Quantifying Structural Diversity of CNG Trinucleotide Repeats Using Diagrammatic Algorithms

Ethan N. H. Phan[2] and Chi H. Mak[1,*]

[1] Department of Chemistry, Centre of Applied Mathematical Sciences and Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA
[2] Department of Chemistry, 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 (TREDs) exhibit complex mechanisms of pathogenesis, some of which have been attributed to RNA transcripts of overexpanded CNG repeats, resulting in possibly a gain-of-function. In this paper, we aim to probe the structures of these expanded transcript by analyzing the structural diversity of their conformational ensembles. We used graphs to catalog the structures of an NG-$(CNG)_{16}$-CN oligomer and grouped them into subensembles based on their characters and calculated the structural diversity and thermodynamic stability for these ensembles using a previously described graph factorization scheme. Our findings show that the generally assumed structure for CNG repeats—a series of canonical helices connected by two-way junctions and capped with a hairpin loop—may not be the most thermodynamically favorable, and the ensembles are characterized by largely open and less structured conformations. Furthermore, a length-dependence is observed for the behavior of the ensembles' diversity as higher-order diagrams are included, suggesting that further studies of CNG repeats are needed at the length scale of TREDs onset to properly understand their structural diversity and how this might relate to their functions.

STATEMENT OF SIGNIFICANCE

Trinucleotide repeats are DNA satellites that are prone to mutations in the human genome. A family of diverse disorders are associated with an overexpansion of CNG repeats occurring in noncoding regions, and the RNA transcripts of the expanded regions have been implicated as the origin of toxicity. Our understanding of the structures of these expanded RNA transcripts is based on sequences that have limited lengths compared to the scale of the expanded transcripts found in patients. In this paper, we introduce a theoretical method aimed at analyzing the structure and conformational diversity of CNG repeats, which has the potential of overcoming the current length limitations in the studies of trinucleotide repeat sequences.

INTRODUCTION

Trinucleotide repeats and the disorders associated with their expansion is a growing field of study. Understanding the pathogenesis of trinucleotide repeat expansion disorders (TREDs) requires insights into the structure-function relationship that relates these repeat sequences to their RNA transcripts and the symptoms attributed to them. Trinucleotide repeats are microsatellites known to exhibit large length variability (1–3). Base on their locations on the genome and the extents of expansion, these repeats are known to cause a variety of neurological disorders such as Huntington's disease and fragile X syndrome. It is often unclear whether an aberrant gene product or the RNA transcript of the repeats themselves is responsible for their cytotoxicity (4–7). Amongst the diverse family of TREDs, those caused by CNG repeats have become a major research target.

While expanded CNG repeats may lead to aberrant protein products with extended sequences of glutamines (polyQ), many of the trinucleotide expansions are not found in coding regions. In these cases, cytotoxicity may be due to the mRNA transcripts, either via a gain or loss of function. Often, disease manifestation is associated with a critical expansion threshold, and as a result, onset of disease appears to be a function of age. Examples of gain of function has been demonstrated in myotonic dystrophy type 1 (DM1) (4), where expanded CTG repeats in the 3'-untranslated regions of the *dystrophia myotonica* protein kinase gene produces a RNA transcript which interacts with CUG-binding proteins (CUGBP1) and muscleblind-like (MBNL1) proteins. These interactions alter protein levels in the cell, which in turn affects their function as splicing regulators, leading to symptoms (8–10). Understanding the *in vivo* structure of the repeats may lead to a better understanding of how RNA with expanded CNG sequences may interact with these proteins.

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 shorts helixes and with a hairpin stem-loop cap (11–16). Many of the studies conducted, however, are based on CNG repeat oligomers, whereas the threshold of TRED disease onset is typically associated with expansions of 60 to 100 units or longer. Additionally, the structures resolved are limited to those which can be isolated and crystalized. As the length of the CNG repeats grow, the diversity of accessible structures could grow rapidly as well, with the necklace structure comprising only one possible subset of motifs out of many. This leaves a gap in our understanding of the structure-function relationship that may be responsible for pathogenesis in CNG-related TREDs. The structural diversity of CNG chains on the order of lengths comparable to the critical expansion thresholds of TREDs remain unclear. In this paper, we present a calculation aimed at addressing this question by using an algorithm based on a diagrammatic approach.

MATERIAL AND METHODS

Backbone Conformational Entropy and Secondary Structures

An open RNA strand is characterized by an ensemble of many diverse conformations and is in a high-entropy state. If this RNA sequence can fold and develop secondary structure(s), the entropy of the chain will decrease as the chain folds because the number of conformations that are consistent with the secondary structures in the folded state is necessarily lower than the unfolded state. This decrease in conformational entropy S of a RNA can be viewed as a result of the constraint(s) imposed by the secondary structural elements present in the fold on the possible conformations of the chain. The entropy loss from the unfolded state to the folded state is:

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

where c is a chain conformation, P_c and P_c' are the normalized probabilities of that conformation with and without the constraint(s) imposed by the secondary structure, and k_B is Boltzmann's constant. If an alternative fold with a different set of secondary structural elements exists, it will in general have a different conformational entropy because the constraints are different. Adding more constraints to the chain in the form of more complex secondary structures will necessarily lead to a more negative ΔS, but the constraints imposed by the different secondary structural elements in a fold are in general not independent, e.g. ΔS with constraints A + B is not necessarily equal to ΔS with constraint A plus ΔS with constraint B.

In order to more easily characterize RNA secondary structures, Schlick et al. have proposed a diagrammatic scheme (17–19). Fig. 1 shows examples of some of the diagrams representing different kinds of secondary structural elements. The bottom row of Fig. 1(a) illustrates the diagram of a three-way junction. Each of the three helices is represented by a black dot. Unpaired loops are represented by curved lines. The bottom row of Fig. 1(b) shows the diagrammatic representation of a pseudoknot. The two black dots now represent the two paired regions, whereas the unpaired loops are represented by straight or curved lines. A triplex is illustrated in Fig. 1(c). A black triangle is used in its diagrammatic representation to represent the three-base interaction in this secondary structural element. Fig. 1(d) shows a quadruplex, and a black square is used to represent the interactions of the four bases in this secondary structure. In these diagrams, the points where two or more lines converge are called "vertices". The lines emanating from each vertex are called "edges" and their number define the degree d_v of vertex v. A dot (representing a duplex) always has four edges and $d_v = 4$. A triangle (representing a triplex) always has six edges with $d_v = 6$, and a square (representing a quadruplex) always has eight edges and $d_v = 8$. Because all RNAs are linear polymers, any graph representing a RNA fold is necessarily Eulerian (20), meaning there is a way to trace through the entire diagram over all its edges only once. This also implies that either the degree of every vertex will be even or only two vertices will be odd while all other are even. In such Eulerian graphs, the number of edges E including the two dangling ends on the 5' and 3' ends is

$$E = 1 + \frac{1}{2}D \qquad (2)$$

where $D = \sum_v d_v$ is the total degree over all vertices in the diagram.

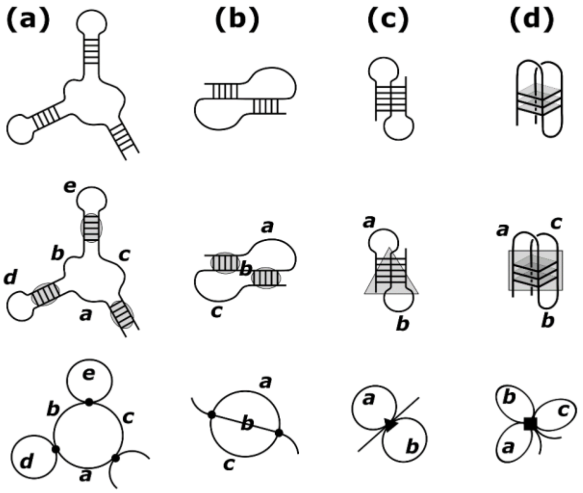


Figure 1.

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.

In addition to their utility in characterizing RNA secondary structures, the diagrammatic representation proposed by Schick et al. is also useful for the calculation of the conformational entropy of folded RNA structures. In a recent paper (21), we described how the constraints imposed by the secondary structural elements of any folded state can be broken into approximately independent sets using a factorization strategy based on how the elements of the diagrams are connected. An example of how this factorization works is illustrated in Fig. 2 for a three-way junction. A "fragile vertex" is defined as any vertex that if removed from the diagram disconnects it into two or more disjoint pieces. Fig. 2 shows that all three vertices in the diagram of a three-way junction are fragile. Disconnecting the diagram at these fragile vertices generates the factorized diagrams on the far right of Fig. 2. When a diagram is completely factorized, it breaks up into irreducible pieces. We have proven that the conformational entropy of the fold is also approximately separable when a diagram is reducible, and ΔS becomes the simple sum of the entropies of all the irreducible pieces. For example, the total entropy ΔS of the three-way junction in Fig. 2 can be reduced to the sum of the three closed diagrams on the right. The big circle with arcs labeled a, b and c represents the unpaired loops in the three-way junction. The two smaller circles labeled d and e represent the loops in the hairpins. The edges corresponding to the 5' and 3' ends of the folded structure contributes nothing to ΔS since it contains no additional constraints and can be omitted from the completely factorized diagram shown on the right. The dots represent duplexes of different helix length (α, β, or γ). Each fragile vertex contains additional enthalpic and entropic free energy contributions depending on its size. The free energy of each fragile vertex can, for example, be estimated using Turner's nearest-neighbor model (22–24) or from computer simulations.

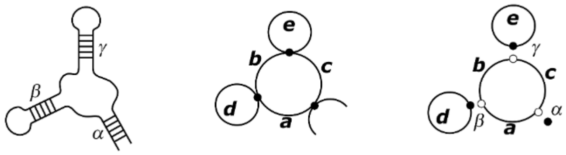


Figure 2.

Factorization of the constraints in a three-way junction into approximately independent contributions. The total entropy ΔS is reduced to the sum of the three closed diagrams on the right.

Monte Carlo Simulations

Diagrammatic factorization provides a simple recipe for the calculation of the conformational entropy of any RNA fold. To make use of it, a library of conformational entropy data must be compiled for every irreducible element representing various types of secondary structural elements (hairpin, junction, duplex, triplex, quadruplex, etc.) of different sizes, as well as for other unfactorizable structures such as pseudoknots. This library can be sourced from experimental data or from computer simulations. In a previous paper (21), we have provided a complete and consistent set of Monte Carlo simulation results for the entropy values of hairpins, two-, three- and four-way functions. These correspond to diagrams in which the vertices are all degree 4 (i.e. dots). While some of the same data are available from melting experiments (22, 24, 25), not everything is. We have relied on extensive computer simulations to compile an internally-consistent data set.

This paper further extends this data library, adding results for non-canonical base pairs and diagrams for quadruplexes and pseudoknots. These new data also serve to demonstrate factorizability questions in pseudoknots and quadruplexes, highlighting comparisons and contrasts with what has already been proven for two-, three- and four-way junctions. This data library provides entropic contributions to the free energy associated with the unpaired loops of each diagram. Free energies of the paired regions associated with the vertices in each diagram are independent of the edges, and they can be added separately. Vertex free energies are not included in the reported data.

Monte Carlo (MC) simulations have been carried out using our in-house Nucleic MC program for high-throughput conformational sampling of RNAs (26). Detailed discussions of the mixed numerical/analytical treatment and closure algorithm used in simulating the sugar-phosphate backbone(26–30) and accounting for steric interactions (31, 32), solvent effect (32–34), and counterions' influence (35, 36) have been presented in previous publications. Using Nucleic MC, we generated thermal ensembles consisting of several million uncorrelated conformations for chains with many different secondary-structural constraints corresponding to a number of different classes of diagrams.

To evaluate the conformational entropy of quadruplexes and pseudoknots, poly-U constructs of many different structures were simulated. For diagrams involving helices with Watson-Crick (WC) base pairs, long hairpin structures in the protein databank were melted to obtain the appropriate starting conformation. Using the same parameters for defining base pairing events from our previous study (21), we then identified and counted spontaneous base pair formations during the simulation to measure the entropic cost of initiating any new base pair constraint within the structure. Multiple base-pairing constraints are associated with some of these structures. To evaluate the entropy of these, we computed the entropy cost for forming the first constraint, and holding the first constraint, we then computed the additional entropy cost for forming the second constraint, etc. Since entropy is a state function, any thermodynamic pathway between the initial (open) and final (folded) states will yield the correct ΔS. For example, the starting structure of a pseudoknot was chosen to produce the proper length for the α helix as shown in Figure 3, with the size of the seeded hairpin chosen to provide a range of lengths in the final assembled pseudoknot structure.

For quadruplexes, parameters for identifying Hoogsteen base pair are needed. The structures of pyrimidine-purine base pairs utilizing the purine's Hoogsteen edge (PDBID 1GQU (37) and 2PS6 (38)) as well as quadruplexes with different topologies (PDBID 1KF1 (39) and 143D (40)) were used to define the base pairing criteria for identifying Hoogsteen pairs. These selection parameters for WC and Hoogsteen pairs are summarized in Fig. 3.

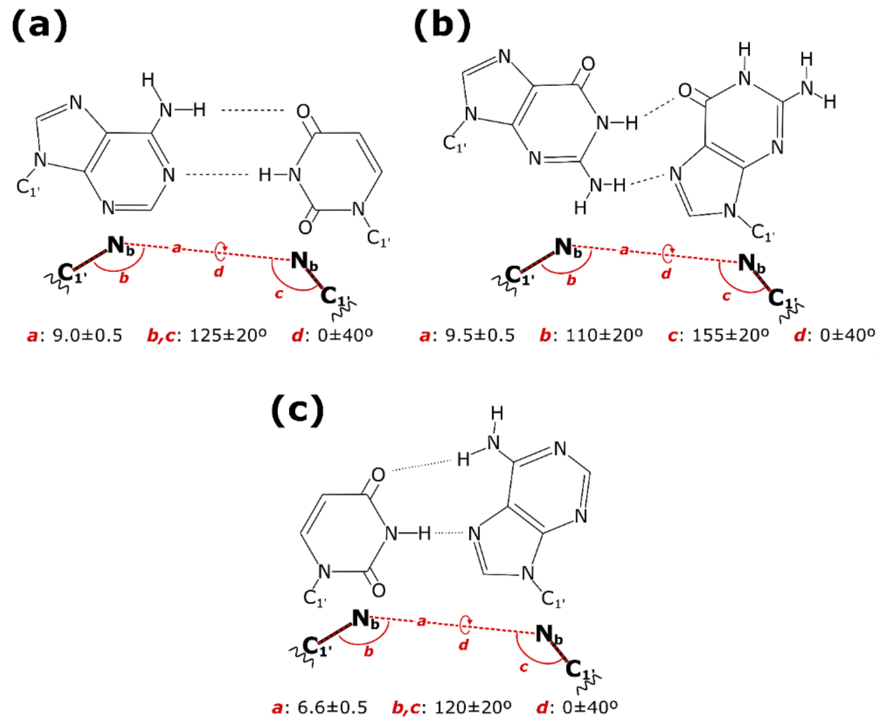

(a)
(b)
(c)
N_b
$C_{1'}$
a
b
c
d
a: 9.0±0.5 b,c: 125±20° d: 0±40°
a: 9.5±0.5 b: 110±20° c: 155±20° d: 0±40°
a: 6.6±0.5 b,c: 120±20° d: 0±40°

Figure 3.

Geometric criteria used in defining (a) Watson-Crick base pairing interaction (41), (b) G-G Hoogsteen interaction (39, 40), and (c) purine-pyrimidine Hoogsteen interaction (38, 39).

Evaluating Conformational Ensembles of CNG Repeats

Using the factorization scheme described above and a library of the entropy values of the irreducible elements, the entropy of the conformational ensemble of any RNA sequence can be evaluated. We illustrate this using the CNG repeat sequence 5'-NG(CNG)$_8$CN-3' as an example. Figs. 4(a) through (d) show four possible secondary structures of this sequence, and their corresponding diagrammatic representations are shown next to each. If we consider only WC base pairs, the longest uninterrupted canonically paired duplex length in CNG repeat sequences is only 2 base pairs (bp). These are highlighted by the blue boxes in Fig. 4. In the corresponding graphs, these are represented by blue dots. Since the structure in (a) has four 2-bp duplexes whereas (b) only has three, the total vertex free energy of (a) should be lower because base pairs are stabilizing. But on the other hand, (b) has fewer secondary structural constraints than (a), and therefore (b) is expected to have a more favorable conformational entropy. In an equilibrium ensemble of this sequence, we expect a thermodynamic competition between maximizing the number of vertices versus maximizing the diversity of the conformational ensemble. Furthermore, assuming 1-nt hairpins are unstable, the structure in (a) is the only conformation consistent with the diagram shown in (a). However, for structure (b), there are multiple alternative structures consistent with the graph shown in (b). These alternative structures can be obtained by permuting the junctions among the various positions along the sequence. For example, permuting the two junctions in the asymmetric bulge leads to a different structure without affecting its topology. Also, transposing a sub-segment within one junction with another junction produces a different structure without altering the topology. For example, one can remove a single (CNG) unit from the hairpin and transpose it into the first junction, making both 4-nt long, to derive a new structure with a symmetric bulge instead of the asymmetric one in (b) without altering the topology. The entropy associated with the configurational diversity of the topological class represented by the graph in Fig. 4(b) is therefore higher than (a), and this favors (b) over (a). In addition to (a) and (b), there are many other structures for the same sequence which belong to other topological classes. Fig. 4(c) and (d) show two additional examples. The structure in (c) corresponds to a three-way junction, while the structure in (d) consists of a pseudoknot plus a hairpin. Whereas (a) and (b) have different number of 2-bp duplex units, structures (b), (c) and (d) all have three duplexes. Because of this, structures (b), (c) and (d) have approximately the same vertex free energy and their competition for relevance within the conformational ensemble of this sequence is controlled by entropy alone. The goal of this study is to quantify the size and diversity of these CNG repeat ensembles using the diagrammatic techniques described above. Notice that each structure is characterized by a certain number of nucleotide (nt) units ℓ distributed over the unpaired regions among the loops and junctions, which in the graphs are associated with E edges. We will see that the problem of calculating the entropy is equivalent to finding all the possible ways of distributing the ℓ unpaired nucleotides over the E edges in the diagram.

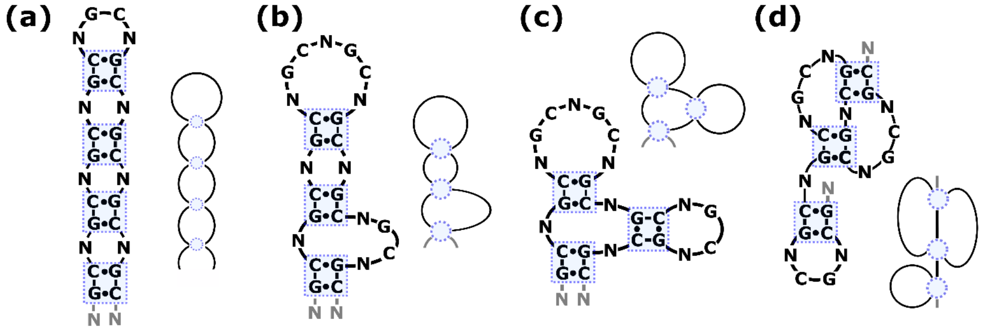


Figure 4.

Examples of different structures of the (CNG)$_M$ repeat sequence belonging to distinct topological classes for M=9.

To evaluate the volume of the conformational ensemble of a (CNG)$_M$ repeat sequence of a certain length M, we divide the ensemble into subsets according to the total degree of the graphs. Recall that the total degree D of a graph is equal to the sum of the degrees over all its vertices. Using this definition, the total degree of the graph in Fig. 4(a) is 16 because each vertex corresponding a duplex is degree 4, since 4 edges emanate from it. On the other hand, the graphs in (b), (c) and (d) all have total degree $D = 12$, because of the three duplexes present in each of those structures. Earlier, we have also mentioned that the total degree of a graph is related to the number of edges E in it by $E = 1 + D/2$. Because of this, we now recognize that even though the graphs in Fig. 4(b), (c) and (d) all belong to distinct topological classes, they all have the same number of edges because they have the same total degree. The diagrams in Fig. 4(b), (c) and (d) all have 7 edges because they are all degree 12. Furthermore, since the vertices are fixed-length duplexes, the total lengths of all the edges for all diagrams of degree D are also the same for sequences containing the same number of (CNG) repeats M. For example, the structures in Fig. 4(b), (c) and (d) all have total edge lengths of $\ell = 16$ nt.

In the last paragraph, we described how the entropy of a certain class of diagrams is derived from the permutation of the edges and the transposition of subsegment lengths among the edges. Restating this more precisely in terms of combinatorics, the structural diversity of a certain topological class is related to the number of possible ways in which the total edge length in a structure consisting of ℓ nts can be distributed among the E edges in the diagram. Because of this, grouping diagrams by total degree is advantageous compared to grouping them according to topological class. Since diagrams of the same degree also have the same number of edges, the combinatoric problem is identical for diagrams across the same degree, regardless of which topological class they belong to. This allows us to recycle the solution of the same combinatorics problem on diagrams of many different classes, as long as their total degrees are the same. This also means that the intrinsic diversities of the different subsets of the ensemble represented by different topological classes of graphs belonging to the same total degree are identical. The only difference between two topological classes belonging to the same total degree lies in the conformational entropies of the irreducible elements, which are different for different types of secondary structures. For example, the probability of observing an 8-nt hairpin loop in the ensemble of all possible conformations is very different from that of observing an 8-nt loop inside a three-way junction, even though they are both loops of the same length. These different entropy values are supplied by the library we complied using the MC simulations described above.

We summarize our solution to the combinatorics problem with a few useful equations here. Referring to Fig. 4, notice that each edge segment has a minimum length of 1 nt and can vary only by multiples of 3 nt. Therefore, the length of the i-th edge in a diagram can be represented by $1 + 3j_i$, where j is a non-negative integer, and there are E of these. The chain contains M (CNG) repeats. We always assume both the 5' and 3' ends have a dangling N nucleotide, so the total length of the sequence is $(3M + 1)$ nt. A degree-D diagram has D nts in the duplexes because all base pairs here come in stacks of 2, so the total edge length is $\ell = (3M + 1) - D$. The number of edges is $E = 1 + \frac{1}{2}D$. In terms of this, $\ell = 3(M + 1 - E) + E$. Subtracting the minimum length of 1 nt for each of the E edges, the number of transposable nts is $3(M + 1 - E)$, but they must occur in 3-nt multiples. Therefore, the combinatorics problem is reduced to finding all sets of non-negative integers $\{j_1, j_2, \cdots j_E\}$ such that $\sum_{i=1}^{E} j_i = J \equiv (M + 1 - E) = (M - \frac{1}{2}D)$, which also implies that the maximum degree for a chain with M repeats is $2M$. The process of dividing up the nucleotides into the edges of the graph is equivalent to creating a string of E non-negative integers with zeroes allowed such that they sum to J. This is the problem of determining all weak compositions of J in combinatorics; for a graph which has E edges, this is the enumeration of all weak E-composition of J (20, 42). For the purposes of this study, the enumeration is done by brute force enumeration with the correct partitions being stored as a valid structure of the graph ensemble. The collection of valid structures α for each graph Ξ forms an ensemble with each structure contributing a weight, $\omega_\Xi(\alpha) = \exp\left(-\frac{F(\alpha)}{kT}\right)$, corresponding to its inherent free energy cost $F(\alpha)$ to the partition function of the graph ensemble, $Z(\Xi) = \sum \omega_\Xi(\alpha)$. For each graph ensemble, we can then define the ensemble-averaged conformational cost, $\Delta F(\Xi) = \sum \frac{\omega_\alpha}{Z(\Xi)} F(\alpha)$, and the entropy $\Delta S(\Xi) = -k_B \sum \frac{\omega_\alpha}{Z(\Xi)} \ln\left(\frac{\omega_\alpha}{Z(\Xi)}\right)$. All graph ensemble calculations were carried out for a RNA strand of 52 nucleotides corresponding to a NG-(CNG)$_{16}$-CN chain with $M = 17$.

RESULTS

Loop Initiation Entropies Involving Hoogsteen Pairs

Previously, we reported data for the entropies of initiating hairpin loops of different sizes seeded by WC pairs. To initiate a n-nt loop, the constraint associated with the base pair suppresses the conformational diversity of the backbone, which suffers an entropy penalty leading to a free energy cost $\Delta G(n)$ which increases with the loop length n. The free energy of formation of loops utilizing WC interactions are shown in Fig. 5 as the black filled circles. RNA triplexes and quadruplexes, on the other hand, must use noncanonical interactions on their Hoogsteen edges to form base pair. The geometric constraints on the backbone needed to facilitate a Hoogsteen pair is different for a purine-purine pair (e.g. G:G) or a purine-pyrimidine pair (e.g. A:C) are shown in Fig. 4(b) and (c), respectively. These Hoogsteen-specific geometric constraints produce higher free energy requirements for loop initiation compared to loops formed via WC interactions. Fig. 5 shows MC results for initiation free energies needed to form a loop using G:G Hoogsteen interactions (solid blue squares) and loops formed via purine:pyrimidine (R:Y) Hoogsteen interactions (open squares). Both types of Hoogsteen-pair mediated loops require higher free energy compared to WC-pair seeded hairpins.

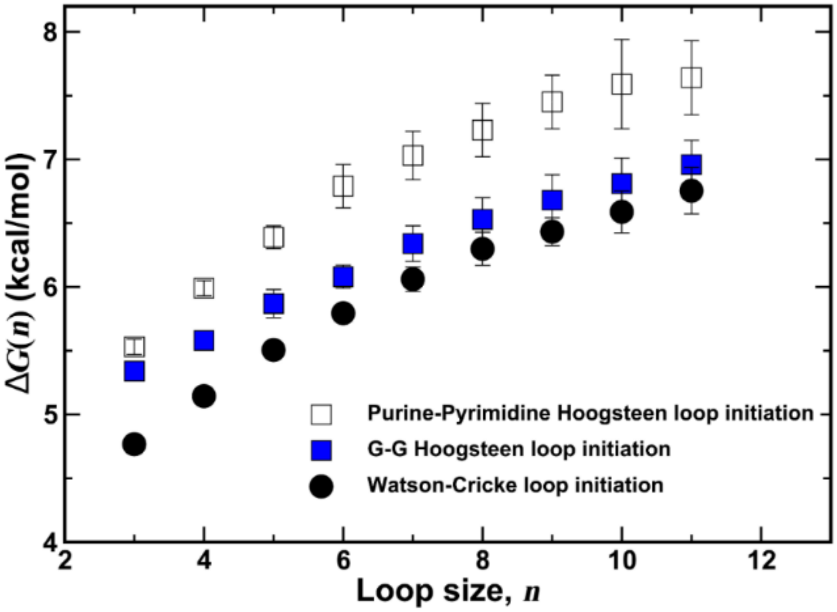


Figure 5.

Comparison of loop initiation using different sets of base pairing criteria. In comparison to the Watson-Crick initiation cost, the purine-pyrimidine Hoogsteen initiation costs are effectively shifted up by a constant. This is consistent with the difference in the targeted inter-base distance and the almost identical range of bond and torsion angles. The G-G Hoogsteen loop initiation, despite sharing the same inter-base distance, suffers a larger cost at small loop lengths that is associated with the base positioning and the different inter-base angles.

Quadruplexes

Quadruplex structures on DNA have been observed on d(GGG-NNN)$_n$ repeats (43). These G-quadruplexes typically consist of a triple-deck sandwich of four Gs on each layer, interacting with each other via G:G Hoogsteen pairs. The d(NNN) sequences act as linkers, connecting the vertices of the triple sandwich. Various linker topologies have been identified. These are exemplified by the structures found in PDB IDs 1KF1 (39) and 143D (40). In 1KF1, the linkers are threaded through the G-quadruplex structure connecting the bottom corner of one edge of the triple sandwich with the top of an adjacent edge. In 143D, the linkers are threaded by connecting either the bottom corner of one edge with the bottom corner of an adjacent edge, or the top corner with the top corner of an adjacent edge.

The type of quadruplexes that are most relevant to $(CNG)_M$ RNA repeats are the double-deck sandwich structures illustrated by Fig. 6. Instead of three layers, the quadruplex structure in Fig. 6 has only two layers. The linker topology shown in Fig. 6 follows a bottom to top threading pattern, analogous to 1KF1. $(CXG)_n$ repeats where X=G can potentially produce quadruplex structures of the type shown in Fig. 6, with each linker being either 1-nt (-C-), 4-nt (-CGGC), 7-nt (-CGGCGGC-) in length, or even longer.

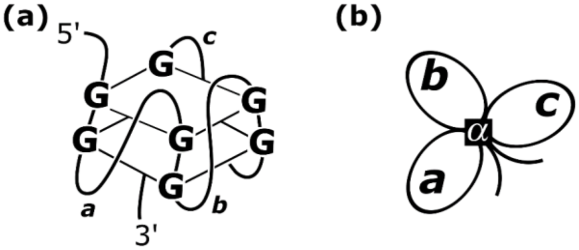


Figure 6.

(a) A possible quadruplex structure relevant to CNG repeat sequences. The structure's entropy is determined by the three loops labeled *a*, *b* and *c*. (b) Diagrammatic representation of a quadruplex, showing its dependence on the three loop lengths *a*, *b* and *c*, as well as the number of layers α.

There are three linker loops in a quadruplex structure. These are labeled *a*, *b* and *c* in Fig. 6 in the 5' to 3' direction. The entropic free energy costs for initiating the first loop *a*, the second loop *b* and the third loop *c* to connect the G on the bottom of one edge of the quadruplex to the G on the top of the next edge are tabulated in Table 1 for a double-deck quadruplex structure. In the MC simulations, loop *a* was initiated first. After this loop was formed, the free energy of initiating loop *b* was computed by holding the first two edges of the quadruplex fixed. After loop *b* was formed, the free energy of initiating loop *c* was then computed by holding the first three edges of the quadruplex fixed. The results in Table 1 suggest that as the linked loops get longer, the free energy cost of forming the loop also increases. This trend is not dissimilar to that observed in Fig. 5 for the hairpin initiation free energies. But as the quadruplex structure was assembled, the loop free energies also become progressively higher from *a* to *b* to *c*. This is presumably due to increased steric congestion in the core of the quadruplex structure, making it more difficult for loop *b* to form compared to *a*, and in turn more difficult for loop *c* to form compared to *b*. For linker loop *c*, the frequency of observing its formation in the MC simulations were too rare to be able to determine their free energies accurately for lengths > 5 or < 2 nt, and these have been left out of Table 1. In addition to the loops *a*, *b* and *c*, there are entropic penalties associated with constraining the backbone to the four edges of the quadruplex. The total free energy cost for this is also given in Table 1.

Loop Length (nt)	ΔG loop *a* (kcal/mol)	ΔG loop *b* (kcal/mol)	ΔG loop *c* (kcal/mol)	Other ΔG (kcal/mol)
1	4.2	5.3	-	all edges 12.6
2	4.9	6.8	-	
3	5.5	7.0	7.0	
4	6.0	7.4	8.4	
5	6.3	7.2	6.9	
6	7.0	7.4	-	
7	7.3	8.7	-	

Table 1. Cost of Initiating the Loops of Fig. 6.

Initiation free energies for the *a*, *b* and *c* loops inside a double-deck quadruplex structure from MC simulations. The typical statistical error on each value is approximately ± 0.05 kcal/mol.

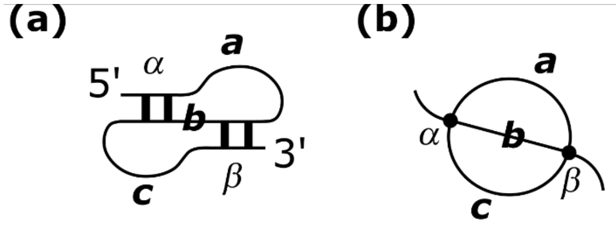
Previously, we found that vertices in graphs such as the helix in a hairpin or a stem in a 2-, 3-, or 4-way junction divide diagrams into approximately independent pieces. For the quadruplex, this independence is not strictly obeyed, because as Table 1 shows, the initiation free energies of the *a*, *b* and *c* loops are asymmetric with respect to exchange and they are no longer independent of each other. To accommodate this, we can modify the graph factorization scheme by simply redefining the entire quadruplex structure together with its *a*, *b*, and *c* loops as one irreducible element, instead of assuming the loops are separable from the quadruplex's core. Since quadruplexes typically have very limited loop lengths, this does not affect the validity or impact the utility of the graph factorization scheme described above.

Pseudoknots

The conformational entropy of a pseudoknot is determined by the lengths of the three loops a, b, and c, as well as the duplex lengths α and β. In the pseudoknot structures most relevant to CNG repeat sequences, α and β are 2 bp, the interhelix length b is 1 nt, and the loop lengths a and c are 1, 4, 7, ... An example of how such pseudoknots fit into a (CGN) repeat sequence is shown in Fig. 4(d). Prior studies in the literature by Cao and Chen suggest that the three loops of a pseudoknot can be treated independently (44–46) when considering their entropies. Our simulation results for the pseudoknot structures most relevant to CNG repeats corroborate this.

To calculate the conformational entropy costs for pseudoknots, we calculated the cost for each steps in a thermodynamic pathway that folds a free chain into the final pseudoknot structure, passing through a hairpin structure along the way (an example of such pathways for the a=b=c=1 case can be found in

Fig.S1 in the Supplemental Information). Consequently, the formation of the pseudoknot's three loops in our method is due to a single base pairing event which turns an existing hairpin structure into a pseudoknot, and this can happen on either the 5' side or the 3' side. Additionally, there are extra entropy costs for extending the helices to reach their target lengths α and β. In Fig. 7, we summarize the entropic cost and standard error of forming the entire pseudoknot for the four smallest pseudoknot structures relevant to $(CNG)_n$ repeats. The cost is calculated as the sum of costs to go from an open chain to the appropriate hairpin and then from the hairpin to the final pseudoknot structure relative to the cost of two 2-bp duplexes. Multiple pathways connecting the initial open chain and the final structure were used and the averages are shown in Fig. 7. The map of the pathway used for each of the structure can be found in the Supplemental Information as Fig. S1-S4, and the costs for each step of the pathways can be found in the Supplemental Information as Table S1-S4.



$b = 1,\ \alpha = \beta = 2$

a	*c*	ΔG (kcal)
1	1	13.2 (0.2)
1	4	13.9 (0.3)
4	1	14.0 (0.3)
4	4	14.7 (0.3)

Figure 7.

(a) The conformational entropy of a pseudoknot is determined by the lengths of the three loops labeled *a*, *b* and *c*, as well as the duplex lengths α and β. In the pseudoknot structures most relevant to CNG repeat sequences, α and β are both 2 nt, the interhelix length *b* is 1 nt, and the loop lengths *a* and *c* are 1, 4, 7, ... (b) Diagrammatic representation of a pseudoknot and calculated average cost for the four smallest pseudoknot structures relevant to CNG repeats.

Hairpins and Multiway Junctions

To carry out the ensemble calculations describe in the remainder of this paper, parameters for hairpin loop initiations as well as two-, three- and four-ways junctions are also required in addition to the library of data presented above. These have been reported in a previous paper (21), and since there are a number of details involved, these earlier results will not be repeated here (but loop initiation free energies using WC pairs are summarized in Fig. 5 compared to R:Y and G:G Hoogsteen pairs). With these results and the data set reported above, the library of loop entropy data needed for the calculations described in the rest of this paper is complete.

We will, however, recall just one particular piece of data, as this relates to the core free energy in a 2-bp duplex, which are the structural elements that provide the stabilization against loop entropy costs inherent in all the structures in the ensemble. In the graphs, 2-bp duplexes are represented by dots. Our previous calculations showed that in a hairpin, the backbone entropy cost, in terms of the free energy associated with propagating the stem once a loop has been initiated, is 5.22 kcal/mol/bp. Since 2-bp

GC|CG duplexes are the only canonically paired motifs in a CNG repeat ensemble, this stem propagation cost must be offset by the intrinsically favorable free energy gain from the pairing and stacking interactions among the bases. Nearest-neighbor estimates for the net free energy of formation of a GC|CG duplex from Turner et al. (24) is -3.42 kcal/mol. Therefore, a reasonable estimate for the stabilization energy due to the core of a GC|CG duplex without the entropy costs associated with the backbone is -8.64 kcal/mol, leading to the overall free energy per duplex core of $\Delta G_{\mathrm{dc}} = -3.42$ kcal/mol.

Ensembles of CNG Repeats

Given a (CNG)$_M$ repeat sequence of fixed length M, we first partitioned conformational space according to the degree of the diagrams, and then by collecting all accessible folded structures which share the same graph representation into a subset ensemble. For the purposes of this study, we have focused on graphs with 2, 3, and 4 vertices. For each structure within a subset ensemble, we calculated its weight using the entropic costs of all its irreducible elements as determined by the library derived from our MC data. The ensemble's partition function of the subset represented by a graph Ξ was then used to calculate its sub-ensemble average conformational cost $\Delta F(\Xi)$, its entropy $\Delta S(\Xi)$, and then the free energy, according to $\Delta G(\Xi) = \Delta F(\Xi) - T\Delta S(\Xi) + \Delta G_0$, where ΔG_0 is the intrinsic free energy associated with the paired regions. The entropy $\Delta S(\Xi)$ is a measure for the diversity of this subset, whereas $\Delta G(\Xi)$ determines the overall thermodynamic stability of this subset relative to other subsets.

The graph ensemble calculations reported below have been carried out for a RNA strand of 52 nucleotides corresponding to a NG-(CNG)$_{16}$-CN chain with $M = 17$. These were done by grouping graphs according to their total degree D, equal to the sum over the degrees of all vertices. Fig. 8 shows all graphs with total degree $D = 8$ and the calculated values of $\Delta F(\Xi)$ and $\Delta G(\Xi)$ in kcal/mol, and $\Delta S(\Xi)/R$ for each. We derived these diagrams from the list of graphs enumerated by Schlick et al. (17, 19), after removing those containing structures for which we have no corresponding data or those requiring non-secondary structural motifs to form. Three motifs are present with total degree of 8: hairpins, two-way junctions, and pseudoknots. The thermodynamic stabilities of the three subsets reported in their ΔG values include the intrinsic stabilization provided by the free energy in the duplexes, $\Delta G_0 = 2 \times -3.42$ kcal/mol. The structures represented by the graphs in Fig. 8 are either marginally stable or unstable.

Data are shown in Fig. 9 for all graphs with total degree $D = 12$ and in Fig. 10 for total degree $D = 16$. Going to higher degrees, the number and diversity of the graphs quickly proliferate, and because of this, explicit enumeration is only possible for total degrees that are not too high. Again, intrinsic stabilization provided by the free energy in the duplexes are included, and ΔG_0 is 3×-3.42 kcal/mol in Fig. 9 or 4×-3.42 kcal/mol in Fig. 10, except for the quadruplex. The intrinsic stabilization free energy due to the quadruplex core is difficult to ascertain. Thermodynamic studies suggest that G-quadruplexes from human telomeric DNAs are marginally stable (47). Since the quadruplex structures most relevant to CNG repeats contain only two layers instead of three, we expect an even weaker stability. Compared to a typical hairpin with a 2-bp GC|CG stem, we do not expect a 2-layer quadruplex to be more prevalent. In the last subsection, using Turner's GC|CG stability of -3.42 kcal/mol, we estimate the stabilization energy due to the core of a GC|CG duplex without the entropy costs associated with the backbone to be -8.64 kcal/mol. The stability of a GG|CG mismatch in DNA has been estimated from melting studies to be $+2.5$ to 2.9 kcal/mol higher than GC|CG (48), whereas a GGC|CGG mismatch is ~ $+2.2$ kcal/mol higher than CGG|GGC in RNA (49). That renders the stabilization energy due to the core of a 2-layer G quadruplex > $4 \times (-8.64 + 2.5) = -24.56$ kcal/mol. For the quadruplex structure in Fig. 10, we have used this estimate in combination with the entropic penalties associated with constraining the backbone to the four edges of the quadruplex given in Table 1, to arrive at a $\Delta G_0 \sim -11.95$ kcal/mol, which is likely a lower bound.

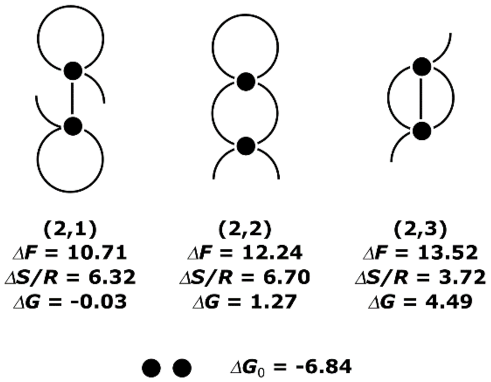


Figure 8.

All graphs for $(CNG)_{17}$ at total degree 8, their RAG-ID (19) and corresponding ensemble-averaged cost, entropy, and the graph free energy. ΔF and ΔG are in kcal/mol. ΔS are reported in units of R, the gas constant.

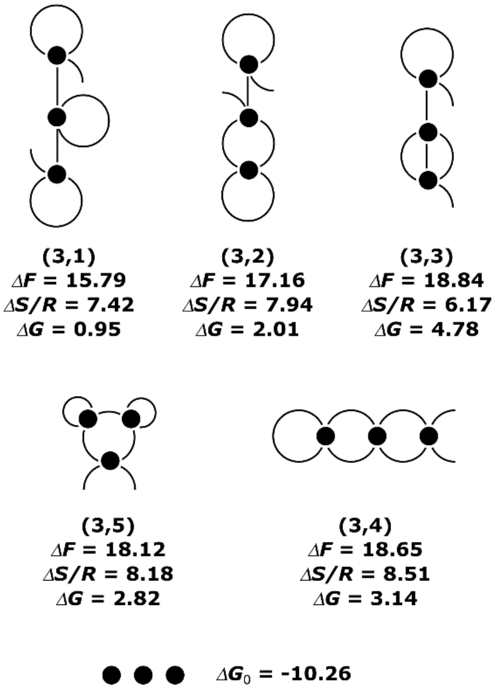


Figure 9.

All graphs for (CNG)$_{17}$ at total degree 12, their RAG-ID and corresponding ensemble-averaged cost, entropy, and the graph free energy. ΔF and ΔG are in kcal/mol. ΔS are reported in units of R, the gas constant.

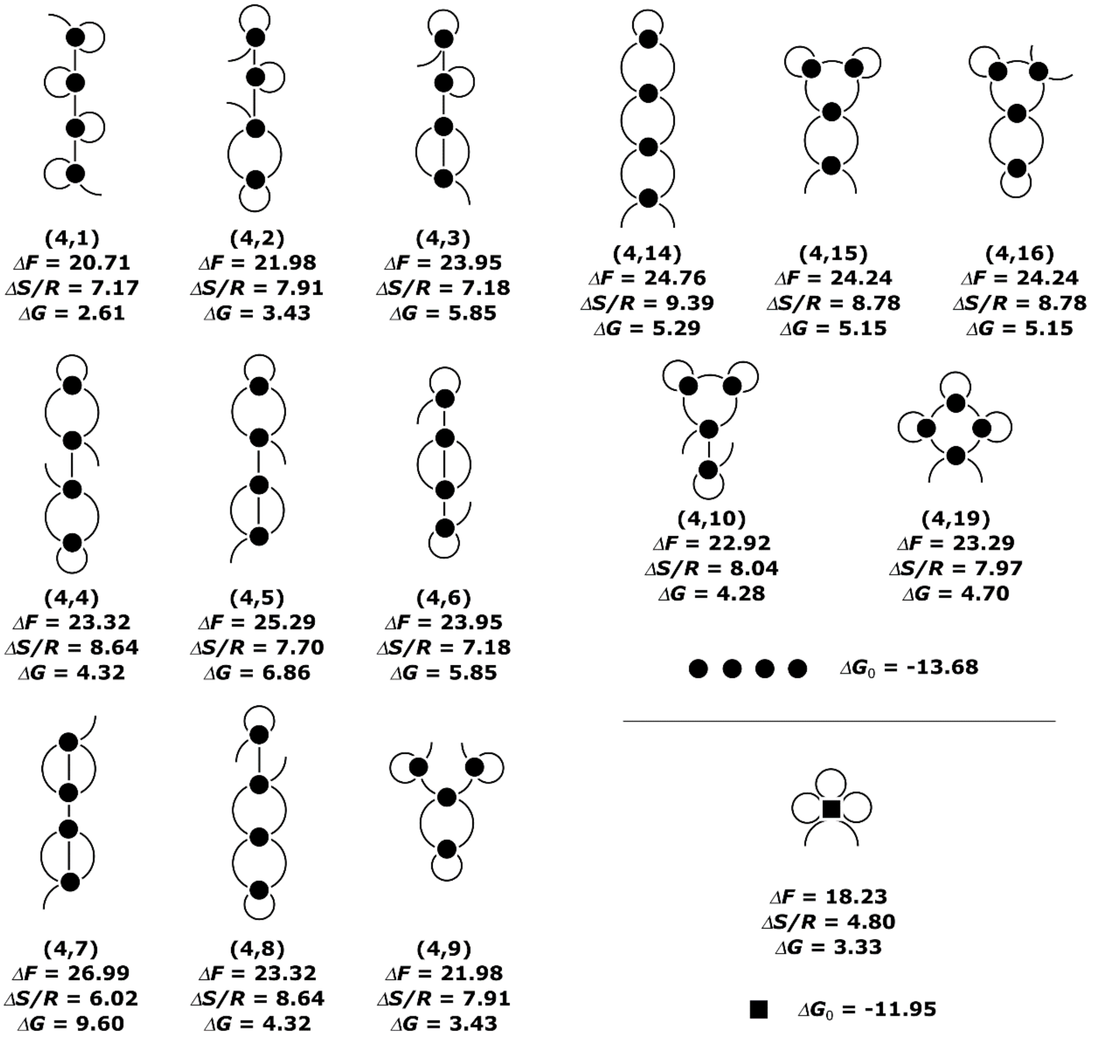


Figure 10

All graphs for $(CNG)_{17}$ at total degree 16, their RAG-ID and corresponding ensemble-averaged cost, entropy, and the graph free energy. ΔF and ΔG are in kcal/mol. ΔS are reported in units of R, the gas constant. The list also includes a quadruplex structure, with an estimate for the intrinsic ΔG of the core, referenced against the same standard state (four 2-bp duplexes) used for the rest of the structures in this figure.

DISCUSSION

Data in Figs. 8, 9 and 10 reveal the basic characters of the structural ensembles typical of CNG repeat sequences. While this direct enumeration approach is limited to graphs of low total degrees, the results for D = 8, 12 and 16 demonstrate some central features that allow us to make some projections about graphs of higher degrees. The results shown in Figs. 8, 9 and 10 are for one example $(CNG)_M$ sequence, where the number of repeats M = 17.

At total degree D = 8, Fig. 8 reveals that the graph with the lowest overall free energy is (2,1). While this is just one graph, it is important to remember that a large number of conformations are represented by it, where the segments in the graph can have variable lengths but they are restricted in such a way that their sum must equal the total length of the chain. The entropy $\Delta S/R$ listed under the graph is a measure for the number of these conformations. The value $\Delta S/R = 6.32$ for graph (2,1) suggests that there are roughly $e^{6.32}{\sim}550$ conformations represented by it. The value ΔF is the sub-ensemble free energy cost associated with suppressing the backbone conformational degrees of freedom to force the chain to conform with the constraints implied by the vertices in the graph, and for (2,1) it is 10.71 kcal/mol. (Notice that for every conformation in this sub-ensemble the backbone conformational cost is different; therefore, the conformation tally of ${\sim}550$ as well as the cost $\Delta F = 10.71$ kcal/mol are ensemble averaged properties.) The overall free energy of graph (2,1) is $\Delta F - (RT)(\Delta S/R) = 10.71 - (0.616)(6.32) + \Delta G_0 = -0.03$ kcal/mol, where ΔG_0 is the intrinsic free energy associated with two 2-bp duplexes, equal to -6.84 kcal/mol as indicated in Fig. 8, and $T = 310$ K.

The graph that has the next higher overall free energy in Fig. 8 is (2,2), whose $\Delta G = 1.27$ kcal/mol, making the relevance of this subset of conformations roughly $e^{(-0.03-1.27)/RT} = 0.12$ that of graph (2,1) at T = 310 K where $(RT) = 0.616$ kcal/mol. The entropy of graph (2,2) is similar to (2,1). This is not surprising because their topologies are similar, except two segments in (2,2) are constrained into a 2-way junction, whereas in (2,1) one of them is constrained inside a hairpin and the other is free. Contrasting this to the graph (2,3) in Fig. 8, which contains a pseudoknot, $\Delta S/R$ is quite a bit lower for (2,3), even though (2,3) contains the same number of segments as (2,2). (Note that Eq.(1) dictates that the total number of edges E and the total degree D are related by $E = 1 + D/2$, all graphs in Fig. 8 necessarily have the same number of edges.) The reason why the structure containing the pseudoknot has a significantly lower entropy compared to (2,1) and (2,2) is related to the loop-length dependence in the cost function of a pseudoknot. Long loop lengths are suppressed in a pseudoknot compared to hairpins or 2-way junctions, and this leads to a lower diversity in the subensemble associated with graph (2,3) compared to (2,1) or (2,2). This also results in a higher overall ΔG for graphs containing pseudoknots.

While each graph represents a subensemble of the conformations at a certain total degree D and its entropy value $\Delta S/R$ reflects the diversity of that subset, an additional ensemble-level entropy is associated with the superset of graphs at each D. This ensemble-level entropy at a degree D is given by $\Delta S_D/R = -\sum_\Xi P(\Xi)\ln P(\Xi)$, where the normalized probability $P(\Xi)$ of graph Ξ is given by $P(\Xi) = e^{-\Delta G(\Xi)/RT}/\sum_\Xi e^{-\Delta G(\Xi)/RT}$. For the graphs in Fig. 8, $\Delta S_D/R$ comes out to be 0.35, suggesting that at the ensemble level, the information content in the superset of $D = 8$ graphs is roughly equivalent to just $e^{0.35}{\sim}1.4$ graphs. Furthermore, the overall free energy of the ensemble ΔG_D can be computed either from $e^{-\Delta G_D/RT} = Z = \sum_\Xi e^{-\Delta G(\Xi)/RT}$ or $\Delta G_D = \langle \Delta G\rangle - T\Delta S_D$, where $\langle \Delta G\rangle = Z^{-1}\sum_\Xi \Delta G(\Xi)e^{-\Delta G(\Xi)/RT}$, both of which yield $\Delta G_D = -0.10$ kcal/mol for all the graphs in Fig. 8. This suggests that the $D = 8$ graphs are collectively marginally stable relative to an open chain.

Moving to the $D = 12$ graphs in Fig. 9, the diversity of the graphs expands. The member of this superset with the lowest overall free energy ΔG is (3,1). The one with the highest free energy is again the structure with a pseudoknot. $D = 12$ is the lowest order at which a 3-way junction appears, as in (3,5).

Similar to the $D = 8$ superset in Fig. 8, the member with the highest subensemble diversity is (3,4), which has the highest entropy $\Delta S/R = 8.51$, corresponding to approximately $e^{8.51}{\sim}5000$ distinct conformations. The ensemble-level entropy for the set $D = 12$ is $\Delta S_D/R = 0.68$ corresponding to ~ 2.0 graphs. The overall free energy of this ensemble is $\Delta G_D = 0.81$ kcal/mol, indicating that the superset of $D = 12$ diagrams is close to neutral compared to the thermodynamics stability of an open chain.

Going to $D = 16$ in Fig. 10, the diversity of the graphs expands even further. Because of this rapid proliferation of the diagrams, enumerating diagrams one-by-one and computing their thermodynamics explicitly quickly becomes infeasible except for only the lowest few degrees. In a companion paper {xxx}, we describe an alternative diagrammatic theory to include diagrams up to infinity total degree, though at the expense of having to introduce some approximations.

All of the graphs in Fig. 10 contain four 2-bp duplexes, except the one with a quadruplex. The intrinsic free energy associated with four 2-bp duplexes is $\Delta G_0 = -13.68$ kcal/mol, which is indicated in Fig. 10. On the other hand, the intrinsic free energy of a 2-layer quadruplex core, according to the reasoning at the end of the last section, is estimated to be $\Delta G_0 {\sim} -11.95$ kcal/mol, which is likely a lower bound. Given this uncertainty, the free energy value of the quadruplex structure relative to the other graphs in Fig. 10 must be considered uncertain also, but the ΔF and $\Delta S/R$ values given for the quadruplex graph are certain, because they do not depend on the value of ΔG_0. Fig. 10 shows that the ΔG value for the quadruplex graph is 3.33 kcal/mol, and because of the uncertainty in ΔG_0, this is the likely the lower bound. Its entropy is low because of a reason similar to the pseudoknots – the quadruplex structure is confined to short loop lengths and reduces the conformational diversity of the subset. Notice that the quadruplex structure is only possible when the sequence is $(\mathrm{CGG})_{\mathrm{M}}$.

Similar to the graphs in Figs. 8 and 9, the graph with the lowest overall free energy is the type associated with graph (4,1) in Fig. 10. We will refer to graphs having this topology as “bubble diagrams”. For graph (4,1) in Fig. 10, the overall free energy is $\Delta G = 2.61$ kcal/mol. Notice that going from $D = 12$ in Fig. 9 to $D = 16$ in Fig. 10, the entropy of the bubble diagram actually decreases from $\Delta S/R = 7.42$ for graph (3,1) in Fig. 9 to 7.17 for graph (4,1) in Fig. 10. This indicates that when going to higher total degree D, the fixed length of the chain becomes a factor limiting the number of combinations of segment lengths that could fit into the total number of nucleotides on the sequence. However, this effect is also dependent on other features of the graphs. For example, the type of graphs with the highest entropy in both Fig. 9 and 10 are the “necklace diagrams”, exemplified by structures (3,4) and (4,14). Going from (3,4) to (4,14), the entropy value of the necklace diagram actually increases from $\Delta S/R = 8.51$ to 9.38, growing from ~5000 to 12000 configurations. Furthermore, some of the other partial necklace diagrams, such as (3,2) in Fig. 9 and (4,4) and (4,8), also increase in diversity going from low to higher order. In addition to these, the additional diagrams associated with 3- or 4-way junctions also seem to expand in diversity. While we do not have $D = 20$ data to compare, it is quite possible that the overall diversity of the ensemble will continue to proliferate when going to even higher orders. The ensemble-level entropy for the set $D = 16$ is $\Delta S_D/R$ =1.67 corresponding to ~ 5.3 graphs, and comparing this to $\Delta S_D/R$ =0.67 for $D = 12$, validates this observation. The overall free energy of this ensemble is $\Delta G_D = 2.15$ kcal/mol, indicating that the superset of $D = 16$ diagrams overall are less stable compared to the $D = 12$ ensemble. However, we believe as one continues to go to higher total degree and longer chains, the graph ensembles at higher total degrees will become progressively more favorable.

Finally, we can measure the overall size of the ensemble at each degree D by computing the total entropy of all the thermally accessible conformations. According to the additivity rule for entropy (50), this measure is conveyed by the system's total entropy $\Delta S_{\text{tot},D} = \langle \Delta S \rangle + \Delta S_D$, where $\langle \Delta S \rangle = Z^{-1} \sum_{\Xi} \Delta S(\Xi) e^{-\Delta G(\Xi)/RT}$, where Z, the partition function, has been defined above. For D = 8, 12 and 16, $\Delta S_{\text{tot},D}/R = 6.71$, 8.23 and 8.85, respectively, corresponding to ~ 800, 3800 and 7000 distinct conformations. We expect this to continue to proliferate going to higher degrees.

CONCLUSION

Our analysis of the graph ensembles of CNG repeat chains at the oligomer scale has provided a first look at a theoretical model for analyzing the structural diversity of trinucleotide repeat chains, as well as observations that will be germane to understanding RNA conformational ensemble. By using a graph factorization method and a data library that has been built from simulations, we were able to group accessible secondary structures together into subsets represented by graphs and have calculated metrics for their thermodynamic stability $\Delta G(\Xi)$, as well as the structure content $\Delta S(\Xi)$ of the subensembles. The results show that most structures are either thermodynamically marginally stable or unfavorable by several kcal/mol. The addition of helices—corresponding in our data to an increase in the total degree of a graph—incurs an additional cost that is offset by an increase in the number of structure accessible to the chain, and it is the balance between these two factors that determines the thermodynamic favorability of the structure. When the total degree (helices) increases, the dominant conformations are associated with the so-called bubble diagram—graphs composed of only hairpins connected by bridging unpaired segments, while the necklace diagrams tend to have higher entropy but also higher free energy costs. The results show that the extent to which the structural diversity of different classes of diagrams can grow as the total degree increases is also dictated by the chain length. Some structures, such as the bubble diagrams, begin to lose structural diversity as the total degree grows past $D = 16$ in the results, while others continue to proliferate.

Altogether, the results show that the structural diversity and propensities for different structural elements on CNG repeat chains are determined by an interplay between the length of the chain, the stabilizing strength of the helices, and the complexity of the graphs in the ensemble. In future studies, we will continue to explore different facets of these questions to understand the structural diversity and stability of long trinucleotide repeat chains on the length scale associated with the onset of TREDs.

AUTHOR CONTRIBUTIONS

CHM designed the study. ENHP carried out simulations, data collection and the calculations. ENHP and CHM wrote the manuscript.

ACKNOWLEDGEMENTS

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

The authors also thank Peter Z. Qin for helpful discussions and comments.

REFERENCES

1. Mirkin, S.M. 2004. Molecular Models for Repeat Expansions. *BIOCHEMISTRY AND MOLECULAR BIOLOGY*. 24.

2. 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.

3. Wells, R.D., R. Dere, M.L. Hebert, M. Napierala, and L.S. Son. 2005. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res*. 33:3785–3798.

4. Li, L.-B., and N.M. Bonini. 2010. Roles of trinucleotide-repeat RNA in neurological disease and degeneration. *Trends in Neurosciences*. 33:292–298.

5. Mirkin, S.M. 2006. DNA structures, repeat expansions and human hereditary disorders. *Current opinion in structural biology*. 16:351–8.

6. Mirkin, S.M. 2007. Expandable DNA repeats and human disease. *Nature*. 447:932–40.

7. Nelson, D.L., H.T. Orr, and S.T. Warren. 2013. The Unstable Repeats—Three Evolving Faces of Neurological Disease. *Neuron*. 77:825–843.

8. Timchenko, L.T., N.A. Timchenko, C.T. Caskey, and R. Roberts. 1996. Novel Proteins with Binding Specificity for DNA CTG Repeats And RNA Cug Repeats: Implications for Myotonic Dystrophy. *Hum Mol Genet*. 5:115–121.

9. 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.

10. Napierala, M., and W.J. Krzyzosiak. 1997. CUG Repeats Present in Myotonin Kinase RNA Form Metastable “Slippery” Hairpins. *J. Biol. Chem.* 272:31079–31085.

11. Kiliszek, A., R. Kierzek, W.J. Krzyzosiak, and W. Rypniewski. 2012. Crystallographic characterization of CCG repeats. *Nucleic Acids Res*. 40:8155–8162.

12. 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.

13. 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.

14. 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.

15. 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.

16. Coonrod, L.A., J.R. Lohman, and J.A. Berglund. 2012. Utilizing the GAAA Tetraloop/Receptor To Facilitate Crystal Packing and Determination of the Structure of a CUG RNA Helix. *Biochemistry*. 51:8330–8337.

17. Izzo, J.A., N. Kim, S. Elmetwaly, and T. Schlick. 2011. RAG: An update to the RNA-As-Graphs resource. *BMC Bioinformatics*. 12:219.

18. 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.

19. 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.

20. Walker, R. 1992. Implementing discrete mathematics: combinatorics and graph theory with Mathematica, Steven Skiena. Pp 334. 1990. ISBN 0-201-50943-1 (Addison-Wesley). *The Mathematical Gazette*. 76:286–288.

21. Mak, C.H., and E.N.H. Phan. 2018. Topological Constraints and Their Conformational Entropic Penalties on RNA Folds. *Biophysical Journal*. 114:2059–2071.

22. 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.

23. Turner, D.H. 1996. Thermodynamics of base pairing. *Curr. Opin. Struct. Biol.* 6:299–304.

24. 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.

25. Diamond, J.M., D.H. Turner, and D.H. Mathews. 2001. Thermodynamics of three-way multibranch loops in RNA. *Biochemistry*. 40:6971–6981.

26. Mak, C.H. 2015. Atomistic Free Energy Model for Nucleic Acids: Simulations of Single-Stranded DNA and the Entropy Landscape of RNA Stem–Loop Structures. *J. Phys. Chem. B*. 119:14840–14856.

27. Mak, C.H. 2008. RNA conformational sampling. I. Single-nucleotide loop closure. *J. Comput. Chem.* 29:926–933.

28. Mak, C.H., W.-Y. Chung, and N.D. Markovskiy. 2011. RNA conformational sampling: II. Arbitrary length multinucleotide loop closure. *J. Chem. Theory Comput.* 7:1198–1207.

29. Mak, C.H., T. Matossian, and W.-Y. Chung. 2014. Conformational entropy of the RNA phosphate backbone and its contribution to the folding free energy. *Biophys. J.* 106:1497–1507.

30. Mak, C.H., L.L. Sani, and A.N. Villa. 2015. Residual Conformational Entropies on the Sugar–Phosphate Backbone of Nucleic Acids: An Analysis of the Nucleosome Core DNA and the Ribosome. *J. Phys. Chem. B*. 119:10434–10447.

31. Weeks, J.D., D. Chandler, and H.C. Andersen. 1971. Role of repulsive forces in determining the equilibrium structure of simple liquids. *J. Chem. Phys.* 54:5237–5247.

32. Mak, C.H. 2016. Unraveling Base Stacking Driving Forces in DNA. *J. Phys. Chem. B*. 120:6010–20.

33. Hummer, G., S. Garde, A.E. Garcia, A. Pohorille, and L.R. Pratt. 1996. An information theory model of hydrophobic interactions. *Proc. Natl. Acad. Sci. USA*. 93:8951–8955.

34. Rury, A.S., C. Ferry, J.R. Hunt, M. Lee, D. Mondal, S.M.O. O'Connell, E.N.H. Phan, Z. Peng, P. Pokhilko, D. Sylvinson, Y. Zhou, and C.H. Mak. 2016. Solvent Thermodynamic Driving Force Controls Stacking Interactions between Polyaromatics. *J. Phys. Chem. C*. 120:23858–23869.

35. Henke, P.S., and C.H. Mak. 2014. Free energy of RNA-counterion interactions in a tight-binding model computed by a discrete space mapping. *J. Chem. Phys.* 141:08B612_1.

36. Mak, C.H., and P.S. Henke. 2012. Ions and RNAs: free energies of counterion-mediated RNA fold stabilities. *J. Chem. Theory Comput.* 9:621–639.

37. Abrescia, N.G.A., A. Thompson, T. Huynh-Dinh, and J.A. Subirana. 2002. Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. *PNAS*. 99:2806–2811.

38. Vedula, L.S., J. Jiang, T. Zakharian, D.E. Cane, and D.W. Christianson. 2008. Structural and mechanistic analysis of trichodiene synthase using site-directed mutagenesis: Probing the catalytic function of tyrosine-295 and the asparagine-225/serine-229/glutamate-233–Mg2+B motif. *Archives of Biochemistry and Biophysics*. 469:184–194.

39. Parkinson, G.N., M.P.H. Lee, and S. Neidle. 2002. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*. 417:876–880.

40. Wang, Y., and D.J. Patel. 1993. Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. *Structure*. 1:263–282.

41. Olson, W.K., M. Bansal, S.K. Burley, R.E. Dickerson, M. Gerstein, S.C. Harvey, U. Heinemann, X.J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.S. Tung, E. Westhof, C. Wolberger, and H.M. Berman. 2001. A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of molecular biology*. 313:229–237.

42. Richmond, B., and A. Knopfmacher. 1995. Compositions with distinct parts. *Aeq. Math.* 49:86–97.

43. Gilbert, D.E., and J. Feigon. 1999. Multistranded DNA structures. *Current Opinion in Structural Biology*. 9:305–314.

44. Cao, S., and S.-J. Chen. 2005. Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA*. 11:1884–1897.

45. Cao, S., and S.-J. Chen. 2006. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res.* 34:2634–2652.

46. Cao, S., and S.-J. Chen. 2009. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA*. 15:696–706.

47. Lane, A.N., J.B. Chaires, R.D. Gray, and J.O. Trent. 2008. Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res*. 36:5482–5515.

48. Tikhomirova, A., I.V. Beletskaya, and T.V. Chalikian. 2006. Stability of DNA Duplexes Containing GG, CC, AA, and TT Mismatches. *Biochemistry*. 45:10563–10571.

49. Kierzek, R., M.E. Burkard, and D.H. Turner. 1999. Thermodynamics of Single Mismatches in RNA Duplexes. *Biochemistry*. 38:14214–14223.

50. Taneja, I.J. 1989. On Generalized Information Measures and Their Applications. In: Hawkes PW, editor. Advances in Electronics and Electron Physics. Academic Press. pp. 327–413.