Diagrammatic Theory of RNA Structures and Ensembles with Trinucleotide Repeats

Chi H. Mak^{1,2,3,*} and Ethan N. H. Phan²

¹ Department of Chemistry, ² Center of Applied Mathematical Sciences and ³ Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA

* To whom correspondence should be addressed. Tel: 1-213-740-4101; Email: cmak@usc.edu

ABSTRACT

Trinucleotide repeat expansion disorders (TRED) are associated with the overexpansion of (CNG) repeats on the genome. mRNA transcripts of sequences with greater than 60 to 100 (CNG) tandem units have been implicated in TRED pathogenesis. In this paper, we develop a diagrammatic theory to study the structural diversity of these (CNG)_n RNA sequences. Representing structural elements on the chain's conformation by a set of graphs and employing elementary diagrammatic methods, we have formulated a renormalization procedure to resum these graphs and arrive at a closed-form expression for the ensemble partition function. With a simple approximation for the renormalization and applied to extended (CNG)_n sequences, this theory can comprehensively capture an infinite set of conformations with any number and any combination of duplexes, hairpins and 2-way junctions. To quantify the diversity of different (CNG)_n ensembles, the analytical equations derived from the diagrammatic theory were solved numerically to derive equilibrium estimates for the secondary structural contents of the chains. The results suggest that the structural ensembles of (CNG)_n repeat sequence with $n \sim 60$ are surprisingly diverse, and they are dominated largely by open segments, with only a small fraction of the nucleotides forming base pairs. At the same time, the variance in the secondary-structural contents on the chains is also quite large, indicating that their structures can undergo strong equilibrium fluctuations and are expected to be rather susceptible to perturbations.

STATEMENT OF SIGNIFICANCE

Trinucleotide repeat expansion disorders (TRED) are associated with the overexpansion of (CNG) repeats on the genome. mRNA transcripts of sequences with critical length greater than 60 to 100 (CNG) tandem units have been implicated in TRED pathogenesis, though their structures remain poorly characterized. Conventional view has tacitly assumed that conformations with maximal C:G base pairing dominate at equilibrium, but here we demonstrate that (CNG) repeat sequences are characterized by diverse ensembles of structurally heterogeneous folds and with a large variance of secondary structural contents. These ensembles of structures also undergo strong equilibrium fluctuations, rendering them rather susceptible to perturbations. These results were based on a novel diagrammatic approach to the ensemble partition function.

INTRODUCTION

A graphical approach for classifying RNA structures has been pioneered by Schlick et al. (1–3). Graphs provide an elegant method for categorizing the many diverse conformational structures that can be adopted by RNA sequences and can help us more easily recognize common topological features in RNA structures that are otherwise difficult to decipher from their 2- or 3-dimensional structures. But more than assisting in cataloging RNA motifs, graph can also provide clues to the “factorizability” of different classes of substructures in RNA folds (4) {companion paper}. Here, we develop these ideas further and formulate a diagrammatic theory to take advantage of these factorizabilities, applying it to evaluate the partition functions of RNA structure ensembles and to study the conformational diversities in their folded structures.

To appreciate what these factorizabilities means for RNA structures, consider an open RNA strand. An unfolded RNA is in a high-entropy state. Its structures are characterized by a diverse ensemble. If c denotes a chain conformation and $P(c)$ its probability, the total entropy content of this ensemble is given by $S = -k_B \sum_c P(c) \ln P(c)$. If the sequence spontaneously folds and develops secondary and/or tertiary structures, the conformational entropy of the chain is suppressed because base complementarity and stacking interactions produce constraints on the chain’s conformations. Under these constraints, the new probability for each conformation in the presence of these constraints $P'(c) = P(c|\text{constraints})$ incurs a penalty, and the loss of entropy is given by:

$$\Delta S = S(\text{with constraints}) - S(\text{no constraints}) = -k_B \sum_c P'_c \ln P'_c - P_c \ln P_c \quad (1)$$

where the sum runs over all conformations. If one can determine how the constraints imposed by the secondary and tertiary structures in the fold transforms $P(c) \rightarrow P'(c)$, ΔS can be determined.

In general, the constraints imposed by secondary/tertiary structures are correlated, but some of them may be independent. “Factorizability” describes how these constraints break up into independent (or approximately independent) subsets. For instance, if the fold introduces 4 constraints A , B , C and D but the effects of A and B produce are separable from C which is also separable from D , then $P'(c) = P(c|A, B, C, D) = P(c|A, B) \cdot P(c|C) \cdot P(c|D)$. Under this factorization, the entropy change in Eq.(1) is simply equal to $\Delta S = \Delta S(\text{with constraints } A, B) + \Delta S(\text{with constraint } C) + \Delta S(\text{with constraint } D)$.

In Schlick’s graph notations, the secondary structures inherent in a RNA fold are represented by edges and vertices (1–3). Using a large body of empirical data derived from Monte Carlo conformational sampling (4) {companion paper}, we have derived a set of general rules for determining which constraints are independent and how they can be divided into uncorrelated subsets. When attempting to summarize these rules visually, we discovered that there is a close correspondence between factorizability and how these constraints are depicted in the graphs proposed by Schlick. By devising an appropriate set of rules, these diagrams can be used to easily factorize the constraints. Once factorized, the diagrams can then be used to represent the partition function of the entire ensemble. In this paper, we illustrate how graphs can also be used to construct resummation techniques to sum the diagrams, leading to an analytical

expression for the partition function. We will apply this technique to examine the diversity of structures in ensembles of $(\text{CNG})_n$ trinucleotide repeat RNA sequences.

In a diverse family of neurological diseases known as trinucleotide repeats expansion disorders (TREDs) (5–9), the onset of illness is associated with the overexpansion of $(\text{CNG})_n$ repeats in the genome (8–10). While most of these expanded repeats occur in noncoding regions and do not appear to translate to aberrant proteins (9, 10), the mRNA transcripts of these overexpanded templates may interfere with cellular pathways leading to cytotoxicity (11, 12). At the same time, $(\text{CNG})_n$ expanded mRNA may also acquire unintended functions in the cell (13). Ascertaining the structures of these sequences is therefore necessary for the understanding of their functions.

Two examples of a short (CNG) repeat with different secondary structures are shown in Fig. 1. Because of their repeat structures, at least one-third of the nucleotides on $(\text{CNG})_n$ sequences cannot form complementary pairs upon folding. TRED disease onset is often associated with a critical expansion threshold of $n > 60$ to 100 (14). The structures most often associated with the gain of function hypothesis for CNG expanded RNA sequences cited in the literature is a necklace-like structure composed of a long stretch of successive two-way junctions interposed by short helices and with a hairpin stem-loop cap (10, 11, 15–19) like that shown in Fig. 1(a). Many of the studies conducted are based on short (CNG) repeats (10, 11) and the structures resolved are limited to those which can be isolated and crystalized (17–20). As the length of the CNG repeats grows, the diversity of accessible structures could grow rapidly as well. In this paper, we use the diagrammatic theory developed here to study the structural ensembles of long (CNG) repeats.

Figure 1.

Examples of a $(\text{CNG})_9$ repeat sequence in two different conformations. The graph representation for each is shown next to it. Each 2-bp duplex is represented by a dot. A hairpin loop is represented by a circle with one dot. A 2-way junction is represented by a circle with two dots. An arc represents two dangling ends. In the graphs, the number adjacent to each edge indicates its length in nt.

The methods described in this paper are applicable to RNA sequences with many transposable repeats. In this context, "transposable" does not mean mobile. Instead, transposable means that if a segment on the sequence is moved to another position, it does not alter the nucleotide sequence. In the formulation below, it will become clear why transposability is an inherent feature of diagrammatic theories. But a chain with a specific nucleotide sequence cannot form base complementarity interactions equally well among every pair of nucleotides across the entire sequence. As a result, the constraints produced by the secondary structures also have an additional sequence-position dependence, and this renders non-repeat sequences non-transposable. The methods developed here are suitable for treating homopolymeric chains or chains with a large number of regular repeats, and the techniques described in this paper have been formulated with these systems in mind. In the future, we will extend this method to more complex, less regular, and more diverse nucleotide sequences.

MATERIALS AND METHODS

Graph Representations

Fig. 2 shows examples of the basic graph elements. The top row shows a 3-way junction in (a), a pseudoknot in (b), a triplex in (c) and a quadruplex in (d). The bottom row shows graph representation of each. Paired regions are represented by dots, triplexes by triangles and quadruplexes by squares. These secondary structures impose constraints on the chains and, as a consequence, produce a loss of conformational freedom in the loop regions. The graphs have been designed to emphasize these features. The middle row in Fig. 2 illustrates pictorially how to transform each of the 2-dimensional drawing in the top row to arrive at the corresponding graph on the bottom row.

Figure 2.

Examples of diagrams representing different secondary structural elements. (a) A three-way junction where each dot represents a helix and loops are represented by lines. (b) A pseudoknot where black dots represent paired regions and unpaired loops are represented by lines. (c) A triplex structure represented diagrammatically by a black triangle. (d) A quadruplex represented by a black square.

We have demonstrated in (4) that most helices are “insulating”, i.e. the two elements immediately adjacent to a helix are approximately uncorrelated. With Fig. 2(a) as an example, the helix on the left and the one on top emanating from the 3-way junction in the center render the two hairpin loops largely uncorrelated with the three unpaired strands on the junction. The open arcs representing the 5' and 3' ends of the sequence experience no constraints, and they therefore cost no entropy. In the companion paper {companion paper}, we provide additional data demonstrating the factorizability of pseudoknots and quadruplexes. The factorizability conditions for pseudoknots and quadruplexes are more complex, but they will not be considered in this paper.

Utilizing these graph elements, Fig. 1(a) and (b) show diagrammatic representations for two different conformations of a (CNG)₈ repeat sequence. The graph associated with Fig. 1(a) shows that it has one hairpin loop and three 2-way junctions. The graph for Fig. 1(b) has one hairpin loop and two 2-way junctions. The loop size in nt is indicated next to each segment. All duplexes in (CNG) repeat structures are 2 bp (4 nt) in length, and since they are uniformly identical for all (CNG) repeat sequences, their sizes are not explicitly labeled in Fig. 1. These graphs resemble pearl necklaces; we will therefore refer to them as “necklace” diagrams. These necklace conformations are the ones generally thought to be most relevant for (CNG) repeat sequences in TREDs (10, 15).

Applying the factorization scheme to the structures in Fig. 1, each graph can be written as a product of independent probabilities. For example, the factorization of the graph in Fig. 1(b) is illustrated in Fig. 3, where the necklace separates into one hairpin loop of length d , two 2-way junctions, one with loop lengths b and f and the other with c and e , two dangling ends with lengths a and g , plus three 2-bp (or 4-nt) duplexes. The composite probability of the graph on the left of Fig. 3 is therefore given by:

$$P_1(a)P_2(b,f)P_2(c,e)P_0(a)P_0(g)[P_\bullet(4)]^3 \quad (2)$$

where $P_1(x)$ is the probability associated with a hairpin loop (or a “1-way junction”) of length x , $P_2(x,y)$ is the probability of a 2-way junction with loop lengths x and y , $P_0(x) = 1$ is the probability associated with an open strand and $P_\bullet$ is the probability of the duplex. For the loops in hairpin and junctions, their probabilities are given by $P = e^{\Delta S/k_B}$, where ΔS is the conformational entropy of a loop relative to an open strand. $P_\bullet(4) = e^{\Delta S_\bullet/k_B - \Delta H_\bullet/k_B T}$, the probability of a 2-bp (4-nt) duplex, has both enthalpic and entropy contributions in it, which involve stacking and base pairing interactions as well as the loss of conformational freedom suffered by the backbone to stack. An example of all the decomposable factors of a necklace diagram is given on the right side of Fig. 3.

Figure 3.

Example showing factorization of the diagram on the left into the factors on the right. The circle with one dot represents a hairpin loop of size d . Circles with two dots represent 2-way junctions. The two open line segments represent open strands. The three filled dots represent 2-bp (4-nt) duplexes. The corresponding expression for the composite probability is given in Eq. (2).

A Crude Approximation for the Partition Function from Necklace Diagrams

Since necklace diagrams are expected to be the ones most relevant to (CNG) repeats, we will use them to illustrate how to calculate the partition function Z from the graphs. The complete set of unbranched necklace diagrams is shown in Fig. 4. Let the length of the chain be l nt. The sum of all segments in every diagram in Fig. 4 must equal l . To sum over all conformations consistent with these graphs, we must trace over all possible combinations of segments subject to the constraint that they sum to l . For example, the second term on the right in Fig. 4 corresponds to:

$$\sum_{a,b,c}^{\{l-(4)\}} P_0(a)P_1(b)P_0(c)P_\bullet(4) \quad (3)$$

where the sums over the lengths of the three segments a , b and c are constrained such that $a + b + c = l - 4$, which is the length of the chain minus the number of nucleotides in the duplex. Similarly, the third term corresponds to:

$$\sum_{a,b,c,d,e}^{\{l-2(4)\}} P_0(a)P_1(b)P_0(c)P_2(d,e)[P_\bullet(4)]^2 \quad (4)$$

Figure 4.

First four terms in an approximation for the partition function Z including only nonbranched necklace diagrams for the (CNG) repeats. The corresponding expression is given in Eq.(5).

Writing out the terms explicitly in Fig. 4, the partition function for a l -nt chain is:

$$\begin{aligned} Z(l) = & P_0(n) + [P_\bullet(4)] \sum_{a,b,c}^{\{l-(4)\}} P_0(a)P_1(b)P_0(c) + [P_\bullet(4)]^2 \sum_{a,b,c,d,e}^{\{l-2(4)\}} P_0(a)P_1(b)P_0(c)P_2(d,e) \\ & + [P_\bullet(4)]^3 \sum_{a,b,c,d,e,f,g}^{\{l-3(4)\}} P_0(a)P_1(b)P_0(c)P_2(d,e)P_2(f,g) + \dots \end{aligned} \quad (5)$$

The constrained sums in Eq.(5) are implied in their corresponding diagrams in Fig. 4.

The constraints in Eq.(5) for the partition function are cumbersome. Instead, consider the generating function of Z :

$$Z(\lambda) = \sum_{l=0}^{\infty} Z(l)e^{-\lambda l} \quad (6)$$

Notice that we have used the same symbol Z for the fixed- l partition function as well as its generating function. This should not cause any confusion because in the rest of this paper only the generating function $Z(\lambda)$ will be considered. The generating function serves the same purpose as the grand canonical partition function, with λ corresponding to the chemical potential divided by $k_B T$. Using the generating function removes the constraints in Z for fixed l , and Eq.(5) becomes:

$$\begin{aligned} Z(\lambda) = & P_0(\lambda) + [P_\bullet(\lambda)]P_0(\lambda)P_1(\lambda)P_0(\lambda) + [P_\bullet(\lambda)]^2P_0(\lambda)P_1(\lambda)P_0(\lambda)P_2(\lambda) \\ & + [P_\bullet(\lambda)]^3P_0(\lambda)P_1(\lambda)P_0(\lambda)P_2(\lambda)P_2(\lambda) + \dots \end{aligned} \quad (7)$$

where all the P functions are now generating functions of those in Eq.(5). In Eq.(7), the chain length is no longer fixed, but it is possible to compute the expectation of the length $\langle l \rangle$, so that the value of λ can be

chosen to match the average $\langle l \rangle$ to the proper target chain length l . The graph representation for Eq.(7) for the generating functions is identical to that shown in Fig. 4, but the interpretation is distinct.

Special Graph Elements

We define a few special graph elements in this section in order to facilitate the renormalization scheme in the rest of the paper.

a. Dangling Ends and Bridges

The RNA constructs under consideration here always have “dangling ends” on both their 5’ and 3’ ends. Since these dangling ends correspond to open strands, they cost no entropy. Diagrammatically, these correspond to the arcs in Fig. 4. Mathematically, they correspond to $P_0(\lambda)$.

Diagrams can also have “bridges”. A bridge is an edge which if removed from the graph will separate the graph into two disconnected pieces. Examples are shown in Fig. 5. Diagrammatically, bridges also correspond to arcs or non-loop edges, and mathematically, they also correspond to $P_0(\lambda)$. Bridges and dangling ends are topologically equivalent. They cost no entropy and correspond to open strands with probability $P_0(\lambda)$ given by:

$$P_0(\lambda) = 1/[1 - e^{-\lambda}] \quad (8)$$

b. Root Vertices

The vertex between two bridges or dangling ends is called a “root”. Examples are shown in Fig. 5. In Ref. 4, we demonstrated that all graphs representing RNA folds are Eulerian and because of this property, the structure of every graph must consist of one or more root vertices connected by bridges, except for the one with just a single edge. Some examples are shown in Fig. 5. A root vertex is simply the beginning and end of a subpart of the diagram that itself forms an Eulerian circuit. Other than the bridges between them, these subparts of the diagram rooted in root vertices do not touch each other.

c. Fragile Vertices

Vertices that are defined as “fragile” are the ones which if they are removed from the diagram factor the diagram into disconnected pieces. Examples are the duplexes in Fig. 3 between the hairpin loop and

the 2-way junctions, as well the duplexes connect one 2-way junction to the next. As the factorization on the right side of Fig. 3 shows, all helices (dots) in that diagram are fragile.

In contrast, example of vertices that are nonfragile are shown in Fig. 6. These include paired regions in a pseudoknot, as well as base contacts in triplexes and quadruplexes. Empirically, these vertices are found not to factorize completely. For example, the three segments in a pseudoknot are correlated. Similarly, the two loops in a triplex are also not independent. The three loops inside a quadruplex are also not independent.

d. Node Vertices

“Nodes” are topologically similar to roots, except they are not adjacent to bridges. Some examples are shown in Fig. 7 to illustrate how they differ from roots. A node is essentially a branch point. It is the beginning and end of a subpart of the diagram that itself forms an Eulerian circuit, which when the node is removed, would completely disconnect from the rest of the graph.

Root Renormalization and Necklace Approximation for the Root Function

The graphs described here share many similar features with those employed in quantum field theory and in liquids, where diagrammatic techniques have been used extensively to manipulate graphs(21).

These techniques will be employed here for RNAs.

Root “renormalization” is a diagrammatic reclassification technique. Under root renormalization, all diagrams in the partition function are regrouped into terms forming a simple series shown in the first equality in Fig. 8. Here, the “root function” $R(\lambda)$ represents the sum over all possible graphs that begin and end on a root node. Some examples of what are included in the root function $R(\lambda)$ are shown in Fig. 9 (which is only a partial list and clearly a very small subset of the full set). How the power series in R can be resummed using this regrouping is demonstrated by the second equality in Fig. 8. This yields the result given diagrammatically on the last line in Fig. 8, producing the following *exact* analytical expression for Z :

$$Z(\lambda) = 1/[1 - e^{-\lambda} - R(\lambda)] \quad (9)$$

$$\begin{aligned} \textcircled{Z} &= \text{---} + \text{---}\textcircled{R}\text{---} + \text{---}\textcircled{R}\text{---}\textcircled{R}\text{---} \\ &\quad + \text{---}\textcircled{R}\text{---}\textcircled{R}\text{---}\textcircled{R}\text{---} + \dots \\ &= \text{---} \left[1 + \textcircled{R}\text{---} + (\textcircled{R}\text{---})^2 + \dots \right] \\ &= \frac{(\text{---})}{1 - \textcircled{R}\text{---}} \end{aligned}$$

Figure 8.

Root renormalization of the partition function Z .

Under root renormalization, the partition function $Z(\lambda)$ is re-expressed in a succinct analytical function, in terms of the root function $R(\lambda)$. While this seems tantalizing, the root function itself is not known in closed form, so the practical utility of Eq.(9) is limited. To make progress, approximations are often invoked to derive different models for $R(\lambda)$. Employing different approximates for $R(\lambda)$ in the renormalization of the root then generates distinct classes of theories.

Here, we will consider a “necklace approximation”, $R_n(\lambda)$, for the root function. This approximation is shown in Fig. 10. Again, loops with one dot represent hairpins, and those with two dots represent 2-way junctions. Since each 2-way junction interposes between two duplex stems and every circuit beginning and ending on a root vertex must have at least one hairpin, the root function $R_n(\lambda)$ includes every diagram with any number of 2-way junctions and a hairpin. Fig. 10 shows how this series is easily resummed, yielding a closed-form expression for the root function:

$$R_n(\lambda) = \frac{P_1(\lambda)P_\bullet(\lambda)}{[1 - P_2(\lambda)P_\bullet(\lambda)]} \quad (10)$$

$$\begin{aligned}
 R_n &= \text{circle with dot} + \text{two circles connected at a dot} + \text{three circles connected in a chain at dots} + \dots \\
 &= \text{circle with dot} \left[1 + \text{circle with dot} + \text{circle with dot}^2 + \dots \right] \\
 &= \frac{\text{circle with dot}}{1 - \text{circle with dot}}
 \end{aligned}$$

Figure 10.
Necklace approximation for the root function and its resummation.

Substituting Eq.(10) into Eq.(9) produces the necklace approximation for the partition function:

$$Z(\lambda) \approx \left(1 - e^{-\lambda} - \frac{P_1(\lambda)P_*(\lambda)}{[1 - P_2(\lambda)P_*(\lambda)]} \right)^{-1} \quad (11)$$

Notice that the partition function $Z(\lambda)$ in Eq.(11), obtained by using the necklace approximation to the root function $R(\lambda)$, contains *many more diagrams* than the necklace approximation to $Z(\lambda)$ in Fig. 4. The much larger set of diagrams included in Eq.(11) contains graphs with any number of roots and bridges connecting them, whereas Fig. 4 is limited to graphs with just one root. Using the necklace approximation $R_n(\lambda)$ for the root function, $Z(\lambda)$ captures an infinite set of diagrams, including every possible combination of hairpins, 2-way junctions and bridges.

The root function obeys a recurrence relationship that can be derived from a similar renormalization procedure applied to the nodes instead. In a future paper, we will address node renormalization strategies in greater detail. But essentially, the result of this on the root function is to generate an equation of the Dyson type(21), shown in Fig. 11. Solving this equation for $R_n(\lambda)$ leads to the same result in Eq.(10). But for cases where the direct resummation of the approximate equation for the root function is not possible, the Dyson equation exemplified by Fig.11 provides an alternate route to the solution.

$$R_n = \text{circle with dot} + \text{circle with dot and } R_n \text{ inside}$$

Figure 11.
Dyson equation for the root function in the necklace approximation.

Specializing to (CNG) Repeat Sequences

The theory above has been formulated for generic RNA sequences. To specialize the formulation to apply to (CNG)_n repeat sequences specifically, we take a closer look at their repeat structure. By “repeat structure”, we are referring to the periodicity of the nucleotide sequence. In our calculations, we employ (CNG)_n constructs with the following architecture:

with n repeating units of (NGC). Formally, this (CNG)_n construct has $l = 3n + 1$ nucleotides instead of $3n$. This is done to ensure that the 5' and 3' ends of the chain do not have to be treated differently, but it does not materially alter the results or the formulation.

As described above, the periodicity of the (CNG)_n sequence permits canonical base pairing producing 2-bp duplexes only, and we assume the base N does not pair with each other or with C or G (though this assumption can be modified and easily incorporated to re-formulate the theory). Because of this repeat structure, unpaired segments on the sequence are limited to lengths equal to 1, 4, 7, 11, ... nt. This means that the sums over l used to define the generating functions, for example the one for the partition function in Eq.(6), should be restricted to multiples of 3 instead. Consequently, instead of summing over l in Eq.(6), we sum over trinucleotide units n . And with just this adaptation, all of the diagrammatic arguments and the equations developed above carry over without any modification, with the proviso that we count in trinucleotide units instead of single nucleotides. To do this, every loop length in the formulation should be replaced by its length divided by 3. For example the lengths $\{a, b, \dots g\}$ in Fig. 3 become $\{a' = a \setminus 3, b' = b \setminus 3, \dots g' = g \setminus 3\}$, where $\setminus$ denotes an integer division without remainder. A loop with length $a' = 0$ is 1-nt long. A loop with $a' = 1$ is 4-nt long, etc. The only exception to this rule is a 2-bp (4-nt) duplex, which is assigned a length of 2 repeat units instead of 1.

RESULTS AND DISCUSSION

The diagrammatic approach formulated here was applied to (CNG)_n repeat sequences, using the necklace approximation in Eq.(11) for the generating function of the partition function $Z(\lambda)$. The inputs, $P_*(\lambda)$, $P_1(\lambda)$ and $P_2(\lambda)$, were obtained from the free energies of formation of duplexes, hairpins and 2-way junctions from Monte Carlo data reported in an earlier paper (4). The functional dependence of the loop free energies on the loop lengths were extended beyond the finite-length data available from the simulations by using scaling relationships, yielding the following expressions at $T = 310$ K:

$$P_*(\lambda) = e^{-\left(2\lambda - \frac{3.42}{0.616} + \lambda_*\right)} \quad (12a)$$

$$P_1(\lambda) = e^{-\left(\lambda + \frac{5.016}{0.616}\right)} + e^{-\left(2\lambda + \frac{5.848}{0.616}\right)} + e^{-\left(3\lambda + \frac{6.159}{0.616}\right)} + e^{-\left(4\lambda + \frac{5.086}{0.616}\right)} \cdot \Phi\left(e^{-\lambda}, 1.75, \frac{13}{3}\right) \quad (12b)$$

$$P_2(\lambda) = Q_2(\lambda) - dQ_2(\lambda)/d\lambda \quad (12c)$$

$$Q_2(\lambda) \equiv e^{-\left(\lambda + \frac{6.528}{0.616}\right)} + e^{-\left(2\lambda + \frac{6.797}{0.616}\right)} + e^{-\left(3\lambda + \frac{6.880}{0.616}\right)} + e^{-\left(4\lambda + \frac{5.587}{0.616}\right)} \cdot \Phi\left(e^{-\lambda}, 1.75, \frac{14}{3}\right) \quad (12d)$$

where Φ is the Lerch transcendent (22). These expressions are based on a baseline free energy of each GC|CG doublet being -3.42 kcal/mol stable (23–25). Depending on the identity of N in the (CNG) repeat unit, there are additional stabilizations resulting from the terminal mismatch, and they can be calculated using Turner's parameters (25). For N = A, C and U, these additional bonuses are approximately $\Delta\Delta G = -0.9$, -0.5 and -1.6 kcal/mol per duplex, respectively. In Eq.(12a), we have added an extra term $\lambda_{\bullet} = \Delta\Delta G/RT$ to the exponent for each duplex (dot) to allow for this. In general, $\lambda_{\bullet}$ can be thought of as a chemical potential for the duplexes. For the case of N = G, the extra stabilization when the duplex is adjacent to a hairpin is -5.2 kcal/mol, but it is different at -7.9 when adjacent to a 2-way junction. This disparity can be accommodated by assigning different $\lambda_{\bullet}$ values to the numerator and denominator in the last term in Eq.(11), but this paper will not address (CGG) repeats as they can potentially also form quadruplexes {companion paper}.

Since the generating function for Z sums over all n , the value of λ that corresponds to the desired target value for $\langle n \rangle$ must first be identified using the relationship $\langle n \rangle = -\partial \ln Z / \partial \lambda$, similar to what is typically done in the grand canonical ensemble. The result is shown in Fig. 12, where the solid blue line shows $\langle n \rangle$ as a function of λ . Since the critical expansion threshold associated with disease onset in TREDS is $n > 60$, we employed 60 as the target value for $\langle n \rangle$. The denominator of $Z(\lambda)$ changes sign at several values of λ , leading to divergences in Z . As Fig. 12 shows, these divergences are also manifested in $\langle n \rangle$. On Fig. 12, the proper value of λ that produces $\langle n \rangle = 60$ is indicated by the open circle on the solid blue line, yielding $\lambda = 0.1135$. The other two curves in Fig. 12 relate to the effects of the stability of the duplex on the average repeat length. They will be discussed later.

Figure 12.

Solid blue curve shows the expectation value $\langle n \rangle$, the number of repeat units, as a function of λ . The point that corresponds to the target value $\langle n \rangle = 60$ is indicated on the solid blue curve by an open circle. The dotted dash orange curve shows the same relationship when the stability of each duple is made more favorable by adding a free energy of $-RT$. The dashed green curve shows the same with an unfavorable bias of $+RT$ added to each duplex.

Using the value $\lambda = 0.1135$, corresponding to a target $\langle n \rangle = 60$, we calculated the expectation values for the number of duplexes, hairpins and 2-way junctions in the ensemble. These were computed by conjugating a chemical potential to each of these elements and differentiating the partition function against each of them. The results are shown in Fig. 13. The blue solid line in Fig. 13 shows the average number of repeats and is identical to the solid blue line in Fig. 12, except with only the relevant domain of λ displayed in Fig. 13. The open circle on the solid blue line in Fig. 13 indicates the target value of $\langle n \rangle = 60$, and the vertical thin dashed line, the value of λ that produces it. The green long dashed line in Fig. 13 shows the average number of duplexes found on the sequence, giving $\langle n_{\bullet} \rangle \sim 4.35$ for repeat length $\langle n \rangle = 60$. The purple short dashed line in Fig. 13 shows the average number of hairpins found on the sequence, and $\langle n_1 \rangle \sim 4.07$ for repeat length $\langle n \rangle = 60$. The red dotted dashed line in Fig. 13 shows the average number of 2-way junctions found on the sequence, and $\langle n_2 \rangle \sim 0.28$ for $\langle n \rangle = 60$. Because of the specific topologies of those diagrams included in the necklace approximation of the root function, the number of hairpins plus the number of 2-way junctions must equal the number of duplexes, and this conservation law is consistent with the observed expectation values.

Figure 13.

Expectation values of different features in an equilibrium ensemble of (CNG) repeat sequences at 310 K as a function of λ . Solid line (blue): (CNG) repeats. Long dashed line (green): duplexes. Dotted dashed line (red): 2-way junctions. Short dashed line (purple): hairpin loops. Vertical thin dashed line indicates the proper value of λ that produces the target value for $\langle n \rangle = 60$, indicated by the open circle on the solid blue line.

The ensemble average of the number of duplexes, $\langle n_{\bullet} \rangle$ is a measure of the secondary structure contents on the chain at equilibrium. The maximum number of duplexes, in the case of (CNG) $_n$ repeat sequences, that can be accommodated on the chain is $n_{\bullet, \text{max}} = \langle n \rangle / 2$. This corresponds to the case when the sequence folds into a single conformation, corresponding to the longest necklace structure possible. This “maximal” necklace conformation has been implicitly assumed to be the dominant structure for (CNG) repeat sequences at equilibrium. But the data in Fig. 13 suggest that this is not the case. For (CNG) $_n$ chains with $n = 60$, the dominant equilibrium conformations are instead characterized by largely open strands, with only a few duplexes, ~ 4 to 5. From the second derivative of Z with respect to the chemical potential of the duplexes, we also computed the variance in the number of duplexes, $\sigma_{\bullet}^2 = \langle n_{\bullet}^2 \rangle - \langle n_{\bullet} \rangle^2$, which turned out to be ~ 24.0 . This suggests that while the average number of duplexes on the chain is ~ 4 to 5, the equilibrium fluctuations are rather large, with a standard deviation ~ 4.9 repeat units. In other words, the equilibrium conformations of the chain in a thermal ensemble at $T = 310$ K are rather diverse, where the expectation value of the number of duplexes is $\langle n_{\bullet} \rangle = 4.35 \pm 4.9$ (CNG) units, suggesting that a rather broad distribution of secondary structural contents is expected to characterize (CNG) repeat sequences in solution. The assumption that the ensemble should be dominated by the conformation with the largest number of 2-way junctions and duplexes corresponding to the maximal-length necklace, is unfounded.

Fig. 14 shows the sensitivity of the chain’s expected duplex content on free energy bias applied to different structural elements on the chain. The three panels in Fig. 14 are similar to Fig. 12, but instead of showing $\langle n \rangle$ on the vertical axis, they show the “duplex fraction”, defined as the average number of duplexes $\langle n_{\bullet} \rangle$ divided by the number of repeats $\langle n \rangle$, plotted as a function of λ , the chemical potential

conjugate to the number of repeats. The duplex fraction is a primary indicator of the level of secondary structure contents on the chain. The theoretical maximum for the duplex fraction is 0.5, corresponding to the maximal necklace structure. Fig. 14(a) shows the effects of adding a bias λ_* to the duplexes: the orange dotted dashed line for a favorable bias $\lambda_* = -1$, the green dashed line for an unfavorable bias $\lambda_* = +1$ and the solid blue line for no bias. Similarly, Fig. 14(b) shows effects of adding a bias to the 2-way junctions, λ_2 (favorable bias $\lambda_2 = -1$ in the orange dotted dashed line, unfavorable bias $\lambda_2 = +1$ in the dashed green line, no bias in the solid blue line). Fig. 14(c) shows the same, but for a bias applied to the hairpins ($\lambda_1 = -1$ for orange dotted dashed line, $+1$ for green dashed line and 0 for solid blue line.) And as a point of reference, the solid blue lines in all three panels are the same curve, since they correspond to 0 bias on all three elements. In each panel, the value of the duplex fraction corresponding to the proper target length $\langle n \rangle = 60$ under each set of biases are identified by circles. These were found by using Fig. 12, in a procedure similar to what has described above. The solid blue line in Fig. 12 describes how $\langle n \rangle$ depends on λ , and the orange dotted dashed line shows the same in the presence of a favorable biased $\lambda_* = -1$ placed on the duplexes, whereas the bias $\lambda_* = -1$ is unfavorable in the green dashed line. The circles in Fig. 12 show the proper values of λ corresponding to the target length $\langle n \rangle = 60$. These values are also highlighted by the circles in Fig. 14(a). A similar procedure was used to identify the proper values of λ when biases are placed on either the junctions or the hairpins, but leaving out the details, these are also highlighted by the circles in Figs. 14(b) and (c).

Fig. 14 demonstrates that no matter whether the bias is applied to the duplexes (λ_1 in (a)), 2-way junctions (λ_2 in (b)) or the hairpins (λ_1 in (c)), it influences the formation of duplexes in the same direction, i.e. favorable biases on either duplexes, junctions or hairpins produced more duplexes, and unfavorable biases on any of these three produced fewer duplexes. This is not surprising, because the formation of hairpins, junctions and duplexes are correlated with each other. Biasing any one of these elements will

have an overall effect on the number of duplexes, which is the primary indicator of the level of secondary structure contents on the chain. However, it is also clear that biasing the duplexes has the largest impact, whereas biasing the 2-way junctions has the least. (Note difference in vertical scales among the three panels.) Compared to the effects resulting from biasing the duplexes or biasing the 2-way junctions, the impact of biasing the hairpins on the chain's secondary structures is slightly weaker than the effect of biasing duplexes, but a bit stronger than that coming from biasing junctions.

Here, we want to make one interesting observation on the graphs in Fig. 14. The same domain of λ values used for Fig. 12 is also used for Fig. 14. But while Fig. 12 shows there are divergences in the expectation value of $\langle n \rangle$ as a function of λ , there are no apparent divergences in the duplex fraction in Fig. 14, after the number of duplexes $\langle n_{\bullet} \rangle$ has been divided by $\langle n \rangle$. This demonstrates an interesting feature in diagrammatic resummation, where the radius of convergence in many of the infinite sums can be effectively extended in the complex plane around singularities when certain thermodynamic observables are considered. For example, in this case of the duplex fraction, there are discontinuities in $\langle n_{\bullet} \rangle / \langle n \rangle$ even at the places where the denominator of the partition function $Z(\lambda)$ changes sign.

The data in Fig. 14 characterizing how secondary structure contents on the chain respond to biases can also be understood in terms of the susceptibility of $\langle n_{\bullet} \rangle$ to variations in the chemical potentials of duplexes, $\lambda_{\bullet}$, junctions, λ_2 and hairpins, λ_1 . For example, the susceptibility $\chi_{\bullet\bullet} \equiv \langle \partial n_{\bullet} / \partial \lambda_{\bullet} \rangle_{\lambda_{\bullet}=0}$ can be computed directly from the partition function $Z(\lambda)$, giving $\chi_{\bullet\bullet} = -23.9$. The other susceptibilities for the effects of junctions and hairpins on the formation of duplexes can be computed analogously, giving $\chi_{\bullet 2} = -1.82$ and $\chi_{\bullet 1} = -22.1$, respectively. Similarly, the susceptibilities of junctions and hairpins can also be calculated the same way, yielding $\chi_{2\bullet} = \chi_{\bullet 2} = -1.82$, $\chi_{22} = -0.399$, $\chi_{21} = -1.43$, $\chi_{1\bullet} = \chi_{\bullet 1} = -22.1$, $\chi_{12} = \chi_{21} = -1.43$ and $\chi_{11} = -20.6$. These susceptibilities agree with the directions and magnitudes of the effects observed in Fig. 14.

In Fig. 15, we show how the duplex fraction varies as a function of different biases, in units of RT , imposed on the duplexes $\lambda_{\bullet}$ (solid lines) or on the hairpins λ_1 (dashed line). Whereas in Fig. 14, the applied biases are either ± 1 , Fig. 15 shows the duplex fraction as a continuous function of the bias on the duplexes $\lambda_{\bullet}$ or the hairpins λ_1 , for $\langle n \rangle$ fixed at 60. When the bias is applied to the duplexes, i.e. when $\lambda_{\bullet}$ is varied, there are actually two families of solutions. These are shown in Fig. 15 as the orange solid line, labeled the “open” branch, and the cyan solid line, labeled the “necklace” branch. In contrast, there is only one physical solution when the bias is applied to the hairpins, i.e. when λ_1 is varied, and this is indicated by the dashed line in Fig. 15.

Up to this point, we have only been discussing solutions on the “open” branch. The data presented in Fig. 12, 13 and 14 above are all associated with $\lambda_{\bullet} = 0$ where there is only one physical solution on the “open” branch. And as we have described above, these solutions on the “open” branch are characterized by largely open chains with a moderate duplex fraction that varies as a function of the bias applied to the duplexes. The majority of the chain’s secondary structures on the “open” branch are hairpins, with a very minor fraction in the form of 2-way junctions.

As we have described under Eq. (12) above, depending on whether N = A, C or U, there are additional free energy bonuses associated with the duplexes, $\Delta\Delta G = -0.9$, -0.5 and -1.6 kcal/mol per duplex, respectively. The values of the $\lambda_{\bullet}$ bias corresponding to these values are indicated in Fig. 15 by the vertical dashed lines, and the corresponding solutions for the duplex fraction highlighted by open circles. Fig. 15 shows that different (CNG) repeat sequences with N = C, A or U are expected to produce somewhat different duplex contents on the chain. Their structural ensembles, however, are physically similar since they are all characterized by solutions on the “open” branch, i.e. they have large structural diversities as well as large equilibrium fluctuations.

Before discussing the second branch, we examine the dashed line in Fig. 15, which describes how the duplex fraction varies as a function of bias imposed on the hairpins λ_1 . These results were obtained

by varying λ_1 , while keeping $\lambda_* = 0$, i.e. no additional bias on the duplexes. These solutions as a function of λ_1 belong to the same physical family of solutions as the “open” branch as a function of λ_* , so we can compare them with each other. For small biases for either λ_* or $\lambda_1 \geq -3$, the effect is only slightly stronger when the bias is applied directly to the duplexes, compared to when the bias is applied to the hairpins. This may seem somewhat puzzling because one would expect a perturbation applied directly to the duplexes should have a stronger effect on the duplexes, but this result is nonetheless consistent with results from Fig. 14. Considering the diagram that are included in the root function (see Fig. 10), it is clear that one duplex is always paired with a hairpin, while at the same time, one duplex is also paired with every 2-way junction. The fact that the effects on duplex formation are similar regardless whether the bias is imposed directly on the duplexes or on the hairpins implies that the number of 2-way junctions must be small in the ensemble, and Fig. 13 also corroborates this. When the magnitude of the bias becomes larger beyond λ_* or $\lambda_1 < -3$, the difference between them becomes more significant. In this region, direct perturbations applied to the duplexes do initiate more duplexes in comparison to when the bias is applied to hairpins. The duplex fraction, in the limit of very large bias imposed on hairpins, appears to approach an asymptotic value of ~ 0.33 , whereas it approaches the theoretical maximum of 0.5 when the bias is applied directly to the duplexes. The bifurcation between these two curves suggests that the number of 2-way junctions in the ensemble becomes more significant when a larger bias is placed on the duplexes directly.

We now turn to the “necklace” branch. This family of solutions are shown as the cyan solid line labeled “necklace” in Fig. 15. As the bias applied to the duplexes λ_* is increased, a second and physically distinct family of solutions appears. This new branch of solutions is characterized by a duplex fraction that grows in rapidly as a function of λ_* . This branch appears abruptly when λ_* reaches the critical value ~ -1.9 , where the duplex fraction on the cyan solid line begins to take on a nonzero value. Once the second branch is seeded, its duplex fraction increases rapidly toward its theoretical maximum value of 0.5. We have also computed the expectation values of hairpins and 2-way junctions along this so-called “necklace” branch, and the average number of hairpins is close to one along the entire branch. The ensembles characterized by solutions on this new branch are therefore made up of the type of necklace structures illustrated in Figs. 1, 3 and 4, with a stretch of successive two-way junctions interposed by shorts helixes and with a single hairpin cap. This is not surprising, because given sufficiently stable duplexes, one would expect the chains to want to maximize base pairing even at the cost of lowering conformational entropy and diversity, in order to fold into these necklace structures with increasing number of junctions, until they grow to maximum length, at which point the duplex fraction becomes 0.5. Fig. 15 shows that this is indeed the asymptotic behavior displayed by this branch of solutions. But while this suggests that the necklace diagrams in Figs. 1, 3 and 4 are indeed predicted by the diagrammatic theory here, we note that the magnitude of the bias needed to be applied to the duplexes are quite large for these necklace solutions to become physically relevant.

Fig. 15 shows that for $N = C$ or A , (CNG) repeat sequences only have a single solution on the “open” branch. But for $N = U$, the stronger stability of duplexes with a U:U terminal mismatch generates a second solution on the “necklace” branch also, and because of this, (CUG)_n repeat sequence could produce physically distinct ensembles compared to (CAG)_n or (CCG)_n chains. But while this is true, Fig. 15 also shows that at the value of $\lambda_s \sim -1.6$ for $N = U$, the duplex fraction on the “necklace” branch is actually lower than on the “open” branch. This suggests that while it is possible for (CUG)_n chains to develop an alternative subensemble dominated by necklace structures, there is actually an overall *higher* level of secondary structure contents in the “open” ensemble. Because of this, we can conclude that for all three $N = C, A$ or U , the equilibrium conformations of long (CNG)_n repeats are dominated by a low level of secondary structures at 310 K.

CONCLUSION

We have formulated a diagrammatic theory to study the conformational ensembles of (CNG)_n RNA sequences. Transcripts of overexpanded microsatellites on the genome containing 60 to 100 (CNG) repeats have been implicated in a number of neurological diseases known as TREDs. To understand the structures of these (CNG) repeats sequences, we performed a series of calculations aimed at characterizing their equilibrium ensembles. Based on a diagrammatic representation of the partition function, our calculations are based on using graphs to annotate secondary structural motifs on the chains, and, in conjunction with evidence from previous simulation studies, these diagrammatic representations allowed us to easily factorize the graphs in order to re-express the free energy of each configuration as a sum of independent terms. Using generating function mathematics and diagrammatic resummation techniques, we were able to derive a closed-form expression for the partition function in terms of a renormalized root function, which is the diagrammatic equivalence of the sum over all self-contained circuit diagrams. Employing a simple approximation for this root function, we derived analytical expressions for the partition function and its corresponding thermodynamic observables. With the so-called necklace approximation for the root function, the partition function captures an infinite set of conformations with any number and any combination of duplexes, hairpins and 2-way junctions. Together with simulation data from a self-consistent library of thermodynamic costs previously obtained for the various graph elements, we solved the resulting equations to arrive at numerical estimates for the ensemble expectation values of duplexes, hairpins and 2-way junction on the chain, enabling us to quantitatively characterize the structural diversity of different (CNG)_n ensembles.

While most studies in the field have implicitly assumed that the ensemble of a (CNG)_n sequence is dominated by a single structure having the maximal number of paired bases forming duplexes interposed by 2-way junctions between them, the results of this study suggest otherwise. The data show that the structural ensembles of (CNG)_n repeat sequence with $n \sim 60$ are surprisingly diverse. The equilibrium number of duplexes, hairpins and junctions on these sequences indicate that their secondary structure contents are far from the expected maximally paired conformation. To the contrary, the ensemble is

dominated by open segments, with approximately 8 to 13 duplexes on the average from (CCG)_n to (CUG)_n chains, whereas the maximally paired structure would have 30 duplexes. But given the exaggerated equilibrium fluctuations on the structure, the number of duplexes could be as high as 16 to 23. The results show that the majority of the loops are in the form of hairpins, but the variance in the secondary-structural contents on the chains are also quite large within these ensembles. And due to these large equilibrium fluctuations, the chains are also highly susceptible to external perturbations. The results show how perturbations in the form of biases on the stabilities of the various structural motifs, duplexes, junctions and hairpins, could affect the secondary structures of the chains in either directions. It appears that the number of duplexes and hairpins are also strongly correlated even though the overall level of secondary structure exhibit strong equilibrium fluctuations, and because of this, it is quite possible that the secondary structures could be easily pinned by tertiary contacts between different parts of the chain, either from the chain interacting with itself or from interactions mediated by the binding or association of the chain to ligand(s). This may in turn have implications on how these (CNG)_n sequences could acquire unintended functions in the cell, leading to their cytotoxicity. Studies are underway to examine whether these perturbations could indeed quench the structural fluctuations of the chains. We are also investigating different classes of approximations for the root function to try to ascertain whether other structural elements, such as multiway junctions or pseudoknots, could also play important roles in controlling the structural ensembles of these (CNG) repeat sequences.

AUTHOR CONTRIBUTIONS

CHM designed the study. CHM and ENHP carried out the work. CHM and ENHP wrote the manuscript.

ACKNOWLEDGEMENT

This material is based in part upon work supported by the National Science Foundation under Grant Number CHE-1664801.

REFERENCES

1. Hin Hark Gan, Daniela Fera, Julie Zorn, Nahum Shiffeldrim, Michael Tang, Uri Laserson, Namhee Kim, and Tamar Schlick. 1987. RAG: RNA-As-Graphs database—concepts, analysis, and features. *Nutr. Health*. 5:1285–1291.
2. Gan, H.H., S. Pasquali, and T. Schlick. 2003. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res*. 31:2926–2943.
3. Fera, D., N. Kim, N. Shiffeldrim, J. Zorn, U. Laserson, H.H. Gan, and T. Schlick. 2004. RAG: RNA-As-Graphs web resource. *BMC Bioinf*. 5:88.
4. Mak, C.H., and E.N.H. Phan. 2018. Topological Constraints and Their Conformational Entropic Penalties on RNA Folds. *Biophysical Journal*. 114:2059–2071.
5. Ranum, L.P., and T.A. Cooper. 2006. RNA-mediated neuromuscular disorders. *Annu. Rev. Neurosci*. 29:259–77.
6. Mirkin, S.M. 2006. DNA structures, repeat expansions and human hereditary disorders. *Current opinion in structural biology*. 16:351–8.
7. Mirkin, S.M. 2007. Expandable DNA repeats and human disease. *Nature*. 447:932–40.
8. Neueder, A. 2019. RNA-Mediated Disease Mechanisms in Neurodegenerative Disorders. *Journal of Molecular Biology*. 431:1780–1791.
9. Khristich, A.N., and S.M. Mirkin. 2020. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem*. 295:4134–4170.
10. Kiliszek, A., and W. Rypniewski. 2014. Structural studies of CNG repeats. *Nucleic Acids Res*. 42:8189–8199.
11. Mooers, B.H.M., J.S. Logue, and J.A. Berglund. 2005. The structural basis of myotonic dystrophy from the crystal structure of CUG repeats. *PNAS*. 102:16626–16631.
12. Miller, J.W., C.R. Urbinati, P. Teng-umnuay, M.G. Stenberg, B.J. Byrne, C.A. Thornton, and M.S. Swanson. 2000. Recruitment of human muscleblind proteins to (CUG)_n expansions associated with myotonic dystrophy. *The EMBO Journal*. 19:4439–4448.
13. Qawasmi, L., M. Braun, I. Guberman, E. Cohen, L. Naddaf, A. Mellul, O. Matilainen, N. Roitenberg, D. Share, D. Stupp, H. Chahine, E. Cohen, S.M.D.A. Garcia, and Y. Tabach. 2019. Expanded CUG Repeats Trigger Disease Phenotype and Expression Changes through the RNAi Machinery in *C. elegans*. *Journal of Molecular Biology*. 431:1711–1728.
14. Orr, H.T., and H.Y. Zoghbi. 2007. Trinucleotide Repeat Disorders. *Annual Review of Neuroscience*. 30:575–621.

15. Sobczak, K., G. Michlewski, M. de Mezer, E. Kierzek, J. Krol, M. Olejniczak, R. Kierzek, and W.J. Krzyzosiak. 2010. Structural Diversity of Triplet Repeat RNAs. *J. Biol. Chem.* 285:12755–12764.
16. Broda, M., E. Kierzek, Z. Gdaniec, T. Kulinski, and R. Kierzek. 2005. Thermodynamic Stability of RNA Structures Formed by CNG Trinucleotide Repeats. Implication for Prediction of RNA Structure. *Biochemistry.* 44:10873–10882.
17. Kiliszek, A., R. Kierzek, W.J. Krzyzosiak, and W. Rypniewski. 2012. Crystallographic characterization of CCG repeats. *Nucleic Acids Res.* 40:8155–8162.
18. Kiliszek, A., R. Kierzek, W.J. Krzyzosiak, and W. Rypniewski. 2011. Crystal structures of CGG RNA repeats with implications for fragile X-associated tremor ataxia syndrome. *Nucleic Acids Res.* 39:7308–7315.
19. Tamjar, J., E. Katorcha, A. Popov, and L. Malinina. 2012. Structural dynamics of double-helical RNAs composed of CUG/CUG- and CUG/CGG-repeats. *Journal of Biomolecular Structure and Dynamics.* 30:505–523.
20. Kumar, A., H. Park, P. Fang, R. Parkesh, M. Guo, K.W. Nettles, and M.D. Disney. 2011. Myotonic Dystrophy Type 1 RNA Crystal Structures Reveal Heterogeneous 1 × 1 Nucleotide UU Internal Loop Conformations. *Biochemistry.* 50:9928–9935.
21. Mattuck, R.D. 1992. A Guide to Feynman Diagrams in the Many-Body Problem: Second Edition. 2nd edition. New York, USA: Dover Publications.
22. Gradshteyn, I.S., and D. Zwillinger. 2014. Table of integrals, series, and products. Eighth edition /. San Diego, CA: Academic Press.
23. Turner, D.H. 1996. Thermodynamics of base pairing. *Curr. Opin. Struct. Biol.* 6:299–304.
24. Mathews, D.H., J. Sabina, M. Zuker, and D.H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–940.
25. Turner, D.H., and D.H. Mathews. 2009. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* 38:D280–D282.