Salt-Dependent Self-Association of Trinucleotide Repeat RNA Sequences

Hiranmay Maity,[†] Hung T. Nguyen,[‡] Naoto Hori,[¶] and D. Thirumalai[*,†,§]

†*Department of Chemistry, University of Texas at Austin, Texas, USA, 78712*

‡*Department of Chemistry, The State University of New York at Buffalo, NY, USA,14260*

¶*School of Pharmacy, University of Nottingham, Nottingham, NG72RD, United Kingdom*

§*Department of Physics, University of Texas at Austin, Texas, USA, 78712*

E-mail: dave.thirumalai@gmail.com

Abstract

Repeat RNA sequences self-associate to form condensates. Simulations of a coarse grained Single-Interaction Site model for $(\mathrm{CAG})_{\mathrm{n}}$ ($n = 30$ and 31) show that the salt-dependent free energy gap, ΔG_S, between the ground (perfect hairpin) and the excited state (slipped hairpin (SH) with one CAG overhang) of monomer for (n even) is the primary factor that determines the rates and yield of self-assembly. For odd n, the free energy (G_S) of the ground state, which is a SH, is used to predict the self-association kinetics. As the monovalent salt concentration, C_S, increases ΔG_S and G_S increase, which decreases the rates of dimer formation. In contrast, ΔG_S for shuffled sequences, with the same length and sequence composition as $(\mathrm{CAG})_{31}$, is larger which suppresses their propensities to aggregate. Although demonstrated explicitly for (CAG) polymers, the finding of inverse correlation between the free energy gap and RNA aggregation is general.

TOC Graphic

JPCL_TOC.pdf

A series of experiments[1–8] have established that low complexity repeat RNA sequences, such as $(\mathrm{CAG})_\mathrm{n}$ and $(\mathrm{CUG})_\mathrm{n}$, (n is the number of repeat units) undergo phase separation. In the two phase region, the high density droplet coexists with the sol or low density dispersed phase. The qualitative features of the phase separation may be understood using the venerable Flory-Huggins theory,[9,10] although in RNA temperature is not as relevant as salt concentration.

Besides their intrinsic biological interest, the transcribed products of $(\mathrm{CAG})_\mathrm{n}$ and $(\mathrm{CUG})_\mathrm{n}$ are implicated in Huntington's disease, muscular dystrophy and amytropic lateral sclerosis.[11–13] Recently, using coarse-grained simulations and theoretical arguments, we provided a conceptual framework for describing condensate formation in repeat nucleotide sequences.[14,15] The driving force for self association arises both from favorable intermolecular Watson-Crick (WC) base pair formation as well as the degeneracy associated with a large number of ways such base pairs can form[14,16] in a droplet. In a recent account,[15] we showed that the propensity of repeat RNA polymers to aggregate can be inferred from the free energy spectrum of the monomer. For even n the ground state of $(\mathrm{CAG})_\mathrm{n}$ polymer is a perfect hairpin (PH) that is stabilized by base stacking and Watson-Crick base pair formation. In this case, we showed that the propensity for self-association between $(\mathrm{CAG})_\mathrm{n}$ polymer is determined by the free energy gap, ΔG_S, between the ground state (GS) and the excited state in which at least one (CAG) unit is exposed, resulting in slipped hairpin (SH) states. The self-association propensity decreases as ΔG_S increases. For odd n, the GS is already in the SH state, which enhances the propensity to self-associate. Our theoretical prediction that $(\mathrm{CAG})_{2\mathrm{n}+1}$ should have higher tendency to aggregate than $(\mathrm{CAG})_{2\mathrm{n}}$ was validated using computer simulations at 0.1 M monovalent salt concentration.[15] A recent all atom molecular dynamics simulations of homopolymeric RNA sequences[17] have also suggested that single chain properties could reveal the propensity to undergo phase separation.

The spectrum of states sampled by the $(\mathrm{CAG})_\mathrm{n}$ polymers may be altered by varying the external conditions. Here, we explored the extent to which aggregation of $(\mathrm{CAG})_\mathrm{n}$ changes as the salt concentration is varied. We simulated the two sequences $(CAG)_{30}$ and $(CAG)_{31}$ both of which undergo phase separation at high densities. The population of aggregation prone hairpin state of the RNAs is suppressed as the salt concentration, C_S, increases from 0.15 M to 0.5 M. The rate of formation of dimer decreases with an increase in C_S, which is in accord with the theory that ΔG_S increases with increasing C_S. In contrast, the propensity to self-associate decreases in shuffled sequences with identical composition because ΔG_S increases substantially.

Salt modulates the population of different hairpin structures: We performed low friction Langevin dynamics simulations (see Supporting Information (SI)) at 37°C by varying the salt concentration, C_S, from 0.15 M to 0.5 M. The secondary structures obtained at these conditions are classified broadly into stem-loop or hairpin-like structures. The terminal nucleotides in the hairpin-like structures are spatially close. Single molecule Foster Resonance Energy Transfer (smFRET) spectroscopy characterized the structures of trinucleotide repeat DNA hairpins ($(CTG)_n$[19] and $(CAG)_n$[20]). It was found that the most populated state for $(CAG)_{14}$ (the $5'$ and $3'$ ends are close) is a perfect hairpin whereas the ground state of $(CAG)_{15}$ is slipped (FRET efficiency is lower compared to the ground state of $(CAG)_{14}$, thus exposing one CAG unit).

Because the FRET efficiency is related to the distribution of the end-to-end distance, we calculated $P(R_{ee})$ for both $A(CAG)_{30}A$ and $A(CAG)_{31}A$ at $C_S = 0.15$ M and 0.5 M (Figures 1A and 1B). We find that $P(R_{ee})$ has more than one peak indicating that multiple stem-loop conformations are sampled. The peak at $R_{ee} = 1.45$ nm is approximately at the equilibrium base pair distance (≈ 1.38 nm) between two beads, showing that the terminal nucleotides are in spatial proximity. The peaks located at $R_{ee} > 1.45$ nm are either due to break down of terminal G-C base pair (bp) or slippage in the strands. As the C_S value is increased to 0.5 M from 0.15 M, the amplitude of the first peak in $P(R_{ee})$ increases for both the sequences, which implies that increment in salt concentration stabilizes the base pair strongly.

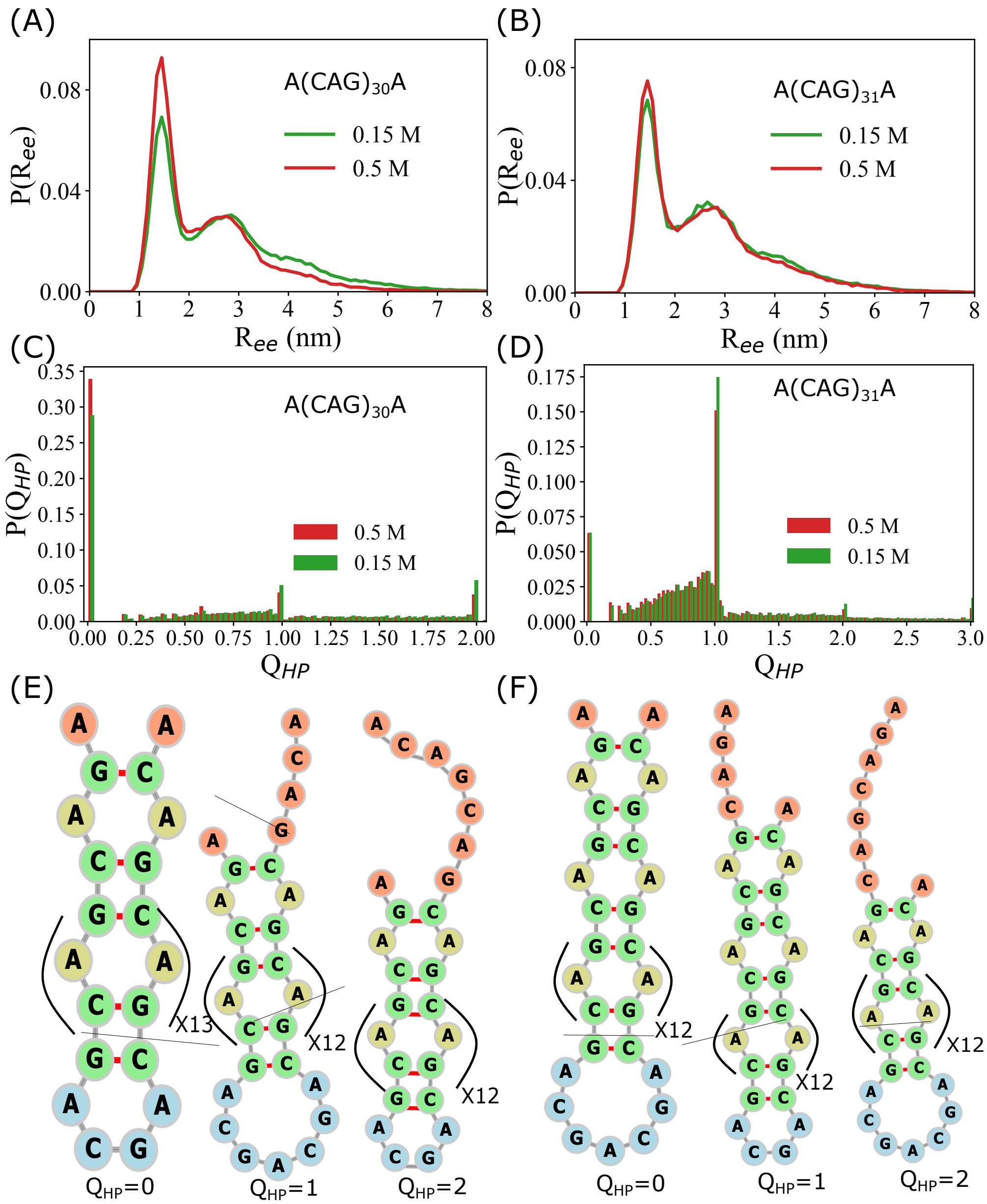

Figure 1: Characterization of hairpin structures: (A) Distribution, $P(R_{ee})$, of the end-to-end distance, R_{ee}, for A(CAG)$_{30}$A at salt concentrations, $C_S = 0.15$ M and 0.5 M are in green and red, respectively. (B) $P(R_{ee})$ as a function of R_{ee} for A(CAG)$_{31}$A at $C_S = 0.15$ M and 0.5 M are in green and red, respectively. Multiple peaks in the $P(R_{ee})$ are signatures of distinct stem-loop conformations in the ensemble of hairpin structures. (C) The probability distribution $P(Q_{HP})$ of Q_{HP} for (CAG)$_{30}$ at $C_S = 0.5$ M and 0.15 M are in red and green, respectively. As C_S increases the probability of $Q_{HP} = 0$ increases whereas $P(Q_{HP})$ with $Q_{HP} > 0$ decreases. (D) $P(Q_{HP})$ for (CAG)$_{31}$ at $C_S = 0.15$ M (green), 0.5 M (red) shows that the ground state population of (CAG)$_{31}$ decreases, i.e. $P(Q_{HP})$ with $Q_{HP} = 1$ decreases. (E) Hairpin structures with $Q_{HP} = 0$, 1, and 2 for (CAG)$_{30}$. (F) Same as (E) except the results are for (CAG)$_{31}$. The hairpin structures are generated using forna.[18]

In order to elucidate the microscopic structures of the stem-loop conformations, we computed the order parameter, Q_{HP}, for a given conformation using Eq. 2 by accounting for the arrangements of the bps in the stem region. Q_{HP} measures the deviation in the arrangement of base pairs in the stem region relative to the the arrangement in the perfect hairpin (PH) structure in which there is no mismatches, except for the unavoidable A-A mismatches. An ensemble of structures with $Q_{HP} = 0$ is, therefore, identical to the PH (Figure 1E and 1F). The set of bps, S_{bp}, representing the stem of a PH can be expressed as, $S_{bp} = (i, j) : i + j = N_T + 1$, where, i and j are the indices of the nucleotides forming the base pair, and N_T is the number of nucleotides in the sequence. Hairpins with $Q_{HP} = m$, where, m is a positive integer, signify the strand slippage by m repeat units of CAG from either the $5'$ or $3'$ end. Fractional values of Q_{HP} indicate the formation of one or more than one bulge at the stem region of the hairpins (Figure S3).

To investigate the effects of salt on the population of different hairpin states, we calculated the distribution, $P(Q_{HP})$, of Q_{HP} at C_S ranging between 0.15 and 0.5 M (Figure 1C, 1D and S1 in SI). The most populated structures or the ground state (GS) conformations of $(CAG)_{30}$ is a PH, whereas it is a slipped hairpin (SH) with one unit of CAG overhang at the terminal for $(CAG)_{31}$. Alternation in GS conformation between PH and SH on going from an even to an odd repeats has been observed in both repeat DNA sequences in experiments[20,21] and in simulations[15] of repeat RNA sequences. Although, the GS remains PH (SH) for $(CAG)_{30}$ ($(CAG)_{31}$) at all C_S values, its population changes as C_S value is varied. For instance, the population of the PH of $(CAG)_{30}$ changes to ≈ 0.33 from ≈ 0.28 as C_S is increased to 0.5 M from 0.15 M. Notably, there is a decrease in $P(Q_{HP} = 2)$ and $P(Q_{HP} = 1)$, which shows that the population of slipped hairpin decreases as C_S increases. Interestingly, we also found a suppression in the population in $P(Q_{HP} = 1)$ for $(CAG)_{31}$. Combining the results for $(CAG)_{30}$ and $(CAG)_{31}$, we conclude that slippage in strands is reduced as C_S increases.

Experiments probing the slippage dynamics of $(CAG)_{14}$ and $(CAG)_{15}$ DNA sequences have observed similar effects on the population of slipped hairpin.[20]

Free energy spectrum of the (CAG)n polymers as a function of C_S: Our theory is that the free energy gap, ΔG_S, separating the GS and the excited state, which contains one or more overhangs of CAG repeats at the terminal, is the key determinant of the association rate between RNA chains. To test the theory, we first computed the free energy spectrum for $(CAG)_{30}$ and $(CAG)_{31}$ for $0.15 \leq C_S \leq 0.5$ M using Eq. 3 (Figure 2 and S2 in SI). As is evident from Fig. 2A and Fig. S2A, ΔG_S separating the GS and the first excited state, with two CAG overhangs, increases with increasing C_S. In contrast, the GS of $(CAG)_{31}$ has one CAG overhang (Fig. 2B), which is susceptible to self-association. In principle, ΔG_S is 0 for $(CAG)_{31}$ as the GS itself contains an overhang region. However, the free energy of the GS increases as C_S increases, which is reflected in the decrease in $P(Q_{HP} = 1)$ (Fig. 1D). As a result, one should expect the dimerization rate for $(CAG)_{31}$ to decrease with an increase in C_S. We show ΔG_S and G_S as a function of C_S for $(CAG)_{30}$ and $(CAG)_{31}$, respectively, in Fig. 3A and Fig. 3B. We find that ΔG_S as a function of C_S increases, which suggests that the association of repeat RNAs through homotypic interactions should decrease with an increase in C_S.

Dimerization of hairpins as a function of C_S: To test our prediction, we performed Brownian dynamics simulations for dimer formation starting from the ground state of the hairpin structures. The simulations were performed by confining four RNA chains in the hairpin conformation inside a sphere of radius, R_0 (see SI for details). We set $R_0 = 100$ Å so that initially the monomers interact only weakly with each other. We ran 100 independent trajectories monitoring the formation of dimer for $0.15 \leq C_S \leq 0.5$ M at 27°C. We calculated the fraction of bps, f_{bp}^{12}, formed between two chains as a function of time to assess if the hairpins formed a duplex (Figure S4). If f_{bp}^{12} exceeds 0.5, the monomers are in the duplex

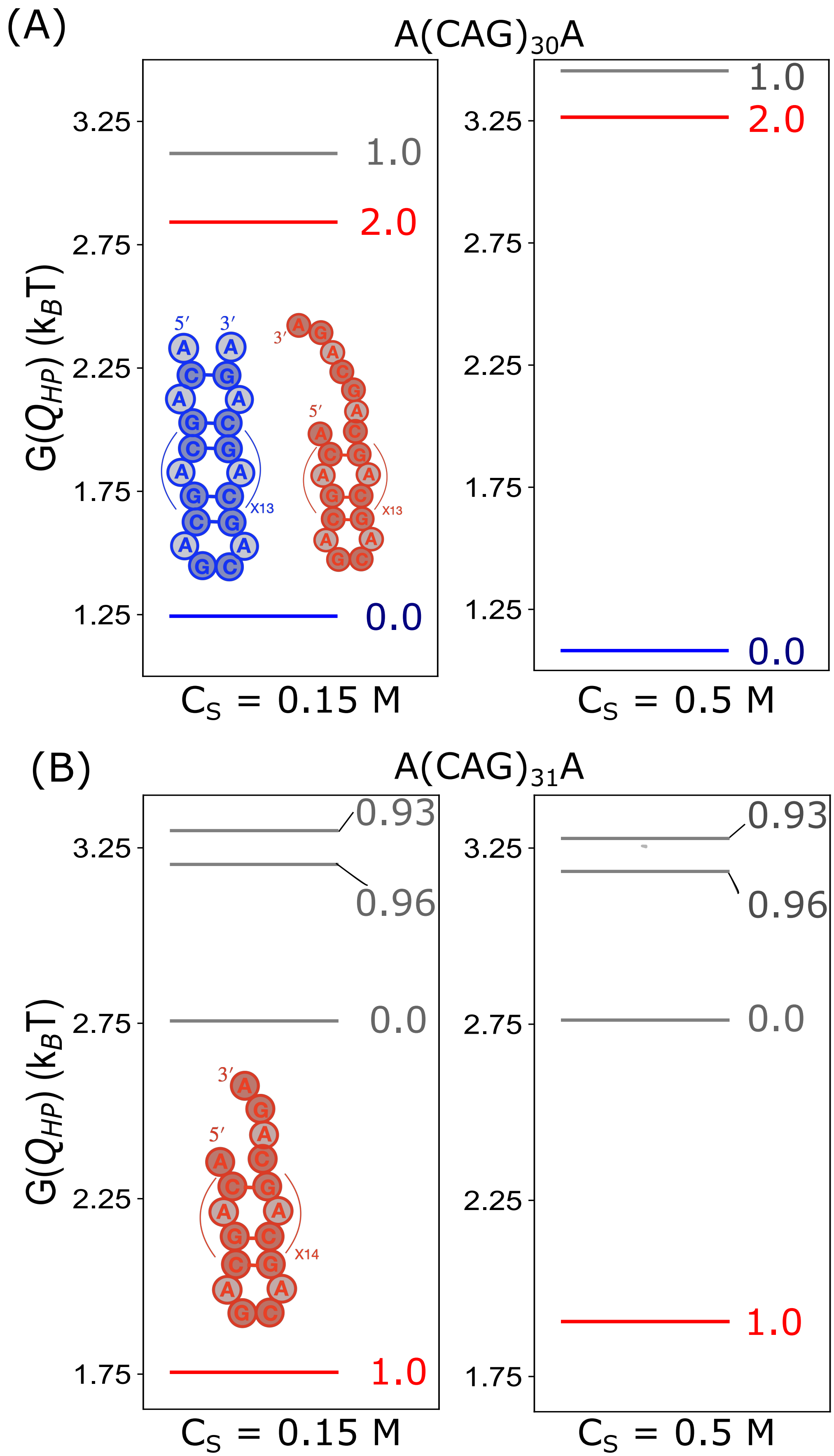

Figure 2: Free energy spectra: (A) $G(Q_{HP})$ as a function of Q_{HP} at $C_s = 0.15$ M (left) and 0.5 M (right) for (CAG)$_{30}$. The free energy gap separating the GS and the first excited state with an overhang region increases as C_S increases. A representative GS structure and the first excited state structure are in blue and red, respectively (left panel inset). (B) $G(Q_{HP})$

structure. We computed the number of trajectories leading to the dimer state, P_D, as a function of time, t (Figure S5). There are ≈ 55 (≈ 42) trajectories in which a dimer formed for $(CAG)_{31}$ ($(CAG)_{30}$) at $C_S = 0.15$ M. The number of dimer forming trajectories for $(CAG)_{31}$ ($(CAG)_{30}$) decreases to 47 (29) as the C_S is increased to 0.5 M from 0.15 M. In the the concentration range ($0.15 < C_S < 0.5$ M), the number of dimer forming trajectories do not vary significantly.

To compare the propensity to form a dimer at different C_S, we calculated the mean first passage time (MFPT), at each salt concentration (Figure 3). The first passage time (FPT), τ_{FPT}, for each dimer forming trajectory is identified with the time at which $f_{bp}^{12} = 0.5$. The MFPT is given by $\sum_{k=1}^{N_D} \tau_{FPT}^{k}/N_D$, where, τ_{FPT}^{k} is the FPT for the k^{th} trajectory and N_D is the total number of dimer forming trajectories. Fig. 3C and 3D show the MFPT as a function of C_S. Our finding that the rate of dimerization ($\propto$ MFPT^{-1}) decreases as C_S increases accords well with the theoretical prediction. The modest change in MFPT in varying C_S is related to the small change in ΔG_S as a function of C_S in the wild type sequences.

Hairpin opening is the rate determining step in RNA aggregation: Salts influence the time (τ_{conv}) required for unwinding the intra-molecular bps during the conversion of hairpins into an anti-parallel duplex structure by modulating the stability of intra-molecular bps. To investigate the effects of salts on τ_{conv}, we computed τ_{conv} for each trajectory at $C_S = 0.15$ M and 0.5 M. Consider the part of the trajectory which leads to the formation of dimer (f_{bp}^{12} $= 0.5$) without re-entering into the hairpin state. We define τ_{conv} as the time required for dimerization to be complete once association between the chains starts (Figure S7A). τ_{conv} for the dimer forming trajectories at $C_S = 0.5$ M is considerably longer compared to $C_S =$ 0.15 M, which is also reflected in the average $\langle\tau_{conv}\rangle$ values. The ratio of τ_{conv} obtained at C_S $= 0.5$ M to 0.15 M is ≈ 2.2 for $(CAG)_{30}$ and ≈ 2.3 for $(CAG)_{31}$ (Figure S7B and S7C). Our results are in accordance with the experimental observation[22,23] that the hairpin opening rate strongly depends on the ionic concentrations.

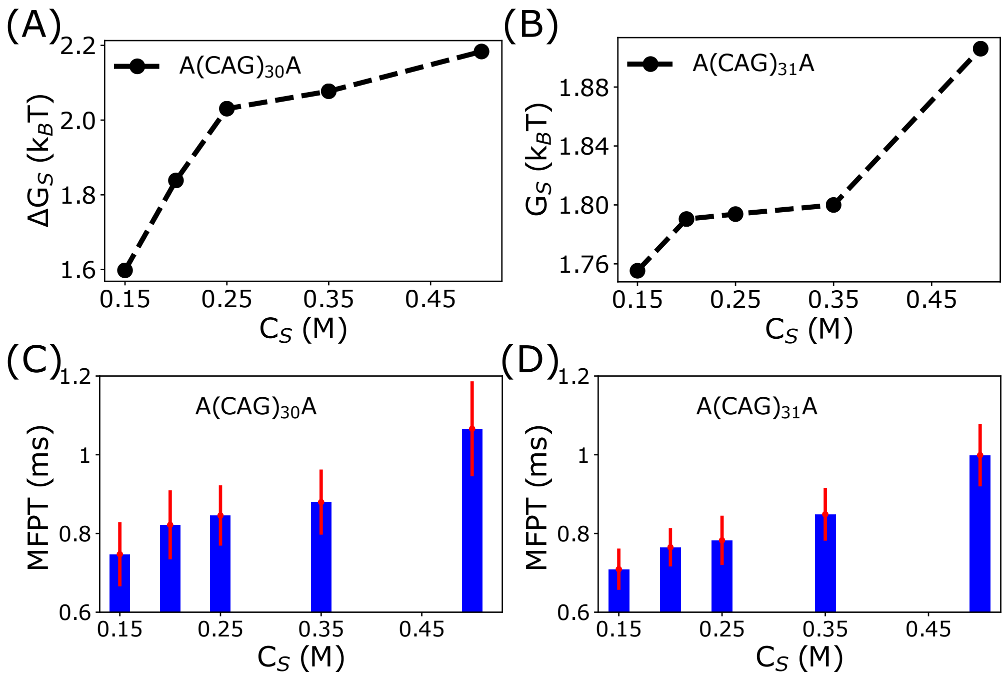

Figure 3: Link between ΔG_S and the mean first passage time, MFPT, for dimerization: (A) ΔG_S as a function of C_s for $(\mathrm{CAG})_{30}$. Increase in ΔG_S as a function of C_S implies that the relative population of aggregation-prone conformation decreases. (B) G_S as a function of C_S is for $(\mathrm{CAG})_{31}$. Increase in G_S with an increase in C_S indicates suppression in population of ground state configurations which are prone to aggregate. (C) The MFPT for $(\mathrm{CAG})_{30}$ at different C_S. The standard error in MFPT computed using the bootstrap sampling technique (see Figure S6 in SI) is in red. (D) MFPT for dimerization of $(\mathrm{CAG})_{31}$ increases with an increase in C_S, thus correlating with G_S.

Reduction in strand slippage in shuffled sequences decreases the self-association propensity: In order to further illustrate the link between phase behavior of RNA to the monomer characteristics, we calculated the free energy spectra for shuffled sequences whose sequence lengths and composition are the same as $(\mathrm{CAG})_{31}$ but the nucleotide positions are shuffled. From the large number of such sequences, we chose two sequences, labelled SS1 and SS2. The sequences are given in Figure 4. Selection of SS1 and SS2 is arbitrary. There are a large number of ways of generating the shuffled sequences. We chose sequences that populate stem-loop or hairpin-like configurations that are similar to the wild type $(\mathrm{CAG})_n$ are chosen to illustrate the theoretical prediction. The free energy spectra at $C_S = 0.15$ M and 0.5 M are given in Figure 4 and Figure S8, respectively. We note that the association of two RNA hairpins requires generating an unpaired region, most preferably near the terminal of the hairpin. An RNA hairpin that does not have an overhang region at the terminal in the ground state (GS) must access the excited state with at least one unpaired CAG repeat unit to facilitate dimer formation. Thus, it is the free energy difference between the GS and the first excited state, with an overhang, that determines the propensity for the association of the RNAs. However, if the GS has an overhang, which is the case for $(\mathrm{CAG})_{31}$, it can form oligomers using directly from the GS. The excitation-free energy difference is zero for $(\mathrm{CAG})_{31}$. In contrast to the GS of $(\mathrm{CAG})_{31}$, the GS of the shuffled sequences does not have overhangs. The value of Q_{HP} is 0 for the GS of both SS1 and SS2. The GS conformations for the sequences are shown in Figure 4B and Figure S9. The terminal nucleotides of the shuffled sequences are engaged in the formation of intramolecular base pairing in the GS conformations, and therefore cannot easily self-associate. However, the RNA chains may aggregate by accessing the excited state with overhangs. The excited states of the SS1 are devoid of such structures because of the consecutive array of GC base pair formation at the terminal. We surmise that the propensities of SS1 and SS2 to self-associate must be small. The rate of dimerization for a set of repeat sequences and their mutant variants is inversely related to the free energy gap separating the GS and the excited state containing the unpaired CAG repeat unit.[15] Based on theoretical considerations, we expect that the MFPT of the sequences would be higher compared to the wild-type sequence.

The absence of multiple peaks in the end-to-end distance distribution ($P(R_{ee})$) further supports our conclusion that the sequence SS1 contains overhang(s) neither in the GS nor in the excited states (Figure S10A). Based on the free energy spectrum, we predict that the sequence SS1 has the lowest tendency to undergo self-association. In contrast, the first excited state of the SS2 contains an overhang at the terminal of the hairpin structure (Figure 4B and S10B). The SS2 sequence is, therefore, susceptible to the formation of higher-order oligomeric structures. However, the propensity to form the oligomers is less compared to the $(CAG)_{31}$ because of enhancement in the stability of the GS for the shuffled sequence (Figure S8). The enhanced stability is also reflected in the distribution of the end-to-end distance (Figure S10B). The $P(R_{ee})$ value at $R_{ee} \approx 1.45$ nm increases significantly as C_S increases to 0.5 M from 0.15 M. The value of ΔG_S also increases as C_S increases. Because the rate of dimer formation correlates inversely with the free energy gap, we predict that the formation of higher-order oligomeric structure in SS2 should decrease substantially as the C_S is increased relative to the wild-type low complexity sequence.

Using the shuffled sequence as a reference, both experiments[1] and simulations[14] investigated the effects of nucleotide position in the sequences in the context of phase separations of CAG repeat RNAs. It was shown there is a suppression in the propensity for aggregation in the shuffled sequences, which is due to the increase in ΔG_S. Shuffling of the nucleotide positions in the sequence also changes the nature of the free energy spectrum, thus modulating ΔG_S, which is the single most important determinant of the phase behavior of RNA sequences .

Conclusions: We established a direct link between changes in the monovalent salt concentration on the free energy spectrum of low complexity RNA sequences and their propensities

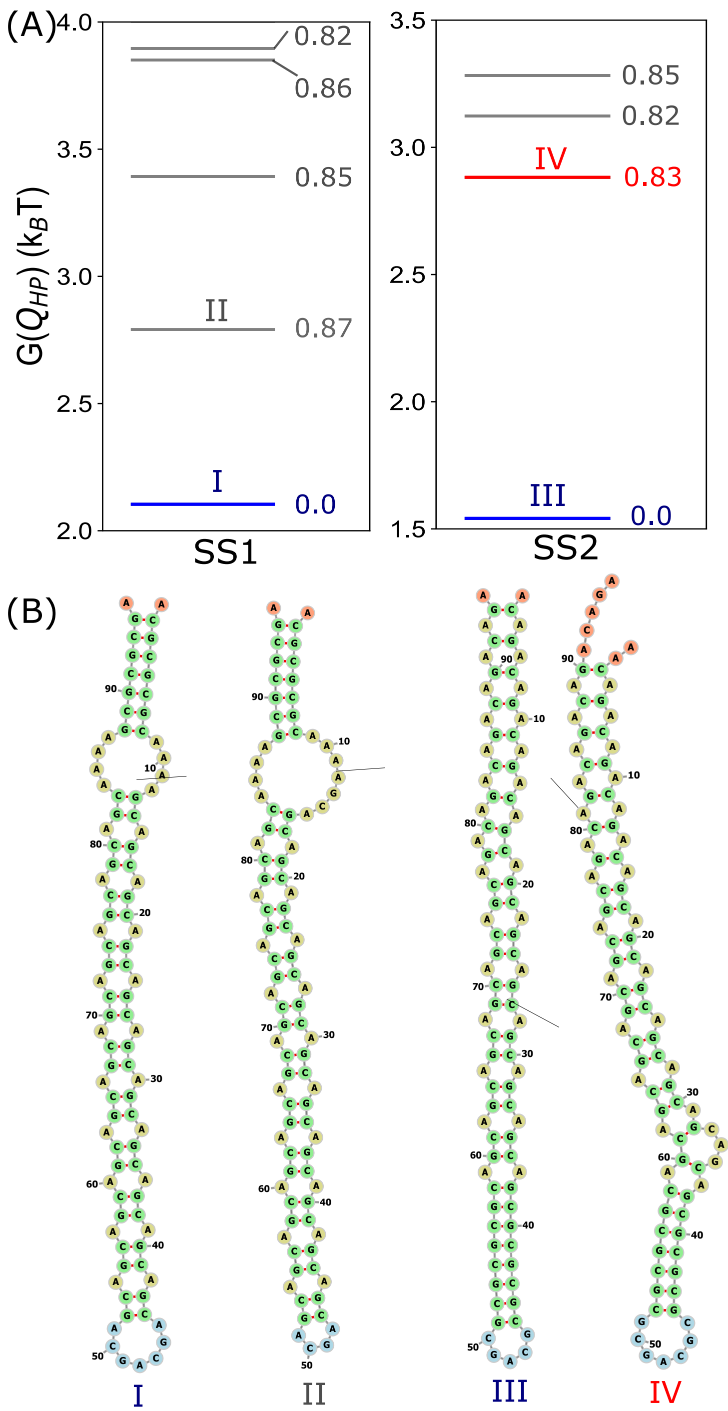

Figure 4: Free energy spectra for shuffled sequences at $C_S = 0.15$ M: (A)$G(Q_{HP})$, as a function of Q_{HP} for the sequence SS1 (left panel). The SS1 sequence is A(CG)$_3$CA$_4$G(CAG)$_{23}$CA$_4$G(CG)$_3$A. The GS and the excited states are in blue and gray lines, respectively. There is no overhang at the terminal, which is required for self-association. Right panel shows $G(Q_{HP})$, as a function of Q_{HP}, for SS2 whose sequence is (ACAG)$_4$(CAG)$_7$(CG)$_4$CAG(CG)$_4$(CAG)$_7$(ACAG)$_4$A. The GS with no overhang and the first excited state with an overhang at the end are in blue and red lines, respectively. (B) GS and the first excited state conformations of the hairpin structures for SS1 (I and II) and SS2 (III and IV).

for self-association. By combining counterion condensation theory, with Debye-Huckel potential for the electrostatic interactions, we accounted for the effects of monovalent salts in coarse-grained SIS model[24] for RNA. The major finding of our study is that the population of the hairpin state, corresponding to self-association prone conformations, is suppressed for both $(CAG)_{30}$ and $(CAG)_{31}$ as C_S is increased. The free energy gap, ΔG_S, separating an aggregation-prone hairpin state from an aggregation inactive state, increases with an increase in C_S. Strikingly, the mean first passage time (MFPT) for dimer formation increases with ΔG_S. More generally, our theory suggests that there is anti-correlation between ΔG_S and rate of RNA association. Because ΔG_S for a given sequence can be altered by changing external conditions, we predict that condensate formation may be drastically changed by tuning temperature and crowding. Our result that the dimer formation rate decreases as the salt concentration is increased can be tested experimentally.

Interestingly, the calculated values of ΔG_S for RNA repeat sequences are similar to the experimentally inferred values for the corresponding DNA sequences. For example, from the population of low and high FRET states of CAG_{14} and CAG_{15} at low salt concentration (see Fig. 2 in Ref.[20]), we find that $\Delta G_S \approx 1.45\ k_BT$. In the RNA repeat sequences value of ΔG_S is $\approx (1.3 - 2.3)\ k_BT$ at a higher value of the salt concentration. Just like the DNA trinucleotide repeats the corresponding RNA sequences are dynamic, which results in multiple states being sampled. It would be valuable to perform single molecule FRET experiments for low complexity RNA sequences in order to verify some of our findings.

An important finding in our study is that ΔG_S increases substantially for the shuffled sequences relative to the repeat sequences. Therefore, we expect that the time for dimerization for the shuffled sequences should be substantially greater for the shuffled sequences. Assuming that the MFPT $\propto exp(-\Delta G_S/k_BT)$,[25] we predict that the time for SS2 to dimerize should be roughly ten times greater. Because ΔG_S depends on the precise sequence it follows that from the astronomically large number of sequences ($= 3^n$) it is possible to construct sequences that would not form condensates at any reasonable external conditions.

Thus, salt-dependent ΔG_S may be used to control aggregation of RNA.

Our predictions are significant in the cellular context. Physiologically relevant ionic concentration varies between cells of different species. For instance, the ionic concentration of potassium ion (K^+) in budding yeast can reach up to 300 mM[26] whereas it is $\approx$ 150 mM in mammalian cells. The free energy spectrum generated for a single RNA molecule in the presence of different ionic concentrations could be used as an important tool to regulate the phase behavior of RNA-rich condensates.

Materials and Methods

Models: Following our earlier studies,[14,15,24] we represent each nucleotide by a single bead. In order to account for counter ion condensation effects,[27,28] we used a reduced value of charge -Q ($0 < Q < 1$) on the bead. In monovalent salt solutions $Q = \frac{b}{l_B(T)}$ where, b = 4.4 Å,[29] and $l_B(T) = \frac{e^2}{4\pi\epsilon(T)k_BT}$, is the Bjerrum length with e being the electron charge, k_B is the Boltzmann constant, and T is the temperature. The T dependence of the dielectric constant[30] is $\epsilon(T) = 87.74 - 0.4008T + 9.398 \times 10^{-4}\, T^2 - 1.410 \times 10^{-6}\, T^3$, and, T is expressed in °C unit. The electrostatic repulsion between the beads, accounting for the interactions between the phosphate groups, is given by,

$$E_{el} = \frac{Q^2 e^2}{4\pi\epsilon(T)} \sum_{i<j} \frac{exp(-\kappa_D r_{ij})}{r_{ij}}, \tag{1}$$

where $\kappa_D = (8\pi\rho l_B)^{1/2}$, ρ is the number density of the monovalent ions. The total energy in the Single Interaction Site (SIS) model is, $E_{TOT} = E_B + E_{HB} + E_{EV} + E_{el}$. The detailed functional forms of the bonded (E_B), hydrogen bond (E_{HB}), and the excluded volume interactions (E_{EV}) along with the parameter values are given elsewhere.[15] Although the SIS model is simple, our previous study[15] showed that the agreement between the calculated and experimentally measured heat capacities for several $(CAG)_n$ constructs is very good.

Free energy spectra: We first calculated the distribution, $P(Q_{HP})$, of the hairpin order parameter, Q_{HP}, which measures the deviation of WC base pairs (GC base pairs in our case) with respect to a perfectly aligned hairpin structure. The Q_{HP} order parameter is defined as,

$$Q_{HP} = \Big[\frac{1}{N_{bp}} \sum_{i,j}^{N_{bp}} \Big(\frac{i + j - N_T - 1}{3}\Big)^2\Big]^{1/2}, \tag{2}$$

where i and j are the nucleotides, N_T is the sequence length, and N_{bp} is the number of base pairs in a given conformation. For a perfect hairpin, it is easy to show that $i + j = N_T + 1$, implying that $Q_{HP} = 0$. The value of $Q_{HP} = 1$ corresponds to a slipped hairpin, corresponding to one unit of unpaired CAG. Fractional values, $0 < Q_{HP} < 1$, represent a variety of conformations containing bulges in the stem. The free energies are calculated by arranging $P(Q_{HP})$ values in descending order. The spectrum is computed using,

$$G(Q_{HP}) = -k_B T \ \ln P(Q_{HP}). \tag{3}$$

Simulations: The thermodynamic properties, including the free energy spectra, are calculated using the trajectories generated by integrating the Langevin equation of motion in the low friction limit.[31] In order to investigate the formation of aggregates in $(CAG)_{30}$ and $(CAG)_{31}$, we performed Brownian dynamics simulations using the Ermack-McCammon algorithm.[32] The details are given in the Supporting Information.

Supporting Information

Simulation details, model and methods, parameters for energy function, probability distribution of Q_{HP}, free energy spectrum, schematics of different hairpin structures, representative dimer forming trajectories, dimer yield as a function of time, distribution of MFPT, time required for completion of a dimer, τ_{conv}, free energy spectrum of shuffled sequences, schematics of shuffled sequences, end-to-end distribution ($P(R_{ee})$) of shuffled sequences.

Acknowledgements: This work was supported by a grant from the National Science Foundation (CHE 2320256) and the Welch Foundation through the Collie-Welch Chair (F-0019).

References

(1) Jain, A.; Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* **2017**, *546*, 243–247.

(2) Hautke, A.; Voronin, A.; Idiris, F.; Riel, A.; Lindner, F.; Lelièvre-Büttner, A.; Zhu, J.; Appel, B.; Fatti, E.; Weis, K; Muller, S.; Schug, A.; Ebbinghaus, S. CAG-repeat RNA hairpin folding and recruitment to nuclear speckles with a pivotal role of ATP as a cosolute. *J. Am. Chem. Soc.* **2023**, *145*, 9571–9583.

(3) Das, M. R.; Chang, Y.; Anderson, R.; Saunders, R. A.; Zhang, N.; Tomberlin, C. P.; Vale, R. D.; Jain, A. Repeat-associated non-AUG translation induces cytoplasmic aggregation of CAG repeat-containing RNAs. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120*, e2215071120.

(4) Ma, Y.; Li, H.; Gong, Z.; Yang, S.; Wang, P.; Tang, C. Nucleobase clustering contributes to the formation and hollowing of repeat-expansion RNA condensate. *J. Am. Chem. Soc.* **2022**, *144*, 4716–4720.

(5) Isiktas, A. U.; Eshov, A.; Yang, S.; Guo, J. U. Systematic generation and imaging of tandem repeats reveal multivalent base-pairing as a major determinant of RNA aggregation. *Cell Rep. Methods* **2022**, *2*, 100334.

(6) Guo, H.; Ryan, J. C.; Song, X.; Mallet, A.; Zhang, M.; Pabst, V.; Decrulle, A. L.; Ejsmont, P.; Wintermute, E. H.; Lindner, A. B. Spatial engineering of E. coli with addressable phase-separated RNAs. *Cell* **2022**, *185*, 3823–3837.e23.

(7) Hauf, S.; Yokobayashi, Y. Chemical control of phase separation in DNA solutions. *Nucleic Acids Res.* **2023**, *59*, 3751–3754.

(8) Wadsworth, G. M.; Zahurancik, W. J.; Zeng, X.; Pullara, P.; Lai, L. B.; Sidharthan, V.;

Pappu, R. V.; Gopalan, V.; Banerjee, P. R. RNAs undergo phase transitions with lower critical solution temperatures. *Nat. Chem.* **2023**, *15*,1–12.

(9) Flory, P. J. Thermodynamics of high polymer solutions. *J. Chem. Phys.* **1942**, *10*, 51–61.

(10) Huggins, M. L. Some properties of solutions of long-chain compounds. *J. Chem. Phys.* **1942**, *46*, 151–158.

(11) Everett, C.; Wood, N. Trinucleotide repeats and neurodegenerative disease. *Brain* **2004**, *127*, 2385–2405.

(12) McColgan, P.; Tabrizi, S.J. Huntington's disease: a clinical review. *Eur. J. Neurol* **2018**, *25*, 24–34.

(13) Depienne, C.; Mandel, J.-L. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **2021**, *108*, 764–785.

(14) Nguyen, H. T.; Hori, N.; Thirumalai, D. Condensates in RNA repeat sequences are heterogeneously organized and exhibit reptation dynamics. *Nat. Chem.* **2022**, *14*, 775–785.

(15) Maity, H.; Nguyen, H. T.; Hori, N.; Thirumalai, D. Odd–even disparity in the population of slipped hairpins in RNA repeat sequences with implications for phase separation. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120*, e2301409120.

(16) Kimchi, O.; King, E. M.; Brenner, M. P. Uncovering the mechanism for aggregation in repeat expanded RNA reveals a reentrant transition. *Nat. Commun.* **2023**, *14*, 332.

(17) Ramachandran, V.; Potoyan, D. A. Mapping energy landscapes of homopolymeric RNAs via simulated tempering and deep unsupervised learning. *bioRxiv* **2023**,

(18) Kerpedjiev, P.; Hammer, S.; Hofacker, I. L. Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. *Bioinformatics.* **2015**, *31*, 3377–3379.

(19) Ni, C.-W.; Wei, Y.-J.; Shen, Y.-I.; Lee, I.-R. Long-range hairpin slippage reconfiguration dynamics in trinucleotide repeat sequences. *J. Phys. Chem. Lett.* **2019**, *10*, 3985–3990.

(20) Xu, P.; Pan, F.; Roland, C.; Sagui, C.; Weninger, K. Dynamics of strand slippage in DNA hairpins formed by CAG repeats: roles of sequence parity and trinucleotide interrupts. *Nucleic Acids Res.* **2020**, *48*, 2232–2245.

(21) Sobczak, K.; de Mezer, M.; Michlewski, G.; Krol, J.; Krzyzosiak, WJ. RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Res.* **2003**, *31*, 5469–5482.

(22) Mitchell, M. L.; Leveille, M. P.; Solecki, R. S.; Tran, T.; Cannon, B. Sequence-dependent effects of monovalent cations on the structural dynamics of trinucleotide-repeat DNA hairpins. *J. Phys. Chem. B.* **2018**, *122*, 11841–11851.

(23) Tsukanov, R.; Tomov, T. E.; Berger, Y.; Liber, M.; Nir, E. Conformational dynamics of DNA hairpins at millisecond resolution obtained from analysis of single-molecule FRET histograms. *J. Phys. Chem. B.* **2013**, *117*, 16105–16109.

(24) Hyeon, C.; Dima, R. I.; Thirumalai, D. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. *Structure* **2006**, *14*, 1633–1645.

(25) Li, M. S.; Klimov, D. K.; Thirumalai, D. Finite size effects on thermal denaturation of globular proteins. *Phys. Rev. Lett.* **2004**, *93*, 268107.

(26) Volkov, V. Quantitative description of ion transport via plasma membrane of yeast and small cells. *Front Plant Sci.* **2015**, *6*, 425.

(27) Oosawa, F. Polyelectrolytes, Marcel Dekker, New York, 1971.

(28) Manning, G. S. Limiting laws and counterion condensation in polyelectrolyte solutions .I. colligative properties. *J. Chem. Phys.* **1969**, *51*, 924–933.

(29) Denesyuk, N. A.; Thirumalai, D. Coarse-grained model for predicting RNA folding thermodynamics. *J. Phys. Chem. B* **2013**, *117*, 4901–4911.

(30) Malmberg, C.; Maryott, A. Dielectric constant of water from 0° to 100° C. *J. Res. Natl. Bureau Standards* **1956**, *56*, 1–8.

(31) Honeycutt, J. D.; Thirumalai, D. The nature of folded states of globular-proteins. *Biopolymers* **1992**, *32*, 695–709.

(32) Ermak, D.; McCammon, J. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.* **1978**, *69*, 1352–1360.