can fold into both structures. But for more than two secondary structures this is not necessarily possible. Moreover, when four or more secondary structures are given the problem to determine the least number of positions, such that after the removal of all incident base pairs, a compatible RNA sequence can be found, is known to be NP-complete ([1]). In this paper we introduce pseudo edges that are used to forbid that certain base pairs can bind and therefore can be used to define the properties of possible RNA secondary structures. We study the complexity of the problem to design an RNA sequence that can fold into different secondary structures each of them is described by a set of required and forbidden base pairs. We refine the NP-completeness results of [1] and show an analogous NP-completeness result for the realization problem concerning the removal of (pseudo) edges. We also present a polynomial time method for checking the realizability of extended shape graphs. Furthermore, we empirically analyze the influence of pseudo edges on the realizability for sets of random RNA sequences and for sets of aptamers.

Keywords—RNA folding; RNA Design; NP-completeness; Secondary Structure; Aptamer

I. INTRODUCTION

RNA molecules that can adopt alternate secondary structures have vital functions in living cells (see, e.g., [2]). Therefore several approaches have been developed for the analysis, the design, and the prediction of RNA structures (see, e.g., [3]). RNA sequences that self-assemble into certain desired three-dimensional structures have also been designed artificially (see, e.g., [4]). RNA switches are a class of RNA molecules which can change between two alternate stable secondary structures with different function. Such conformational switching of RNA molecules can be induced by diverse signals, e.g., the presence or absence of other certain molecules or specific physical conditions like temperature. There is some evidence that the codon-anticodon arrangement and the proper recognition of tRNAs at the ribosomal A site are controlled by an RNA switch [5]. RNA switches have also been designed artificially (for an overview see [6], [7]). An example is an RNA molecule that is triggered by ligand binding and has a switching mechanism which is similar to the one proposed for the ribosomal A site ([8]).

To find an RNA sequence which can fold into given multiple secondary structures is an important RNA design problem. A secondary structure of an RNA molecule of length \( n \) can be characterized by a so called shape graph. A shape graph \( G = (V,E) \) of size \( n \) has a vertex set \( V = \{v_1, ..., v_n\} \) and satisfies some restrictions on the edge set \( E \) (see Section II). Such a shape graph represents a secondary structure of an RNA molecule when vertex \( v_i \) represents the \( i \)th base of the RNA and each edge in \( E \) represents a Watson-Crick base pair or a wobble base pairs of the secondary structure.

The RNA design problem considered here is to find an RNA sequence \( s \) of length \( n \) that can fold into secondary structures \( S_1, ..., S_k \) represented by their shape graphs \( G_1, ..., G_k \). All base pairs defined by the edge set of \( G_i \) form a Watson-Crick base or wobble base pair in \( S_i \). If we have found such a sequence \( s \) we say that \( s \) realizes the given shape graphs or is compatible to them.

It is known ([9]) that for any two shape graphs \( G_1 \) and \( G_2 \) with vertex set of size \( n \) there exists an RNA sequence that is compatible with both shape graphs. While there is not necessarily an RNA sequence that is compatible with more than two shape graphs. It was shown in [1] that the decision version of the problem to find for given shape...
graphs $G_1, \ldots, G_k$, $k \geq 4$ of same size $n$ the minimal number of positions, such that after removal of all base pairs incident to these positions, there exists a compatible RNA sequence is NP-complete.

So far the RNA design problem has been considered only for the case that the RNA molecule to be found forms at least the base pairs given by the shape graph. In this paper we argue that the RNA design problem is interesting also under additional restrictions. Since the function of an RNA molecule is determined by its structure it might be also interesting to prohibit certain bases from binding. This is the type of additional restriction that we consider in this paper. This restriction can be used for example to require that certain loops in the secondary structure have a minimum size or that certain stems of the molecule have a maximum size or that certain stems of the molecule have a maximum size and that certain loops in the secondary structure have a minimum size. All these requirements fit easily into the combinatorial formulation of the RNA design problem. With this paper we hope to inspire more work on this topic.

II. Basic Definitions

An RNA sequence $s = s_1 \ldots s_n$ is a word over the alphabet $\mathcal{A} = \{A, C, G, U\}$ (this is also called the primary structure of the RNA). The secondary structure of an RNA molecule of length $n$ is a set $\text{Sec}$ of pairs $(i, j)$, where $1 \leq i < j \leq n$, such that: i) if $(i, j) \in \text{Sec}$, then $s_is_j \in \mathcal{B} = \{AU, UA, GC, CG, GU, UG\}$ (B is the set of allowed base pairings), ii) if $(i, j) \in \mathcal{B}$, then it is not the case that $i < k < j < l$, iii) if $(i, j), (k, l) \in \text{Sec}$ and $i \in (k, l)$ then $i = k$ and $j = l$ holds, iv) if $(i, j) \in \mathcal{B}$, then $j > i + \theta$, where $\theta$ is a fixed integer, usually taken to be 3.

A shape $S$ of size $n$ is a graph $S = (V_n, E)$ with set of vertices $V_n = \{v_1, \ldots, v_n\}$ and set $E$ of independent edges such that for any two edges $\{v_i, v_j\}, \{v_k, v_l\} \in E$, where $i < j$ and $k < l$, it is not the case that $i < k < j < l$. The fact that the edges are independent implies that each vertex is connected by at most one edge. A shape can represent a secondary structure where vertex $v_i$ represents the $i$th base of the corresponding RNA sequence and the edges of the shape correspond to the base pairs of the secondary structure. For given shapes $S_1, \ldots, S_k$ of the same size $n$, we define the graph of shapes $G(S_1, \ldots, S_k)$ as $G(V_n, \bigcup_{i=1}^k E(S_i))$ where $E(S_i)$ is the edge set of shape $S_i$. A string $s = s_1 \ldots s_n \in \{A, C, G, U\}^n$ realizes a shape $S$ of size $n$ if and only if for all $\{v_i, v_j\} \in E(S) \subseteq \{s_i s_j \in \mathcal{B} = \{AU, UA, CG, GC, GU, UG\}\}$, for a shape $S$ let $R(S)$ be the set of all sequences that realize $S$. Figure 1 shows an example of an RNA sequence that realizes two different shapes. With this definitions the Realizability problem can be defined [1]: For given shapes $S_1, \ldots, S_k$ of size $n$ find a string $s = s_1 \ldots s_n \in \{A, C, G, U\}^n$ that realizes all these shapes.

As already mentioned it is known that for any two shapes $S_1$ and $S_2$ of size $n$ there exists an RNA sequence which realizes these shapes. Formally, this is stated in the following theorem.

Theorem 1 (Intersection Theorem, [9]): For any two shapes $S_1$ and $S_2$ of same size holds: $R(S_1) \cap R(S_2) \neq \emptyset$.

Clearly, if there exists a string, that realizes all edges in the shapes $S_1, \ldots, S_k$, then this string realizes all edges in the graph $G(S_1, \ldots, S_k)$. The following generalization of the intersection theorem was presented in [10].

Theorem 2 (Generalized Intersection Theorem, [10]): Let $S_1, \ldots, S_k$ be shapes of length $n$. It holds...
\[ \bigcap_{i=1}^{k} R(S_i) \neq \emptyset \iff G(S_1, \ldots, S_k) \text{ is bipartite.} \]

\section{Extended Shapes}

In this section we introduce extended shapes which are used to model the requirement that in the sought RNA sequences the bases at certain positions can not form a base pair. Let \( S = G(V_n, E) \) be a shape of size \( n \). The set \( V'(S) = \{ v_i \in V_n \mid \exists v_j \in V_n \text{ such that } \{ v_i, v_j \} \in E \} \) is the set of all vertices that are not paired. In the following we extend shapes by adding a second type of edges. A \textit{pseudo edge} is an edge \( \{ v_i, v_j \} \) connecting two unpaired nodes \( v_i, v_j \in V'(S) \). The introduction of pseudo edges into extended shapes is restricted to the fact that for all nodes \( v \in V_n \) the degree of node \( v \) is at most one and pseudo-knots (see for example [11]) are not allowed, i.e., (pseudo) edges in an extended shape are not allowed to cross.

Formally, an \textit{extended shape} (or \textit{shape extension}) of shape \( S \) is defined by \( \hat{S} := (V_n, \hat{E}, \gamma) \) where \( S = (V_n, E) \) is a shape, \( E \subseteq \hat{E} \), \( \hat{E} := \hat{E} \setminus E \) is the set of pseudo edges, and \( \gamma : \hat{E} \to \{0, 1\} \) is a signing function such that

\[ \gamma(\{ v_i, v_j \}) = \begin{cases} 0 & \text{if } \{ v_i, v_j \} \in E \\ 1 & \text{if } \{ v_i, v_j \} \in \hat{E} \end{cases} \]

The edges in \( E = E(S) \) will also be called \textit{regular} edges to distinguish them from the pseudo edges. Examples of a shape and an extended shape are depicted in Figure 2.

Similar to a graph of shapes we are now able to define a graph of shape extensions. Let \( \hat{S}_1 = (V_n, \hat{E}_1, \gamma_1), \ldots, \hat{S}_k = (V_n, \hat{E}_k, \gamma_k) \) be the shape extensions of shapes \( S_1, \ldots, S_k \). The shape graph \( G(\hat{S}_1, \ldots, \hat{S}_k) \) of the extended shapes is defined by \( (V, \hat{E}, \gamma) := (V_n, \bigcup_{i=1}^{k} \hat{E}_i, \gamma) \) with \( \gamma : \hat{E} \to \{0, 1, 2\} \), such that

\[ \gamma(\{ v_i, v_j \}) = \begin{cases} 0 & \text{if } \{ v_i, v_j \} \in \bigcup_{i=1}^{k} (E(S_i)) \text{ and } \{ v_i, v_j \} \notin \bigcup_{i=1}^{k} (\hat{E}(S_i)) \\ 1 & \text{if } \{ v_i, v_j \} \in \bigcup_{i=1}^{k} (\hat{E}(S_i)) \text{ and } \{ v_i, v_j \} \notin \bigcup_{i=1}^{k} (E(S_i)) \\ 2 & \text{else} \end{cases} \]

If \( \gamma(\{ v, w \}) = 2 \) holds for two vertices \( v, w \) then \( \{ v, w \} \in E(S_i) \cap \hat{E}(S_m) \) for some \( i, m \in \{1, \ldots, k\} \), this is called an \textit{overlay} (of edges).

A binary string \( s = s_1 \ldots s_n \in \{0, 1\}^n \) realizes a shape \( S \) of size \( n \) if for all \( \{ v_i, v_j \} \in E(S) \) holds \( s_i \neq s_j \). In [11] it was shown that for given shapes \( S_1, \ldots, S_k \) there exists an RNA sequence that realizes these shapes if and only if there exists a binary string that realizes these shapes \( S_1, \ldots, S_k \).

It can be shown that an analogous equivalence does not hold in general for extended shapes and arbitrary sets for pairings of pseudo edges. Due to space limitations we consider here only the most interesting case to realize extended shapes with RNA sequences where the allowed pairings for regular edges are \( \mathcal{B} := \{AU, UA, CG, GC, GU, UG\} \) and the allowed pairings for pseudo edges are \( \mathcal{B}_{PE} := \{AA, UU, GG, CC\} \). For this case an equivalence between realizations with RNA sequences and realizations with 0-1-sequences can be shown (proof omitted due to space limitations). The \textit{realization of extended shapes} is defined as follows.

\textbf{Definition 1:} A string \( s = s_1 \ldots s_n \in \{A, C, G, U\}^n \) realizes an extended shape \( \hat{S} \) of length \( n \) if and only if i) for all edges \( \{ v_i, v_j \} \in E(S) \), holds: \( s_i \neq s_j \) in \( \mathcal{B} \) and, ii) for all edges \( \{ v_i, v_j \} \in \hat{E}(S) \) holds: \( s_i = s_j \), i.e., \( s_i s_j \in \mathcal{B}_{PE} \).

A string \( s_1 \ldots s_n \) has a \textit{character change} at position \( i \) if \( s_i \neq s_{i+1} \) for \( i \in \{1, \ldots, n - 1\} \). The following theorem which is built on the next lemma holds (the proofs are
Let $s_1, s_2$ be two signed graphs $G(G)$ can be realized by replacing some edges of a graph $G$. The number of character changes is such that we have considered here (it might not hold if all these paths have odd length).

**Problem 1:** Min-Node-Deletion in Shapes (MNS): Compute the minimum number of vertices, that must be removed from the graph $G(S_1, \ldots, S_k)$ such that $S_1, \ldots, S_k$ are realizable by a single RNA sequence.

**Problem 2:** Min-Edge-Deletion in Shapes (MES): Compute the minimum number of edges, that must be removed from the graph $G(S_1, \ldots, S_k)$ such that $S_1, \ldots, S_k$ are realizable by a single RNA sequence.

The corresponding problems for a given extended shape graph are called Min-Node-Deletion in Extended Shapes (MNeS) and Min-Edge-Deletion in Extended Shapes (MEdS). It has been shown in [1] that (the decision version of) MNS is NP-complete for $k \geq 4$.

Obviously, MNS is a subproblem of MNeS and therefore MNS is NP-complete for $k \geq 4$. In the following it will be shown that MES is NP-complete for $k \geq 3$. Since MES is a subproblem of MEdS this implies that MEdS is NP-complete for $k \geq 3$.

**A. Homeomorphic Extension**

Similar as in [1] we use a known theorem for homeomorphic extensions of graphs for the NP completeness proof.

**Definition 2:** Let $G = (V, E)$ and $G' = (V', E')$ be two graphs. $G'$ is a homeomorphic extension of $G$ if $G'$ can be obtained from $G$ by replacing some edges $\{v', v''\} \in E$ with paths $P(\{v', v''\}) = v', u_1, \ldots, u_k, v''$ where $u_1, \ldots, u_k$ are new nodes of degree 2 not yet contained in $G$. The graph $G'$ is called odd homeomorphic extension of $G$ if all these paths have odd length.

**Theorem 4 ([1]):** Let $k \geq 3$. Then for any graph $G$ with maximum vertex degree $k$ there exist $k$ shapes $S_1, \ldots, S_k$ such that $G(S_1, \ldots, S_k)$ is isomorphic to an odd homeomorphic extension of $G$ and $||G(S_1, \ldots, S_k)|| = O(||G||^2)$ where $||H||$ denotes the number of edges of a given graph $H$.

**B. Testing Realizability**

In the following we provide a polynomial time method that tests shapes and extended shapes w.r.t. their realizability. For this purpose we utilize the definition of signed graphs, which were first introduced in [12].

A signed graph, or briefly an $s$-graph, $H = (V, E, \varphi)$ consists of a graph $G = (V, E)$ and a signing function $\varphi : E \to \{+1, -1\}$. Let $P = [v_1, \ldots, v_n, v_{n+1}]$ be a path and $C = [v_1, \ldots, v_n, v_1]$ be a cycle in $G$. The sign $\varphi(P)$ (of a path) is defined as the product of the signs of its edges, i.e., $\varphi(P) := \prod_{i=1}^{n} \varphi(e_i)$ (respectively $\varphi(C) := \prod_{i=1}^{n} \varphi(e_i)$). A path $P$ (cycle $C$) is called positive (respectively negative), if the signs $\varphi(P)$ (respectively $\varphi(C)$) are positive (respectively negative). An
s-graph $H = (V, E, \varphi)$ is balanced, if all its cycles are positive.

**Theorem 5:** [12] An s-graph $H = (V, E, \varphi)$ is balanced if and only if $V$ can be partitioned in two disjoint subsets $V_1$ and $V_2$ such that

$$\varphi(\{u, v\}) = \begin{cases} -1, & \text{if } u \in V_1 \text{ and } v \in V_2 \\ +1, & \text{otherwise} \end{cases}$$

To prove the following lemma we use this theorem (the proof is omitted).

**Lemma 2:** Let $H = (V, E, \varphi)$ be an s-graph. Let $G = (V, E, \gamma)$ with $V(G) = V(H)$ be the graph obtained from $H$ by using the signing function $\gamma(e) = 0$ (respectively $\gamma(e) = 1$) if and only if $\varphi(e) = -1$ (respectively $\varphi(e) = +1$), i.e., edges with negative signs become regular edges and edges with positive signs become pseudo edges. The following holds: $H$ is balanced if and only if $G$ does not contain cycles with an odd number of regular edges and $G$ does not contain overlays.

In [13] an algorithm with time complexity $O(|E| + |V|)$ was presented to test if a graph is balanced. Thus we can determine the realizability of a given graph in the same time by applying Lemma 2 and Theorem 3.

**C. NP Completeness of MES and MEEs**

In this subsection we show the NP-hardness results. In [14] it was shown that the problem of removing a minimal number of edges to make a graph bipartite is NP-complete. Note that a graph is bipartite if and only if it does not contain any odd cycles.

**Theorem 6:** MES is NP-complete for $k \geq 3$ shapes.

*Sketch of proof:* Theorem 2 shows that shapes $S_1, \ldots, S_k$ can be realized if and only if $G(S_1, \ldots, S_k)$ has no odd cycles, i.e. if and only if $G(S_1, \ldots, S_k)$ is bipartite. The minimal number of edges that have to be removed from $G$ to make the resulting graph bipartite is equal to the number of edges that have to be removed from an odd homeomorphic extension $G'$ of $G$ to make $G'$ bipartite, which is shown in Lemma 3 (proof omitted). Furthermore, there exists a simple non-deterministic polynomial time algorithm for checking a solution of MES: we guess a subset of edges in the graph and check in $O(|E| + |V|)$ time whether after removing this subset the remaining part of the graph is balanced, applying Lemma 2 and the mentioned algorithm in [13], which completes the proof.

**Lemma 3:** Let $G = (V, E)$ be an odd homeomorphic extension of $G$ and let $E^*(G) \subseteq E(G)$ ($E^*(G') \subseteq E(G')$) be an edge set of minimum size that has to be removed from $G$ (respectively $G'$) such that the resulting graph is bipartite. Then it holds $|E^*(G)| = |E^*(G')|$. The next theorem follows now directly from the observation that MES is NP-complete.

**Theorem 7:** MEEs is NP-complete for $k \geq 3$ shapes.

---

**V. Empirical Results**

**A. Data Sets, Parameters**

In the following three different data sets will be used. For the artificial data set RAND we generated 200 random sequences (each base $A, C, G, or \ U$ chosen with the same probability) for sequence lengths $5, 6, 7, \ldots, 100$. Our biological data sets are based on aptamer sequences, i.e. strands of oligonucleotides that bind a specific target molecule. The first biological data set called APT includes all RNA aptamer sequences for which for a given sequence length more than 100 aptamers were available in the aptamer database [15] (The enquiry if we want to have modified sequences was answered with: no, yes, and no statement). The second biological set called APT$_u$ includes all RNA aptamer sequences of the set APT, but multiple identical sequences were deleted except for one representant. Thus no redundant sequences are contained in this set. Only those sequences with equal or more than 100 instances for a fixed length were left in APT$_u$. In the set APT the following sequence lengths are present: 14, 16, 20, 21, 22, 25, 26, 28, 29, 30, 31, 34, 35, 36, 38, 39, 40, 41, 49, 50, 59, 60, 65, 68, 80. Sequence lengths in the set APT$_u$ are 16, 20, 26, 30, 36, 39, 40 and 60.

From each set RAND, APT, and APT$_u$ we have chosen randomly four different sequences $S_1, S_2, S_3, S_4$ of the same length, folded them with the program RNAfold from the Vienna RNA package [16] to get the minimum free energy secondary structures, and computed the respective shape graphs $G(S_1, \ldots, S_4)$ which are denoted in the following by $G_{\text{RAND}}, G_{\text{APT}},$ and $G_{\text{APT}_u}$. This was repeated 100 times for each length in each data set. Hence all results in the following are averages over 100 test runs. Note that the sequence lengths of the input sequence is identical to the number of vertices in shape graphs.

**Number of Regular Edges:** The first investigation analyzes the correlation between the number of regular edges in shape graphs.
graphs and the length of the input sequences. Results are depicted in Figure 4. With an increasing number of vertices the average number of regular edges increases approximately linearly starting at a sequence length of \( \approx 10 \). The measured Spearman's rank correlation coefficient is \( \rho = 0.999 \) for \( \text{RAND} \), \( \rho = 0.984 \) for \( \text{APT} \) and \( \rho = 0.976 \) for \( \text{APT}_u \), at a significance level of 0.01 indicating that the relation in all three test cases are well characterizable by a monotone function. A notable fact is that the average number of regular edges in the shape graphs of the biological data set \( G_{\text{APT}} \) has a significantly different value for some sequence lengths. For lengths 34, 49, 50, 68, and 80 the corresponding average numbers of regular edges are 18.5, 41.4, 41.8, 62.3, and 74.7. These values are considerably smaller then the average values of the random sequences. The values for the set \( \text{APT}_u \) are similar as those for the set \( \text{APT} \) at the respective instance lengths.

**Realizability of (Extended) Shape Graphs:** In the following we investigate the influence of pseudo edges on the realizability of shape graphs. Note that realizability can be checked with the method presented in Section IV-B in time \( O(|V| + |E|) \). The relative frequency of realizable shape graphs for all data sets and all sequence lengths is given in Figure 5. This investigation was repeated, but this time we randomly included pseudo edges. For each test set of 4 sequences we iteratively included pseudo edges in the folded sequences at randomly chosen possible positions \( AA, UU, CC \) and \( GG \) as long as possible, while keeping the conditions of extended shapes. Then we tested again the property of realizability of these extended shapes. The results are given in Figure 6. The relative frequency of realizable shape graphs becomes clearly smaller with an increasing number of vertices. (The statistical significance analysis is omitted here). There is a monotone relation between the lengths of the shapes and the possibility to find a single sequence which realizes the respective four chosen structures. The reason for that is that with an increasing number of vertices the number of regular edges increases, and therefore the chance for building odd length cycles or overlays becomes higher, and thus the chance for destroying the realization property increases. The realizability is more limited in extended shape graphs, because the additional pseudo edges make overlays and cycles with an odd number of regular edges more likely.

The aptamer data sets show again a different behavior than the random data set. It can be seen in Figure 6 that none of the 100 extended shapes based on the \( \text{RAND} \) data set was realizable for sequence lengths greater than 37. The largest investigated and still realizable shape graph based on the \( \text{APT} \) data was based on sequence length 80. Remarkable are also the rather high frequencies of realizability of non extended shape graphs in Figure 5 for the \( \text{APT} \) data set in comparison to the random case at length 34, 50, 68 and 80. The corresponding average frequencies are 0.91, 0.54, 0.30 and 0.19 for the \( \text{APT} \) data set and 0.65, 0.38, 0.14, and 0.11 for the \( \text{RAND} \) data set.

**Number of Pseudo Edges to Hinder Realizability:** Figure 7 depicts the number of pseudo edges that could be inserted with the random addition strategy as explained above, without destroying the realizability of the shape graph (in some cases this was possible, until no additional pseudo edges could be added). The number of addable edges increases strongly for sequence lengths increasing from 5 to 10. This is due to the fact, that in these shapes none or only very few regular edges exist (see Figure 4), and thus the corresponding shape graphs are always realizable even with an maximum of inserted pseudo edges.

For sequence lengths \( \geq 20 \) the number of addable pseudo edges falls, simply due to fact that with an increasing number of vertices in the shape graph the number of realizable shape graphs decreases. Moreover, since the number of
regular edges increases with an increasing sequence length. The assumption is reasonable that additional pseudo edges in larger shapes lead faster to shape graphs with overlays or cycles with an odd number of regular edges, and thus to non-realizable structures.

Also in this investigation, the shape graphs based on the aptamer sequences lead to notable outliers. For example, for shape graphs of the APT data set with 34 vertices, the average number of edges that could be added was 8.7, whereas it was only 3.0 for the corresponding shapes based on random sequences.

VI. CONCLUSIONS

We have argued that the problem to design RNA sequences with secondary structures that can realize certain shapes is an interesting combinatorial problem for different types of restrictions to the shapes. The motivation for this lies in the need for proper RNA sequence designs that should satisfy desired properties, and to design RNAs that can fold into 2 or more different shapes. As an example for an interesting type of restrictions we introduced pseudo edges between base pairs to define properties of unpaired nodes. We investigated the complexity of realization problems for the resulting extended shape graphs. A method was presented that can check the realizability of extended shapes graphs in polynomial time. It was shown that the Min-Edge-Deletion in Shapes (MES) problem is NP hard for 3 or more given (extended) shapes. This result complements the result of [1] that the corresponding node deletion problem is NP hard for 4 or more given shapes. Moreover, we have empirically analyzed random sequences and aptamer sequences with respect to the realizability problem.

ACKNOWLEDGEMENT

We thank Peter F. Stadler and Steve Hoffmann for the helpful comments and all the fruitful discussions.

REFERENCES


