Published online 26 July 2021

Nucleic Acids Research, 2021, Vol. 49, No. 14 ***7901–7908***
https://doi.org/10.1093/nar/gkab616

Predicting stability of DNA bulge at mononucleotide microsatellite

Jin H. Bae [1] and David Yu Zhang [1,2,*]

[1]Department of Bioengineering, Rice University, Houston, TX 77005, USA and [2]Systems, Synthetic, and Physical Biology, Rice University, Houston, TX 77005, USA

Received February 02, 2021; Revised June 28, 2021; Editorial Decision June 30, 2021; Accepted July 07, 2021

ABSTRACT

Mononucleotide microsatellites are clinically and forensically crucial DNA sequences due to their high mutability and abundance in the human genome. As a mutagenic intermediate of an indel in a microsatellite and a consequence of probe hybridization after such mutagenesis, a bulge with structural degeneracy sliding within a microsatellite is formed. Stability of such dynamic bulges, however, is still poorly understood despite their critical role in cancer genomics and neurological disease studies. In this paper, we have built a model that predicts the thermodynamics of a sliding bulge at a microsatellite. We first identified 40 common bulge states that can be assembled into any sliding bulges, and then characterized them with toehold exchange energy measurement and the partition function. Our model, which is the first to predict the free energy of sliding bulges with more than three repeats, can infer the stability penalty of a sliding bulge of any sequence and length with a median prediction error of 0.22 kcal/mol. Patterns from the prediction clearly explain landscapes of microsatellites observed in the literature, such as higher mutation rates of longer microsatellites and C/G microsatellites.

INTRODUCTION

Microsatellites, which are also known as short tandem repeats of DNA, are clinically and forensically important. Due to their high mutability, microsatellites have been widely used as predictive (1–4), diagnostic (5–8), prognostic (9–12) and forensic biomarkers (13–16). They also act as markers in population studies (17–20) and have diverse functional roles (21). When genetic hypermutability is caused by an impaired DNA mismatch repair system, a condition called microsatellite instability (MSI) arises, where the number of tandem repeats grows or shrinks (22). This emphasizes the role of MSI as a molecular phenotype of tumor, but it also promotes oncogenesis (23).

In order for natural mutagenesis at a microsatellite to initiate, a strand slippage error occurs first during replication (24) to result in a special bulge with structural degeneracy (Figure 1A). Moreover, such bulges can be formed again when a hybridization probe or a PCR primer binds to a microsatellite with such mutation. We named this laterally sliding DNA motif a sliding bulge to distinguish from other static bulges. Characterization of sliding bulges is necessary to design effective probes and primers, understand how MSI occurs and, consequently, better exploit microsatellites as biomarkers. Because microsatellites in human genome are primarily mononucleotides (25) that go up to 83 repeats according to the reference genome assembly (GRCh38/hg38), the scope of this study is set to sliding bulges at mononucleotide microsatellites.

Despite the importance of sliding bulges, thermodynamics behind their stability is still poorly understood. One reason why there has been no systematic study is perhaps that various lengths of microsatellites give rise to a huge number of possible sliding bulges to be characterized. According to the nearest-neighbor (NN) model (26), which is the currently accepted DNA thermodynamic approximation, even the number of sliding bulges shorter than 30 tandem repeats is already over 1000. Another reason may be a lack of an accurate method for distinguishing small energy differences among many sliding bulges. Errors of DNA melting analysis were too significant (27) to confidently analyze the delicate thermodynamics of sliding bulges owing to structural degeneracy.

Here, we describe how we constructed and validated a predictive model of the thermodynamics of sliding bulges with any length and sequence. Toehold exchange energy measurement (TEEM) (28), which utilizes toehold exchange reactions in parallel to infer $\Delta\Delta G^{\circ}$ over a range of temperature independently (Supplementary Section S1), was used to accurately measure the free energy of bulges. We first resolved sequence-specific effects to systematically study destabilization of sliding bulges, and then observed how structural degeneracy affects destabilization. Based on the partition function (29), we characterized 40 common bulge states that can be assembled to any sliding bulge. We

*To whom correspondence should be addressed. Tel: +1 626 390 2242; Email: dyz1@rice.edu

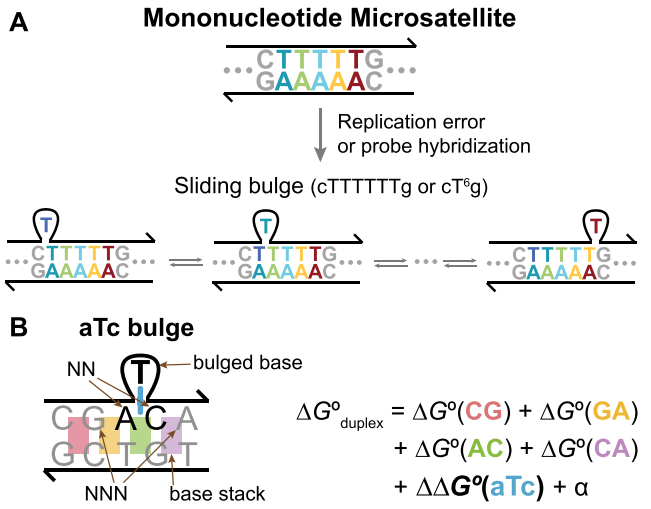

Figure 1. **(A)** A mononucleotide microsatellite is hypermutable short tandem repeats of a single base. When an extra T is added in this example, a bulged base is formed that can replace a neighboring T, causing a domino effect of sliding. We named this laterally sliding motif a sliding bulge. **(B)** Bulge thermodynamics based on the nearest-neighbor (NN) model. The name aTc bulge indicates that the bulged base is T, and its NNs are A and C. The bases next to the NN are called the next-nearest neighbors (NNNs). $\Delta G^\circ_{\text{duplex}}$ is calculated as a sum of ΔG° of its base stacks (colored rectangles), a destabilizing $\Delta\Delta G^\circ$ of a bulge (blue rod) and a few constant terms.

successfully validated our model by comparing it to experimental results and the literature.

MATERIALS AND METHODS

Materials

Phosphate-buffered saline (PBS, pH 7.4), Tris–EDTA (TE) and Tween 20 were purchased from Sigma-Aldrich. All oligonucleotides (oligos) were synthesized at the 100 nmol scale, dissolved in TE buffer (pH 8.0) to 100 μM and HPLC-purified by Integrated DNA Technologies (IDT). Chemical modifications on oligos were prepared by IDT as well. The concentrations of oligo stocks were verified with Nanodrop (Thermo Fisher), and then diluted to 10 μM in PBS. All oligos were stored in darkness at 4°C. The sequences of all oligos are listed in Supplementary Table S1. Solution fluorescence for TEEM was measured using a QuantStudio 7 Flex instrument (Applied Biosystems). Samples were loaded in MicroAmp Fast Optical 96-Well Reaction Plates, 0.1 ml (Applied Biosystems), and the loaded plate was sealed using MicroAmp Optical Adhesive Film (Applied Biosystems).

TEEM

TEEM is described in detail and validated thoroughly in our previous publication (28). Briefly, it utilizes a toehold exchange reaction of C, P and X oligos, where both P and X oligos can hybridize with C oligo (Supplementary Section S1). Because P and X oligos are shorter than C oligo and aligned to its opposite ends, they can displace each other from C oligo. As a result, CP and CX duplexes exist in equilibrium according to their ΔG°, and we can calculate the free energy by measuring the concentrations of the duplexes. In order to infer the concentrations, we designed only CX duplex to emit fluorescence by functionalizing C and P oligos with a ROX fluorophore and an Iowa Black RQ quencher, respectively. Inferring $\Delta\Delta G^\circ$ of a motif requires toehold exchange reactions with X(reference) and its variation, X(bulge), oligos. The only difference between X(reference) and X(bulge) oligos is the presence of a bulge motif of interest, and subtracting their ΔG° of the reaction cancels everything out except $\Delta\Delta G^\circ$ of a bulge.

TEEM begins with diluting C, P, X(reference) and X(bulge) oligos from stock solutions to 0.5, 0.5, 0.4 and 10 μM, respectively, in PBS with 0.1% (v/v) Tween 20. C and P oligos were mixed first and then X(reference) oligo was added afterward to make their working concentrations 20, 30 and 40 nM, respectively. For reactions with X(bulge) oligos, their working concentration was 1 μM. Positive and negative control samples for characterizing maximum and minimum fluorescence signals were prepared by adding PBS in place of the P or X oligos, respectively.

To verify whether fluorescence measurements reflect actual equilibrium conditions, we measured fluorescence at every integer temperature between 20 and 70°C twice: once as the solution is being gradually cooled, and another time when the solution is being gradually warmed. Between the cooling and heating phases, temperature was maintained at 20°C for 1 h to check again whether equilibration is complete. The heating phase was the reverse of the cooling phase, and all temperature change was done at the rate of 2°C/s. Consistency between the two measurements implied that equilibrium was established.

RESULTS

Effect of the NNNs

The NN model approximates the free energy of hybridization $\Delta G^\circ_{\text{duplex}}$ as a sum of their $\Delta G^\circ_{\text{base stack}}$ and destabilization $\Delta\Delta G^\circ$ of a bulge as shown in Figure 1B. This bulge is called aTc bulge because AC base stack was disturbed by a bulged T, and now A and C are the NNs of the bulge. Using their relationship, bulge destabilization energy, or thermodynamic penalty, can be calculated as a ΔG° difference between a bulged duplex and its corresponding canonical duplex, hence $\Delta\Delta G^\circ$.

To systematically measure $\Delta\Delta G^\circ$ of bulges, we first supplemented the NN model by reducing sequence-specific effects of local context. Although the thermodynamics of a perfectly matched duplex can be approximated well by considering only NNs, the same cannot be assumed for non-canonical structures like bulges. A bulge can disrupt a local double-helix structure (30), and this destabilization may reach bases beyond NNs. We thus defined NNNs as bases adjacent to NNs of a bulged base, and checked how NNNs affect bulge $\Delta\Delta G^\circ$ by measuring $\Delta\Delta G^\circ$ of four bulges (cAc, gTa, aTc and aCt) 12 times, each with different NNNs (Figure 2A). The measurement was performed at 43 different temperatures from 25 to 67°C, resulting in 2064 $\Delta\Delta G^\circ$ values in total. As expected, different NNNs showed different effects on bulge $\Delta\Delta G^\circ$.

To keep consistency in $\Delta\Delta G^\circ$ measurement, we decided to select the representative NNNs and use them through-

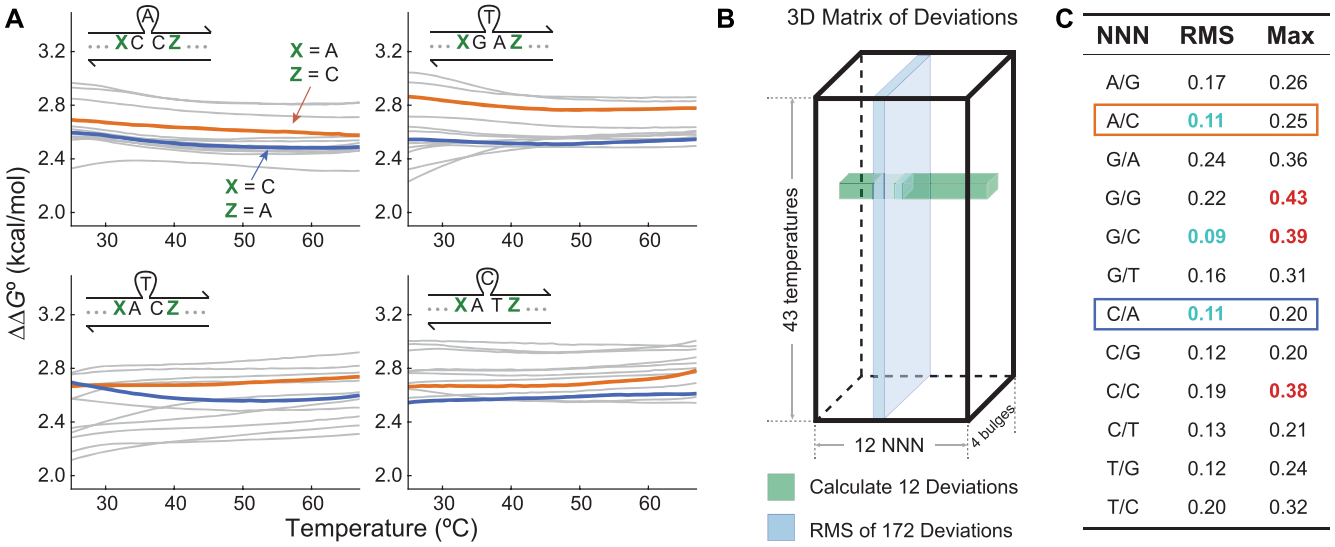

Figure 2. Process of selecting the representative NNNs for consistency in $\Delta\Delta G^\circ$ measurement. **(A)** $\Delta\Delta G^\circ$ of four non-slide bulges, each with 12 different NNNs (denoted as X and Z) at 43 different temperatures, were measured by TEEM (2064 $\Delta\Delta G^\circ$ values). After (B) and (C), A/C (orange) and C/A (dark blue) were to be selected as the representative NNNs for their best representativeness. **(B)** To find the NNNs that represent the others best, we compared $\Delta\Delta G^\circ$ values of each bulge with 12 different NNNs. A deviation of each NNN is defined as a difference between each $\Delta\Delta G^\circ$ and a mean of 12 $\Delta\Delta G^\circ$ values (green row). To evaluate overall representativeness of each NNN, a root mean square (RMS) of 172 deviations (blue plane) was calculated. A lower RMS means better representativeness. **(C)** As shown in (A), we selected two NNNs, A/C and C/A, with low RMS (teal) and without high maximum deviations (red).

out this work. With a given NN and a bulged base, the representative NNNs should result in bulge $\Delta\Delta G^\circ$ close to a mean $\Delta\Delta G^\circ$ of all 12 NNNs. We introduce the concepts of a NNN deviation and an RMS of deviations to find NNNs with the highest representativeness across all temperatures and four bulges. An NNN deviation is defined as a difference between a $\Delta\Delta G^\circ$ value and a mean of 12 $\Delta\Delta G^\circ$ values from different NNNs. Because we tested four bulges at 43 temperatures, each NNN had 172 deviation values in total (Figure 2B), and RMS of each NNN evaluated the representativeness. Naturally, a lower RMS means better representativeness across different temperatures and bulge motifs. Because NNNs with low RMS could still have several huge deviations that damage the systematic approach, we also used maximum deviation as a secondary criterion to avoid any extreme cases. By picking three NNNs with the lowest RMS and avoiding three NNNs with the highest maximum deviations, A/C (A and C next to 5′ and 3′ ends of the bulge NN, respectively) and C/A were selected as the representative NNNs to measure $\Delta\Delta G^\circ$ twice and take an average for further experiments (Figure 2C).

Bulge $\Delta\Delta G^\circ$ measurement

Using the selected representative NNNs, we investigated how structural degeneracy of a sliding bulge affects the thermodynamics. As the reference, $\Delta\Delta G^\circ$ of all 36 bulges with a single dominant state and no sliding (non-slide bulge) were first measured at 43 different temperatures from 25 to 67°C. There are 36 bulges because a bulged base can be any of four bases, while the nearest non-bulged bases have to be different from the bulged base, leaving them three choices each. Figure 3A summarizes the results with each arrow starting at $\Delta\Delta G^\circ_{25^\circ C}$ and ending at $\Delta\Delta G^\circ_{65^\circ C}$ of a bulge, and black dots on the arrows indicate $\Delta\Delta G^\circ_{45^\circ C}$. Complete datasets with all $\Delta\Delta G^\circ$ values are provided in Supplementary Section S2. We grouped the bulges by their bulged bases to highlight differences in slopes among them, and there was no other grouping method that revealed other differences (Supplementary Table S2). It is notable that $\Delta\Delta G^\circ$ of bulges with purine bases (A and G) are generally affected more by temperature, shown by the longer arrows. Moreover, their directions are always downward, implying that they have lower thermodynamic penalties at higher temperature.

As a comparison, we then measured $\Delta\Delta G^\circ$ of two-slide bulges that have sliding between two degenerate states (Figure 3B). There are 36 two-slide bulges, which are formed when a bulged base and one of the neighboring bases are the same, allowing them to replace each other. In addition to having the same trends observed in non-slide bulges, two-slide bulges with a purine base showed slightly lower $\Delta\Delta G^\circ$ than pyrimidine bulges (C and T) in this dataset. Both non- and two-slide bulges generally showed lower $\Delta\Delta G^\circ$ than their RNA counterparts (31,32), confirming that DNA bulges were less destabilizing. To observe the effect of structural degeneracy, $\Delta\Delta G^\circ$ of two-slide bulges were plotted against those of their corresponding non-slide bulges (e.g. aTc versus aTTc) at 37°C (Figure 3C). Most of 36 bulge pairs were below the diagonal line with five exceptions (brown), implying that the degeneracy generally decreased $\Delta\Delta G^\circ$ and stabilized two-slide bulges.

Interestingly, all five exceptions have a pyrimidine base as a bulged base, whereas the motifs on the opposite side (teal) have a purine base. This polarized result between pyrimidine and purine bulges can be explained when $\Delta\Delta G^\circ_{37^\circ C}$ values of each motif are grouped by the ring types (Figure 3D).

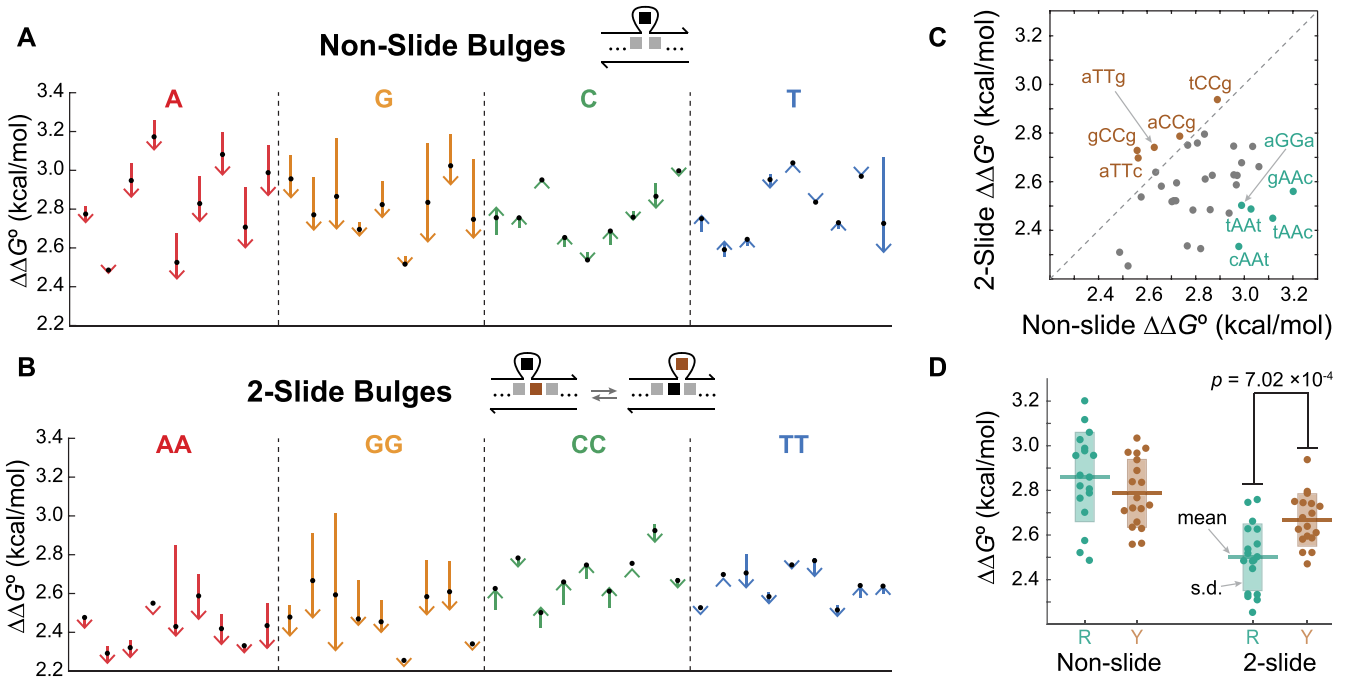

Figure 3. **(A)** $\Delta\Delta G^\circ$ of all 36 non-slide bulges at 43 different temperatures were measured as the reference and summarized by arrows. Each arrow starts at $\Delta\Delta G^\circ_{25^\circ C}$ and ends at $\Delta\Delta G^\circ_{65^\circ C}$ with each black dot indicating $\Delta\Delta G^\circ_{45^\circ C}$. Bulges with the same bulged base are organized in alphabetical order. **(B)** Summarized $\Delta\Delta G^\circ$ of all two-slide bulges. **(C)** $\Delta\Delta G^\circ$ of two-slide bulges plotted against $\Delta\Delta G^\circ$ of the corresponding non-slide bulges at 37°C. $\Delta\Delta G^\circ$ of a two-slide bulge is generally lower than that of a non-slide bulge. The exceptions are colored in brown, which all have the pyrimidine bases (C and T) as bulges, whereas the examples in the opposite end (teal) have the purine bases (A and G). **(D)** $\Delta\Delta G^\circ_{37^\circ C}$ grouped by ring types of a bulged base. R and Y denote purine and pyrimidine, respectively. Only two-slide bulges show a statistically significant difference (Welch's *t*-test) between purine and pyrimidine, which explains the polarized result in (C).

While $\Delta\Delta G^\circ$ of both purine and pyrimidine bulges were decreased by structural degeneracy, the change was more significant for purine bulges. Welch's unequal variance *t*-test results suggest that non-slide bulges do not show any statistical difference between purine and pyrimidine bulges, but two-slide bulges do.

Sliding bulge model construction

To build a predictive model of sliding bulge stability from the measured data, we laid groundwork of the model construction with the NN model of DNA thermodynamics. $\Delta\Delta G^\circ$, which is a measure of destabilization, shows the ratio of the new equilibrium constant K_2 to the old K_1:

$$\Delta\Delta G^\circ = \Delta G^\circ_2 - G^\circ_1 = -RT\ln K_2 - (-RT\ln K_1) = -RT\ln\left(\frac{K_2}{K_1}\right).$$

The partition function, which derives the thermodynamic properties of the equilibrium conformational ensemble (29), can be readily applied to predicting overall destabilization from degenerate states of a sliding bulge. Our model infers $\Delta\Delta G^\circ$ of a sliding bulge by combining $\Delta\Delta G^\circ$ of each degenerate bulge state into a partition function. Figure 4 shows examples of the energy calculation and how $\Delta\Delta G^\circ(\mathrm{cT}^N\mathrm{c})$ can be calculated from $\Delta\Delta G^\circ$(cTTc) and $\Delta\Delta G^\circ$(cTTTc). First, $\Delta\Delta G^\circ$ and the partition function Z of cTTc bulge are expressed with $\Delta\Delta G^\circ$ of its two states (Figure 4A):

$$\Delta\Delta G^\circ(\mathrm{cTTc}) = -RT\ln(Z) = -RT\ln\left(\mathrm{e}^{-\Delta\Delta G^\circ(\mathrm{cTt})/RT} + \mathrm{e}^{-\Delta\Delta G^\circ(\mathrm{tTc})/RT}\right).$$

$\Delta\Delta G^\circ$(cTTTc) can be expressed in a similar way (Figure 4B):

$$\Delta\Delta G^\circ(\mathrm{cTTTc}) = -RT\ln\left(\mathrm{e}^{-\Delta\Delta G^\circ(\mathrm{cTt})/RT} + \mathrm{e}^{-\Delta\Delta G^\circ(\mathrm{tTc})/RT} + \mathrm{e}^{-\Delta\Delta G^\circ(\mathrm{tTt})/RT}\right).$$

The only difference between $\Delta\Delta G^\circ$(cTTc) and $\Delta\Delta G^\circ$(cTTTc) is $\Delta\Delta G^\circ$(tTt) term (green), which we named T triplet state, so it can be numerically separated from the equations (Supplementary Section S3).

T triplet state plays the key role in expanding the prediction to a longer sliding bulge $\mathrm{cT}^N\mathrm{c}$. Using the separated $\Delta\Delta G^\circ$(tTt), the following equation can be used to predict $\Delta\Delta G^\circ(\mathrm{cT}^N\mathrm{c})$ with any N value (Figure 4C):

$$\Delta\Delta G^\circ(\mathrm{cT}^N\mathrm{c}) = -RT\ln\left(\mathrm{e}^{-\Delta\Delta G^\circ(\mathrm{cTTc})/RT} + (N-2)\,\mathrm{e}^{-\Delta\Delta G^\circ(\mathrm{tTt})/RT}\right),$$

If we apply the same reasoning to other sequences, $\Delta\Delta G^\circ$ of 40 common bulge states, which are 36 two-slide bulges and 4 triplet states, work as common building blocks that can be assembled into $\Delta\Delta G^\circ$ of any sliding bulge.

With $\Delta\Delta G^\circ$ data of all two-slide bulges acquired, the only missing building blocks for the predictive model were $\Delta\Delta G^\circ$ of the triplet states. Because $\Delta\Delta G^\circ$ of each triplet

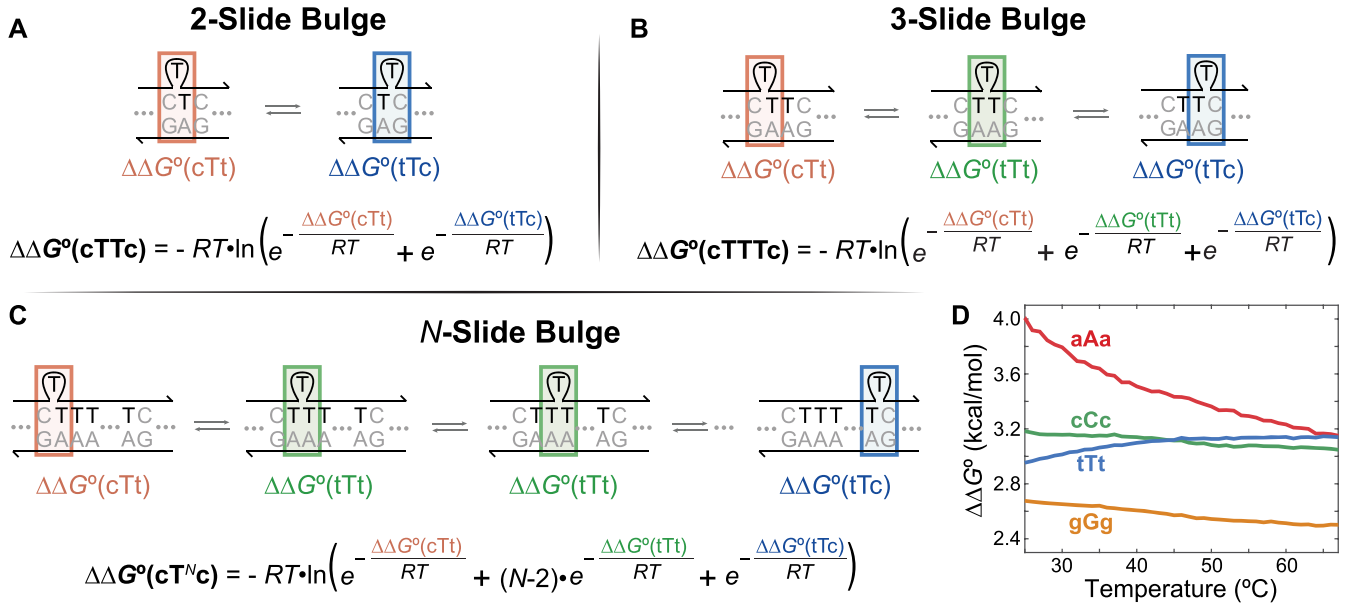

Figure 4. An example of a thermodynamic model of sliding bulges based on the partition function. **(A)** $\Delta\Delta G^\circ$ of a two-slide bulge cTTc can be expressed by $\Delta\Delta G^\circ$ of two degenerate states (orange and blue boxes). **(B)** Compared to cTTc bulge, a three-slide bulge cTTTc has one more state (green box) that we named T triplet state. **(C)** Likewise, an N-slide bulge is formed when a bulged base and $N-1$ of neighboring bases are identical. Because the only difference between cT^{N}c and cTTc bulges is the number of T triplet states, inferring $\Delta\Delta G^\circ$(tTt) enables the model to predict $\Delta\Delta G^\circ$(cT^{N}c) with any N value. T triplet state can be isolated by comparing $\Delta\Delta G^\circ$(cTTc) and $\Delta\Delta G^\circ$(cTTTc). **(D)** Inferred mean $\Delta\Delta G^\circ$ of each triplet state. Measured $\Delta\Delta G^\circ$ of three three-slide bulges used for the calculation are shown in Supplementary Section S3.

state can be extracted from $\Delta\Delta G^\circ$ of a two-slide bulge and its corresponding three-slide bulge (Supplementary Section S3), we measured $\Delta\Delta G^\circ$ of three three-slide bulges each. They were lower than $\Delta\Delta G^\circ$ of two-slide bulges in general, and $\Delta\Delta G^\circ$ of the G bulges were especially lower than the others (Supplementary Section S2). Figure 4D shows $\Delta\Delta G^\circ$ of the triplet states calculated from the mean $e^{-\Delta\Delta G^\circ(\text{triplet})/RT}$ of three three-slide bulges. $\Delta\Delta G^\circ$(aAa) is much higher than $\Delta\Delta G^\circ$ of any two-slide bulge at lower temperature, suggesting it has a minor role in equilibrium conformational ensemble. However, as temperature goes up, $\Delta\Delta G^\circ$(aAa) goes down, which makes $\Delta\Delta G^\circ$ of sliding A bulges more temperature dependent. In contrast, low $\Delta\Delta G^\circ$(gGg) makes the partition function Z larger and overall $\Delta\Delta G^\circ$ smaller, especially when multiple G triplet states stack in longer bulges. This can be interpreted as more stable G triplet state contributing significantly to stabilizing sliding G bulges.

Predicting thermodynamics of sliding bulges

Using the predictive model based on the partition function and the $\Delta\Delta G^\circ$ data, we tested our prediction power by comparing predicted and measured $\Delta\Delta G^\circ$ values of four longer sliding bulges (Figure 5A). The lengths of C and G bulges were shorter than those of A and T bulges because sequences with long consecutive C or G had shown unreliable results under TEEM in our previous experience (28), which we attribute to secondary structures and limitations of oligo synthesis. A median and a maximum absolute value of residuals, which are differences between predicted and measured $\Delta\Delta G^\circ$, were 0.22 and 0.37 kcal/mol, respectively. It is notable that temperature did not affect the residuals significantly because of the similar slopes between the predicted and the measured values.

As a quick reference, we plotted mean $\Delta\Delta G^\circ$ of all bulges with a given length and a bulged base (Figure 5B). A number next to each plot indicates the length of homopolymeric repeats. Sliding A bulges have steep $\Delta\Delta G^\circ$ slopes at lower temperature due to steep $\Delta\Delta G^\circ$(aAa) slopes as displayed in Figure 4D, whereas $\Delta\Delta G^\circ$ of sliding G bulges have lower values because the low $\Delta\Delta G^\circ$(gGg) contributes more to their stability. Of note, DNA with a long stretch of C or G may form secondary structures such as i-motif or G-quadruplex, resulting in deviations from the model. To make the prediction publicly available, we have created and attached a MATLAB function that only requires the sequence of a sliding bulge and temperature.

DISCUSSION

In this work, we have built a model of sliding bulges at mononucleotide microsatellites to predict their $\Delta\Delta G^\circ$. The model construction started with the theoretical work using the partition function to identify 40 common bulge states of sliding bulges, followed by careful $\Delta\Delta G^\circ$ measurements with TEEM. We first tested the effect of NNNs on bulge $\Delta\Delta G^\circ$ and selected the representative NNNs for the systematic study. Based on the groundwork, $\Delta\Delta G^\circ$ values of sliding bulges necessary to the model were measured.

Although Zhu and Wartell (31) did recognize structural degeneracy of a sliding bulge as the reason for its lower free energy, they failed to develop the observation into a proper model. After testing 10 sliding bulges, they used a single empirical constant to account for two- and three-slide bulges without realizing that the length of a sliding bulge affects

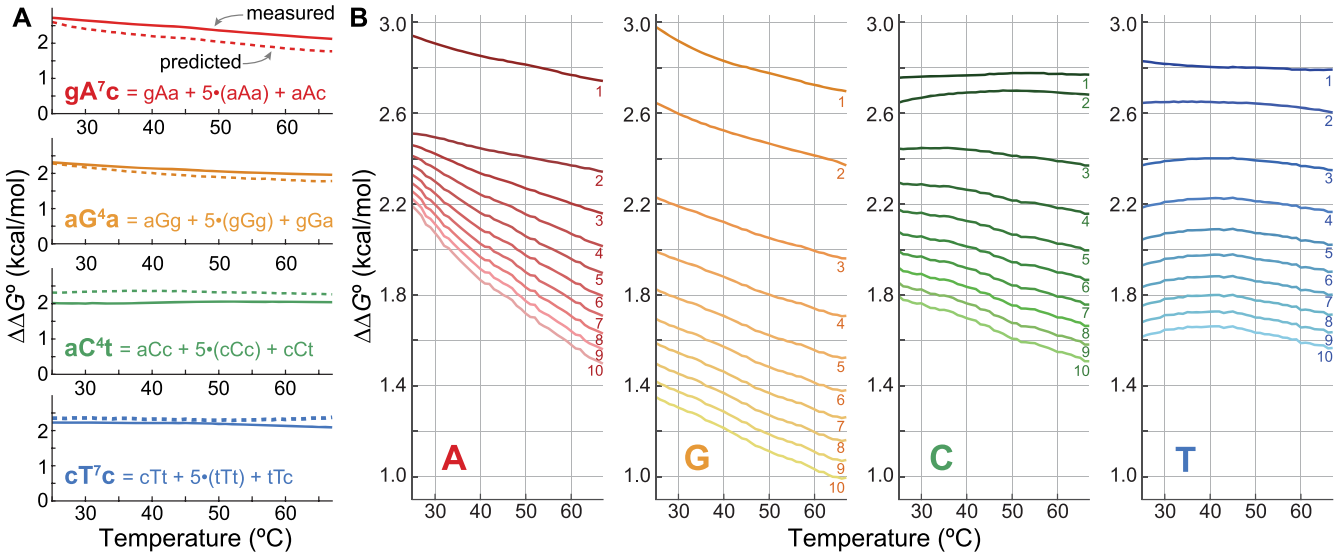

Figure 5. Validation and summary of sliding bulge $\Delta\Delta G^\circ$ predictions. **(A)** $\Delta\Delta G^\circ$ of four longer sliding bulges predicted by our model (solid lines) and experimentally measured by TEEM (dotted lines). A median and a maximum absolute value of residual were 0.22 and 0.37 kcal/mol, respectively. **(B)** Summary of predictions. Each line shows a mean of $\Delta\Delta G^\circ$ of all bulges with a given length and a bulged base, and a number next to each plot denotes the number of tandem repeats of a mononucleotide. Steep $\Delta\Delta G^\circ$ slopes of A bulges and low $\Delta\Delta G^\circ$ of G bulges reflect characteristics of their triplet states' $\Delta\Delta G^\circ$.

$\Delta\Delta G^\circ$. Thus, our model is the first to predict the free energy of longer sliding bulges, and it was based on the comprehensive analysis of the thermodynamics of sliding bulges. Moreover, our model construction principle can be further expanded to studying dinucleotide or trinucleotide sliding bulges that cause various neurological diseases (33–36).

With the theoretical background and the systematic data collection with TEEM, our model provides explanations to the experimental results of the literature on MSI. For example, researchers have observed that longer microsatellites tend to have higher mutation rates (37,38), which has been clearly elucidated by our thermodynamic model of sliding bulges based on the partition function. Our model also implies that sliding G bulges with the lowest penalty will dramatically stabilize longer microsatellites and drive mutation rates of C/G microsatellites up. Indeed, the longest C/G mononucleotide microsatellite from exome sequencing on 24 colorectal tumors was twice as long as the longest A/T mononucleotide microsatellite (81 bp versus 40 bp) (38). The same study revealed that the mutation rates of C/G microsatellites were 7.5 times higher on average than that of A/T when they had the same length and <10% margin of error. Other studies on MSI with human cancer cell lines reported similar results with C/G microsatellites showing 7 times (39) or 4.4 times (40) higher mutation rates.

In addition to offering theoretical explanations to landscapes of MSI, our model can help design experiments studying MSI. From a high-level point of view, the predictions made by the model can be used as a general guideline. Figure 5B shows how sequence and temperature affect $\Delta\Delta G^\circ$ of sliding bulges at a microsatellite, and such difference in penalty should be considered according to detail of an experiment. For instance, genotyping indels at poly(G) with a probe or a primer would be more difficult than genotyping poly(C) on the opposite strand due to low $\Delta\Delta G^\circ$ of sliding G bulges. And when probe hybridization protocol involves a temperature change due to washing steps, targeting poly(T) will be more consistent than targeting poly(A) because its $\Delta\Delta G^\circ$ is less dependent on temperature. The predicted $\Delta\Delta G^\circ$ values can also be used for fine-tuning probe or primer hybridization if more sophisticated approach is desired. By definition, adding $\Delta\Delta G^\circ$ of a sliding bulge to hybridization ΔG° of an original sequence without a sliding bulge gives the actual hybridization ΔG° of a sliding bulge formation. An equilibrium constant of hybridization can then be calculated from the resulting ΔG° and temperature, providing an estimation for a probe or a primer binding yield.

The purpose of the model construction was to predict $\Delta\Delta G^\circ$ of sliding bulges, but the non-slide bulge datasets acquired in the process are also useful by themselves. Those bulges can be easily formed during molecular biology experiments when primers or probes are hybridized with indel variants or non-specific targets. Thus, it is important to first predict DNA hybridization to properly design the experiments. There are a few data on the thermodynamics of non-slide bulges, but they were either a simple approximation with no experimental data (41) or measured at biologically irrelevant salinity with significant errors (42). In contrast, our non-slide bulge data from TEEM do not suffer from any of these problems (Supplementary Section S4) and can be readily applied to such predictions.

It is interesting that sliding bulges show clear differences according to whether they have a purine or a pyrimidine as a bulged base. As we constantly observed in non-, two- and three-slide bulges, $\Delta\Delta G^\circ$ of purine bulges decrease as temperature goes up. The temperature dependence of purine bulge $\Delta\Delta G^\circ$ implies that the entropy may be behind it. One possibility is that a bulkier purine base, which is a pyrimidine ring fused to an imidazole ring, causes more disorder

when bulged out from a perfect double-helix structure. For example, larger purine bases could create larger hydration shells that limit mobility of water and, as a result, increase entropic impact. As a matter of fact, the literature on crystallographic data of unpaired RNA bases (43) and statistical analysis of hydration levels (44) showed purine bases had more water molecules around them than pyrimidine bases (Supplementary Table S3).

Another outcome of this work is a possibility of a sliding bulge that is more stable than a typical double-helix structure. In theory, a sliding bulge with enough tandem repeats could have so high structural degeneracy that its stabilizing effect overcomes thermodynamic penalty of having a bulge. Our model predicts that this singularity will happen when there are >41 G, >91 C, >101 A or >104 T repeats. Although such bulges will rarely appear *in vivo* or *in vitro* due to their lengths, their implication for a method of designing stable structures with structural degeneracy is intriguing.

An important limitation of our model is that errors from using the representative NNNs cannot be avoided. Because a bulge may disrupt a local helix structure, we used the representative NNNs throughout this work for consistency. This strategy effectively standardized bulge $\Delta\Delta G^\circ$ for the systematic model construction and prevented extreme $\Delta\Delta G^\circ$ deviation. However, some errors will always exist when the actual NNN is different from the representative NNNs unless we measure $\Delta\Delta G^\circ$ of every bulge with every NNN, which is impractical in terms of time and cost. In Figure 2A, the mean and the standard deviation of gaps between measured $\Delta\Delta G^\circ$ and their corresponding representative $\Delta\Delta G^\circ$ were 0.15 and 0.008 kcal/mol, respectively. See Supplementary Section S5 for a summary of all $\Delta\Delta G^\circ$ error values from the NNN choices. This error should be considered when the NNN is different from the representative NNNs used in this study.

DATA AVAILABILITY

All experimental data are plotted in the Supplementary Materials, and numerical data files are available upon request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Alice Lee for editorial assistance.
Author contributions: J.H.B. conceived the project, designed and conducted the experiments, analyzed the data and wrote the paper. D.Y.Z. conceived the project, analyzed the data and wrote the paper.

FUNDING

National Human Genome Research Institute [R01HG008752 to D.Y.Z.]. Funding for open access charge: National Human Genome Research Institute [R01HG008752].
Conflict of interest statement. D.Y.Z. declares a competing interest in the form of consulting for and equity ownership in NuProbe USA, Torus Biosystems and Pana Bio.

REFERENCES

1. Le,D.T., Uram,J.N., Wang,H., Bartlett,B.R., Kemberling,H., Eyring,A.D., Skora,A.D., Luber,B.S., Azad,N.S., Laheru,D. *et al.* (2015) PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.*, **372**, 2509–2520.
2. Ribic,C.M., Sargent,D.J., Moore,M.J., Thibodeau,S.N., French,A.J., Goldberg,R.M., Hamilton,S.R., Laurent-Puig,P., Gryfe,R., Shepherd,L.E. *et al.* (2003) Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *N. Engl. J. Med.*, **349**, 247–257.
3. Kim,G.P., Colangelo,L.H., Wieand,H.S., Paik,S., Kirsch,I.R., Wolmark,N. and Allegra,C.J. (2007) Prognostic and predictive roles of high-degree microsatellite instability in colon cancer: A National Cancer Institute–National Surgical Adjuvant Breast and Bowel Project Collaborative Study. *J. Clin. Oncol.*, **25**, 767–772.
4. Bertagnolli,M.M., Niedzwiecki,D., Compton,C.C., Hahn,H.P., Hall,M., Damas,B., Jewell,S.D., Mayer,R.J., Goldberg,R.M., Saltz,L.B. *et al.* (2009) Microsatellite instability predicts improved response to adjuvant therapy with irinotecan, fluorouracil, and leucovorin in stage III colon cancer: Cancer and Leukemia Group B Protocol 89803. *J. Clin. Oncol.*, **27**, 1814–1821.
5. Sznajder,Ł.J., Thomas,J.D., Carrell,E.M., Reid,T., McFarland,K.N., Cleary,J.D., Oliveira,R., Nutter,C.A., Bhatt,K., Sobczak,K. *et al.* (2018) Intron retention induced by microsatellite expansions as a disease biomarker. *Proc. Natl Acad. Sci. U.S.A.*, **115**, 4234–4239.
6. Lagerstedt Robinson,K., Liu,T., Vandrovcova,J., Halvarsson,B., Clendenning,M., Frebourg,T., Papadopoulos,N., Kinzler,K.W., Vogelstein,B., Peltomäki,P. *et al.* (2007) Lynch syndrome (hereditary nonpolyposis colorectal cancer) diagnostics. *J. Natl Cancer Inst.*, **99**, 291–299.
7. Blouin,J.-L., Dombroski,B.A., Nath,S.K., Lasseter,V.K., Wolyniec,P.S., Nestadt,G., Thornquist,M., Ullrich,G., McGrath,J., Kasch,L. *et al.* (1998) Schizophrenia susceptibility loci on chromosomes 13q32 and 8p21. *Nat. Genet.*, **20**, 70–73.
8. Lynch,H.T. and de la Chapelle,A. (2003) Hereditary colorectal cancer. *N. Engl. J. Med.*, **348**, 919–932.
9. Popat,S., Hubner,R. and Houlston,R.S. (2005) Systematic review of microsatellite instability and colorectal cancer prognosis. *J. Clin. Oncol.*, **23**, 609–618.
10. Vilar,E. and Gruber,S.B. (2010) Microsatellite instability in colorectal cancer—the stable evidence. *Nat. Rev. Clin. Oncol.*, **7**, 153–162.
11. Phipps,A.I., Limburg,P.J., Baron,J.A., Burnett-Hartman,A.N., Weisenberger,D.J., Laird,P.W., Sinicrope,F.A., Rosty,C., Buchanan,D.D., Potter,J.D. *et al.* (2015) Association between molecular subtypes of colorectal cancer and patient survival. *Gastroenterology*, **148**, 77–87.
12. Lochhead,P., Kuchiba,A., Imamura,Y., Liao,X., Yamauchi,M., Nishihara,R., Qian,Z.R., Morikawa,T., Shen,J., Meyerhardt,J.A. *et al.* (2013) Microsatellite instability and BRAF mutation testing in colorectal cancer prognostication. *J. Natl Cancer Inst.*, **105**, 1151–1156.
13. Hagelberg,E., Gray,I.C. and Jeffreys,A.J. (1991) Identification of the skeletal remains of a murder victim by DNA analysis. *Nature*, **352**, 427–429.
14. Jobling,M.A. and Gill,P. (2004) Encoded evidence: DNA in forensic analysis. *Nat. Rev. Genet.*, **5**, 739–751.
15. Gill,P., Ivanov,P.L., Kimpton,C., Piercy,R., Benson,N., Tully,G., Evett,I., Hagelberg,E. and Sullivan,K. (1994) Identification of the remains of the Romanov family by DNA analysis. *Nat. Genet.*, **6**, 130–135.
16. Algee-Hewitt,B.F.B., Edge,M.D., Kim,J., Li,J.Z. and Rosenberg,N.A. (2016) Individual identifiability predicts population identifiability in forensic microsatellite markers. *Curr. Biol.*, **26**, 935–942.
17. Rubinsztein,D.C., Amos,W., Leggo,J., Goodburn,S., Jain,S., Li,S.-H., Margolis,R.L., Ross,C.A. and Ferguson-Smith,M.A. (1995) Microsatellite evolution—evidence for directionality and variation in rate between species. *Nat. Genet.*, **10**, 337–343.
18. Nybom,H. (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Mol. Ecol.*, **13**, 1143–1155.
19. Weir,B.S., Anderson,A.D. and Hepler,A.B. (2006) Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.*, **7**, 771–780.

20. Bowers,J., Boursiquot,J.-M., This,P., Chu,K., Johansson,H. and Meredith,C. (1999) Historical genetics: the parentage of Chardonnay, Gamay, and other wine grapes of northeastern France. *Science*, **285**, 1562–1565.
21. Bagshaw,A.T.M. (2017) Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biol. Evol.*, **9**, 2428–2443.
22. de la Chapelle,A. (2003) Microsatellite instability. *N. Engl. J. Med.*, **349**, 209–210.
23. Woerner,S.M., Yuan,Y.P., Benner,A., Korff,S., von Knebel Doeberitz,M. and Bork,P. (2010) SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology. *Nucleic Acids Res.*, **38**, D682–D689.
24. Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.
25. Hause,R.J., Pritchard,C.C., Shendure,J. and Salipante,S.J. (2016) Classification and characterization of microsatellite instability across 18 cancer types. *Nat. Med.*, **22**, 1342–1350.
26. SantaLucia,J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. U.S.A.*, **95**, 1460–1465.
27. Wang,C., Bae,J.H. and Zhang,D.Y. (2016) Native characterization of nucleic acid motif thermodynamics via non-covalent catalysis. *Nat. Commun.*, **7**, 10319.
28. Bae,J.H., Fang,J.Z. and Zhang,D.Y. (2020) High-throughput methods for measuring DNA thermodynamics. *Nucleic Acids Res.*, **48**, e89.
29. Dirks,R.M. and Pierce,N.A. (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, **24**, 1664–1677.
30. Joshua-Tor,L., Frolow,F., Appella,E., Hope,H., Rabinovich,D. and Sussman,J.L. (1992) Three-dimensional structures of bulge-containing DNA fragments. *J. Mol. Biol.*, **225**, 397–431.
31. Zhu,J. and Wartell,R.M. (1999) The effect of base sequence on the stability of RNA and DNA single base bulges. *Biochemistry*, **38**, 15986–15993.
32. McCann,M.D., Lim,G.F.S., Manni,M.L., Estes,J., Klapec,K.A., Frattini,G.D., Knarr,R.J., Gratton,J.L. and Serra,M.J. (2011) Non-nearest-neighbor dependence of the stability for RNA group II single-nucleotide bulge loops. *RNA*, **17**, 108–119.
33. Spada,A.R. La, Wilson,E.M., Lubahn,D.B., Harding,A.E. and Fischbeck,K.H. (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, **352**, 77–79.
34. Verkerk,A.J.M.H., Pieretti,M., Sutcliffe,J.S., Fu,Y.-H., Kuhl,D.P.A., Pizzuti,A., Reiner,O., Richards,S., Victoria,M.F., Zhang,F. *et al.* (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, **65**, 905–914.
35. Harley,H.G., Brook,J.D., Rundle,S.A., Crow,S., Reardon,W., Buckler,A.J., Harper,P.S., Housman,D.E. and Shaw,D.J. (1992) Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. *Nature*, **355**, 545–546.
36. Klement,I.A., Skinner,P.J., Kaytor,M.D., Yi,H., Hersch,S.M., Clark,H.B., Zoghbi,H.Y. and Orr,H.T. (1998) Ataxin-1 nuclear localization and aggregation: role in polyglutamine-induced disease in SCA1 transgenic mice. *Cell*, **95**, 41–53.
37. Sia,E.A., Kokoska,R.J., Dominska,M., Greenwell,P. and Petes,T.D. (1997) Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol. Cell. Biol.*, **17**, 2851–2858.
38. Kondelin,J., Gylfe,A.E., Lundgren,S., Tanskanen,T., Hamberg,J., Aavikko,M., Palin,K., Ristolainen,H., Katainen,R., Kaasinen,E. *et al.* (2017) Comprehensive evaluation of protein coding mononucleotide microsatellites in microsatellite-unstable colorectal cancer. *Cancer Res.*, **77**, 4078–4088.
39. Boyer,J.C., Yamada,N.A., Roques,C.N., Hatch,S.B., Riess,K. and Farber,R.A. (2002) Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. *Hum. Mol. Genet.*, **11**, 707–713.
40. Zhang,L., Yu,J., Willson,J.K. V, Markowitz,S.D., Kinzler,K.W. and Vogelstein,B. (2001) Short mononucleotide repeat sequence variability in mismatch repair-deficient cancers. *Cancer Res.*, **61**, 3801–3805.
41. SantaLucia,J. and Hicks,D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 415–440.
42. Tanaka,F., Kameda,A., Yamamoto,M. and Ohuchi,A. (2004) Thermodynamic parameters based on a nearest-neighbor model for DNA sequences with a single-bulge loop. *Biochemistry*, **43**, 7143–7150.
43. Kirillova,S. and Carugo,O. (2011) Hydration sites of unpaired RNA bases: a statistical analysis of the PDB structures. *BMC Struct. Biol.*, **11**, 41.
44. Schneider,B. and Berman,H.M. (1995) Hydration of the DNA bases is local. *Biophys. J.*, **69**, 2661–2669.