DEPARTMENT OF PHYSICS

University of Strathclyde Science

PH 570 Project Submission Page

Project Title: Coarse-Grained Modelling of DNA Triplet-Repeat Slip-out Migration: A Case for Well-Tempered and Transition-Tempered Metadynamics.

Student Registration Number: 201817988

Date of submission: 02/02/2023

By submitting this work I acknowledge that this work is my own and accept that the submission will be processed through Turnitin to detect plagiarism.

NB: Please use your student number as the Author on the project paper, with the University of Strathclyde affiliation.

Coarse-Grained Modelling of DNA Triplet-Repeat Slip-out Migration: A Case for Well-Tempered and Transition-Tempered Metadynamics.

Lewis MacLeod Russell[1]

[1]*SUPA Department of Physics, University of Strathclyde, Glasgow G4 0NG, United Kingdom.*

Repeat sequences of DNA bases have been shown to form secondary hairpin structures, and although these have been attributed to a number of degenerative diseases, details of their structure and dynamics remain largely elusive. Three-way double-stranded DNA (dsDNA) junctions containing slip-out hairpins of CAG or CTG repeats have been recently shown to undergo branchpoint migration, where the hairpin effectively transverses along the main dsDNA strand. Herein, coarse-grained modelling of CAG hairpins are explored and appropriate metadynamic sampling is considered. Recent experimental observations provide evidence of structural heterogeneity in these hairpins, as was comparatively perceived in the coarse-grained modelling. Our results for larger hairpins also suggests the junction does not fully denature and re-hybridise in one direct branchpoint migration process, but rather the process occurs incrementally; the hairpin denatures and re-hybridises through smaller intermediate and even metastable transitional states. These original insights are of particular relevance when considering their potential impact on DNA processing pathways, and it may hold important information for future understanding and treatments of triplet-repeat expansion diseases.

Introduction.– DNA is formed of molecules formally known as nucleotides, each of which comprises of a sugar-phosphate backbone and a nitrogen base. These nitrogen bases are further categorised into four types with specific complimentary binding: adenine (A) pairs with thymine (T), and guanine (G) with cytosine (C). Expansion of simple repeat sequences of these bases are currently known to be responsible for almost 50 human genetic disorders [1], the vast majority of which are not currently preventable or curable. Collectively, these types of disorders are known as "repeat-expansion-diseases" (REDs); almost all of which are severe, degenerative, and significantly reduce both lifespan and quality of life.

These REDs can be further categorised into 13 different types of tandem repeats, with the most common sequence being CAG, or complementary CTG [1]. Combined, these account for some of the more common and well-known REDs, each affecting from approximately 1 in 10,000 individuals to 1 in 100,000 individuals [2, 3]: these include Huntington's Disease [2, 4], Spinocerebellar Ataxias [1, 3, 5], and Myotonic Dystrophy [6] for reference.

Research into the expansion of these trinucleotide repeats (TNR) has been substantial over the past two to three decades [7], though despite their well-recognised association to neurological and neuromuscular diseases, many components eluding to their fundamental molecular dynamics remain obscured. Nevertheless, some significant higher-level determinations have been made. Multiple studies have shown that the number of repeats correlates with both disease severity and an earlier age of symptomatic onset [7–11]. And as for the formation of these tandem repeats, sufficiently longer strands are capable of forming secondary structures [8–10], with hairpin slip-outs being the most widely studied [9, 12–14].

These hairpins create a three-way dsDNA junction (3WJ), illustrated in Figure 1. Two important findings of these partially self-complementary hairpins have been recently reported [10]. When the TNR extends into the adjacent duplex, both localised conformation and branchpoint migration can occur. The first of these statements describes the form of the junction itself, where it can be T-shaped (as in Figure 1), Y-shaped, or take a less energetically favourable and distorted formation. As for the longer-range interconversion associated to branchpoint migration, denaturing and rehybridisation of base-pairs occurs and the junction can relocate along the main dsDNA strand.

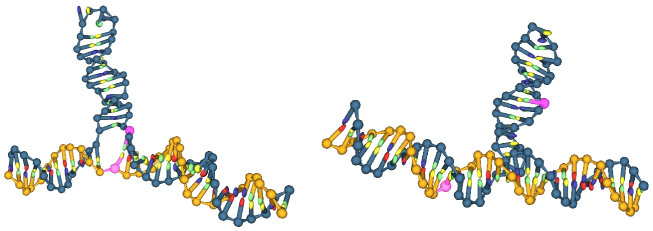

FIG. 1. An example visualisation of a three-way DNA junction, with the hairpin slip-out represented in the blue strand. Providing the main dsDNA strand contains a sequence of complementary base-pairs to that found in the hairpin, the location of the junction can reposition along the main dsDNA segment as shown left to right in the Figure; a process known as "branchpoint migration". The two nucleotides marked in purple are highlighted as a visual guide.

The significance of these dynamics is not yet determined, though is believed to potentially affect disease progression and could act as a target for therapeutic solutions [10]. Bianco et al [9] recently performed further analysis of these 3WJs and hairpins, experimentally employing single-molecule-FRET (smFRET) analysis and complementing this with coarse-grained modelling [15–18]. smFRET analysis probed the state-to-state kinetics

and dwell times on three closely related triplet-repeat sequences: CAG_{10}, CTG_{30}, and CAG_{40}. However, the coarse-grained modelling was confined only to the CAG_{10} sequence. In this Report, we explore additional modelling techniques on the CAG_{10} sequence, propose how best these techniques can be utilised, and discuss some of the current insights and limitations found when extending the TNR range to CAG_{20} and beyond.

Our coarse-grained model of choice is oxDNA2 [15–18]. Applying a top-down approach to coarse-grained modelling, each nucleotide is represented as a three-interaction-site rigid body. Specifically, these sites are: a hydrogen-bonding/excluded volume site, a stacking site, and a backbone/electrostatic interaction site [17]. All bonded and pair interaction potentials can be summarised as:

$$V = \sum_{<ij>} (V_{backbone} + V_{stack} + V_{exc}) + \sum_{i,j \neq <ij>} (V_{HB} + V_{cr.st.} + V'_{exc} + V_{coax} + V_{dh}) \quad (1)$$

The first sum in Equation 1 represents nearest neighbour nucleotides, and the second sum comprises all remaining pairwise interactions. The terms listed represent: backbone connectivity ($V_{backbone}$), stacking between adjacent bases on a strand (V_{stack}), excluded volume (V_{exc} and V'_{exc}), hydrogen bonding between complementary bases (V_{HB}), duplex axis cross-stacking ($V_{cr.st.}$), coaxial stacking across a nicked backbone (V_{coax}), and a Debye-Hückel potential (V_{dh}). The Debye-Hückel potential was a new addition implemented in oxDNA2 [16, 18], and defines an implicit electrostatic potential representing different salt concentrations within the modelling solution [16]. The backbone potential is a Finitely Extensible Nonlinear Elastic (FENE) bond mimicking the covalent bonds along the strand, and the excluded volume and backbone interactions are distance-dependent functions between repulsion sites. All other potentials are dependent on the relative orientation of nucleotides as well as the distances between the stacking and hydrogen-bonding interaction sites.

The use of Langevin dynamics are also applied within the model; this mimics the effects of the surrounding thermal collisions acting on the DNA molecule, as well as the viscosity effects from the solution. This is defined in Equation 2 as:

$$m\ddot{\boldsymbol{x}} = \boldsymbol{F}_c + \boldsymbol{F}_f + \boldsymbol{F}_r = -\frac{dV(\boldsymbol{r})}{d\boldsymbol{r}} - \gamma m\boldsymbol{v} + \sqrt{2m\gamma k_b T}\boldsymbol{\eta}(t) \quad (2)$$

Where: $\boldsymbol{F}_c$ is a conservative force and is potential-dependent, $\boldsymbol{F}_f$ is the velocity-dependent frictional force, and $\boldsymbol{F}_r$ is a temperature-dependent random force representing thermal collisions. The Langevin friction coefficient γ controls the viscosity of the solution, $\boldsymbol{\eta}(t)$ represents the stochastic force contribution, and k_bT is the Boltzmann constant multiplied by system temperature. It should be stated that the mean average value of the stochastic term $\boldsymbol{\eta}(t)$ should be zero. The value may be negligible or significant at each timestep, although within a sufficiently large number of timesteps it should average to zero. This accurately describes the random collisions from the solution, without introducing an unwanted bias.

Whilst Equation 2 is given in its translational form $V(\boldsymbol{r})$, the rotational equivalent is also applied to the orientation-dependant potentials from Equation 1. The overall form remains the same, although the friction coefficient γ is modelled ten times larger for orientational degrees of freedom as it is for the translational counterparts.

Comparing Metadynamic Flavours on CAG_{10} and CAG_{20} Hairpins.– In its double-stranded form, the hydrogen-bonding of nucleotides results in DNA being far more stable than when it is in the more flexible single-stranded (ssDNA) state. This is commonly characterised by persistence length, a measure of a polymer's bending stiffness and its susceptibility to bending under thermal fluctuations. The difference between ssDNA and dsDNA is substantial, around 5 and 150 base-pairs in length respectively [19, 20]. Physically, this represents a naturally stable structure for dsDNA.

Additionally, our coarse-grained modelling is limited to time-scales up to milliseconds [21], whereas branchpoint migration occurs in the range of seconds [9]; if significant conformational change is to be seen in dsDNA-hairpin modelling, it is necessary to bias the system's free energy. The branchpoint migration process itself requires the 3WJ to overcome large energy barriers, and so to achieve practical simulation times, we apply the use of metadynamics in order to enhance the sampling of these rare states [22–26].

The general principle of metadynamics is as follows, and requires the preliminary identification of Collective Variables (CVs); suitable CVs are those that can characterise the transitions between conformations of interest. Firstly, the system evolves as per its unbiased dynamics, then a Gaussian energy potential is introduced that alters its free-energy landscape. These energy packets, known as "HILLS", are built-up and allow the system to evolve into new states - conformations with naturally higher free energies will allow evolution under lesser biasing that those of lower free energies. The HILLS themselves are placed at the average value of the CVs during a given time interval, and thus discourages the system from further exploring these areas. In essence, the metadynamics biasing can be considered converged once the system has an appropriately adapted free-energy landscape conducive to complete and free exploration.

One of the largest drawbacks to Standard (Direct) Metadynamics is that it lacks effective adaptation of hill-heights. This leaves the user with the difficult decision of when to terminate a run, as the free energy does not converge to a definite value, but rather fluctuates around a general region [22]. It can be unclear if runs have reached convergence, or if regions of the free-energy landscape are under or over represented. The Standard Metadynamics algorithm, applying the biased energy potential $V_G(S(x),t)$, is as follows [23]:

$$V_G(S(x),t) = \omega \sum_{t'=\tau_G,2\tau_G,\ldots}^{t'<t} exp\left(-\frac{(S(x)-s(t'))^2}{2\delta s^2}\right) \quad (3)$$

In Equation 3, and in those to follow, $S(x)$ denotes a function of co-ordinates (Contact Maps - *"CM"*), $s(t')$ the values of the CVs at a given time, ω and δs are the Gaussian height and width respectively, and the time interval at which Gaussians are added is defined by τ_G.

Given the primary motivation of our simulations is to witness branchpoint migration in the dsDNA-hairpin, metadynamics governed by Equation 3 would require each individual run to be terminated whenever conformations in such free-energy regions become diffusive. This would be cumbersome, as each run would have a different number of timesteps and require tailored post-processing of data. Furthermore, the lack of effective hill-height adaptation means that we'd either have to run significantly longer simulations will small hill-heights, or increase the hill-heights to larger values and risk losing conformational detail in the region of slip events. For these reasons alone, Standard Metadynamics was never tested in our runs. Instead, we first consider the Well-Tempered flavour of metadynamics (WTMetaD), in line with the method of choice from Bianco et al [9].

In this version of metadynamics, the Well-Tempered flavour adds an addition scaling factor to each timestep, given in Equation 4 [23] by:

$$V_G(S(x),t) = \omega \sum_{t'=\tau_G,2\tau_G,\ldots}^{t'<t} exp\left(-\frac{V_G(S(x),t')}{k_b\Delta T}\right) exp\left(-\frac{(S(x)-s(t'))^2}{2\delta s^2}\right) \quad (4)$$

Where k_b is the Boltzmann Constant, and ΔT is given in units of temperature. The result of this additional exponential decay is that convergence of the biasing is smoothed; the hills are first applied in a relatively severe manner, and then reduced over time so that each applied hill has less subsequent contribution over the free-energy landscape. In addition, the metadynamics library used in our tests (PLUMED [27, 28]) also re-scales the Gaussian heights based on a defined bias factor. This allows the user to tune the aggressiveness of this time-dependent hill-height reduction to their desire.

For the CAG_{10} runs using WTMetaD (herein referred to as "WT_CAG_{10}"), an initial test suite of bias factors was configured; 6,8,10,12, and 14. Although WTMetaD has much improved convergence over Standard Metadynamics, the same issue of viable aggressiveness still applies - the bias factor must be large enough that conformational exploration is witnessed within a practical simulation time, though small enough that substantial detail is retained. It was found that a bias factor of 8 was appropriate for CAG_{10}, in accordance with Bianco et al [9]. Using various random seeds across multiple runs, full slip events where commonly witnessed.

These slip events shared common similarities. The transition left to right in Figure 1 involves incremental migration via partial denaturing and rehybridisation - the hairpin rarely denatures more than three or four base-pairs at a time. In the latter stages of the migration process, it is not uncommon for the denatured base-pairs to actually form more energetically favourable stacked arrays or bulges. Even if these stacked segments are only two or three base-pairs in length, this can be enough to prevent full branchpoint *and* hairpin migration. This leaves some post-slippage base-pairs unable to hybridise with one-another, and the migrated hairpin takes on a slightly more frustrated form to that shown in the right of Figure 1.

Along side the use of trajectory visualisation tools, a heatmap of the free-energy landscape (shown in Figure 2) can also help explain this behaviour. These heatmaps provide beneficial insight into the thermodynamic properties of the migration process. Figure 2 plots the averaged free energy against the two CVs used in our runs; the "Denatured bps" representing base-pairs that must open to allow slippage, and "Hybridised bps" defining those that have re-hybridised into a post-slip conformation. The two deep wells around (3,3) and (15,15) account for our initial pre-slipped then post-slipped conformations, respectively. Both wells feature a relatively accessible region surrounding one or two CVs, meaning that the system is able to arrange one or two base-pairs without great difficulty.

Starting from the initial well (3,3) and continuing straight upwards within the confines of the bottom $-45k_bT$ contour, we see that base-pairs first denature until the free-energy gradient increases, at which point re-hybridisation becomes the more favourable option. This can be seen within the right-hand side confines of the bottom $-40k_bT$ contour. At this stage, more base-pairs open up and unpaired nucleotides increase. This leads to the two outcomes described previously: either the hairpin begins to re-hybridise supportively of post-slip pairing, or the unpaired nucleotides instead form the aforementioned and metastable stacked bulges. These states

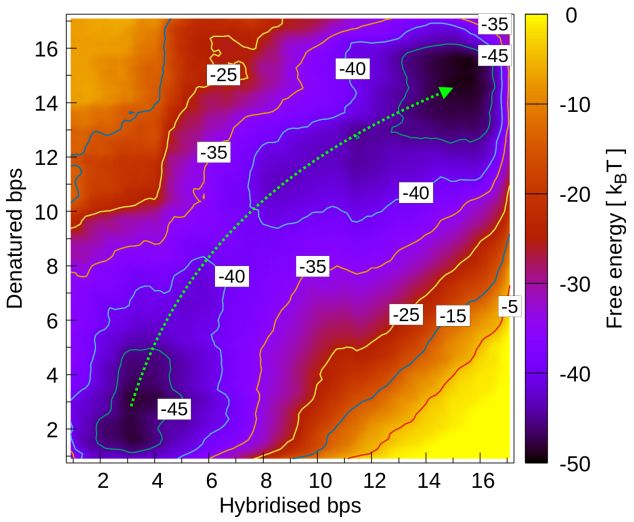

FIG. 2. The Free-Energy Landscape Sampled Across Simulations of the CAG_{10} Hairpin. This was sampled and averaged across WTMetaD runs that displayed clear branchpoint migration. The initial state of the system is represented by the bottom energy well around (3,3), with slipped states belonging to the upper well around (15,15). The presence of the large free energy barrier signifies the necessity of the applied metadynamics, although the dashed green arrow shows an approximate path of least resistance along the diagonal. Intuitively, this represents transitional conformations where the branchpoint and hairpin first denature and then partially slip into re-hybridised intermediate conformations.

are located roughly in the centre of the dashed green arrow. States encouraging post-slip pairing will continue into the upper deep well (15,15), with metastable states tending to linger within the confines of the upper $-40k_bT$ contour. These steps described will differ between independent runs with different random seeds, although are generically indicative of the dynamics observed across WT_CAG_{10} runs.

For reference, each of the WTMetaD simulations required 500M timesteps in order to sufficiently explore post-slip conformations, totalling around 50 hours of simulation time per run on our compute cluster. This is a relatively large run time considering our small system size for CAG_{10} (48 base-pairs). Despite this being adequate once averaged across all runs, some still display poor exploration of the free-energy-landscape due to the stochastic variance between randomised seeds.

This is a well-known caveat of WTMetaD. Overall, it is a practical algorithm for adaptively enhanced sampling, though its reliance on pre-defined parameter tuning leaves some runs lacking effective state sampling. As the length of the hairpin is increased and thus too the CV definition, this effect is only exacerbated. This restriction in WTMetaD inspired the development of the Transition-Tempered flavour of metadynamics (TTMetaD), promising substantial speed and accuracy improvements over WTMetaD [26]. This method focuses on "Fill First, Then Converge", and achieves rapid and asymptotic convergence within the free-energy-landscape whilst avoiding the overcompensation of hills. Unlike WTMetaD, the TTMetaD hills have a variable bias factor that introduces a dependence on the free-energy-landscape and CVs. TTMetaD is well-suited to studying branchpoint and hairpin migration within dsDNA, as it is specifically designed to study systems in which conformations of interest are a prior known condition, though transition mechanics are not.

TTMetaD determines convergence via analysis of transition paths. More specifically, the algorithm detects maximally biased paths defined where the minimum bias within these paths is maximal. Consider first the initial phases of a TTMetaD run in which basins are not entirely filled by the applied bias. This means that the bias surface will contain a number of hills within partially filled basins, however some regions will still be unbiased.

Taking the minimum bias of these paths connecting partially filled basins will of course return zero, since hills are still separated by unbiased regions. Formally, this additional biasing condition can be written in Equation 5 as [26]:

$$V^*(t) \equiv \max_{s(\lambda)\in\mathcal{P}} \min_{\lambda\in[0,1]} V(s(\lambda),t) \tag{5}$$

Where $s(\lambda) \in \mathcal{P}$ represents the set of all continuous paths between basins, and λ indexes the points along each path. By this definition, the algorithm requires positional CM co-ordinates for the basins of interest. More specifically, these where defined in our modelling to the initial and post-slip conformations of: CAG_{10}, ([3,3],[15,15]); and CAG_{20}, ([3,3],[26,26]).

Substituting Equation 5 into the TTMetaD equation (Equation 6) reveals that tempering of the biased potentials will only occur once all regions along paths contain overlapping hills. Otherwise, biasing is unaffected as $V^*(t')$ returns zero and subsequently $exp\left(-\frac{V^*(t')}{k_b\Delta T}\right) = 1$.

$$V_G(S(x),t) = \omega \sum_{t'=\tau_G,2\tau_G,\ldots}^{t'<t} exp\left(-\frac{V^*(t')}{k_b\Delta T}\right) exp\left(-\frac{(S(x)-s(t'))^2}{2\delta s^2}\right) \tag{6}$$

Figure 3 highlights the superior convergence time of TTMetaD over WTMetaD, illustrated here plotting the contact maps (CM) against time. Both plots in this Figure share the same random seed, though the transient period in reaching exploratory post-slip dynamics is notably improved in the TTMetaD approach. TTMetaD obtains high contact map values within 40M timesteps, though comparatively is this far larger for WTMetaD

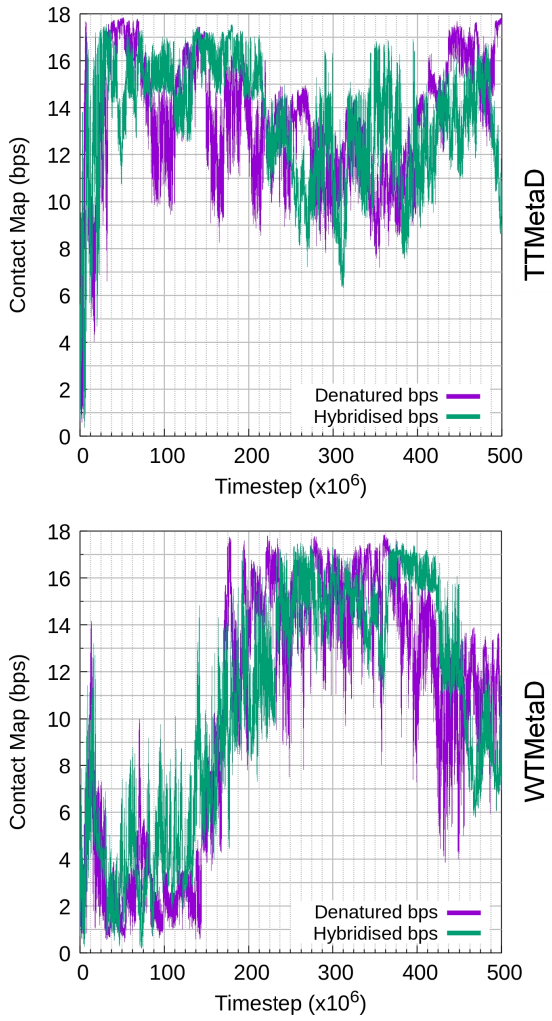

FIG. 3. Contact Maps Highlighting the Effects on System Evolution Between WTMetaD (bottom) and TTMetaD (top). The rapid increase in the contact map values for TTMetaD over WTMetaD highlights its effectiveness in applying large hills within the first energy basin (initial condition). When the contact map has reached the second energy basis defining the post-slip region, the hill heights are significantly reduced only once all paths between the two basins have entered previously unbiased regions.

taking around 250M timesteps. An average of the free-energy-landscape for TTMetaD (Figure 4) runs also indicates that the free-energy estimates are consistent with WTMetaD (Figure 2), albeit with slightly more jagged contour lines.

It can be concluded for CAG_{10} that TTMetaD is the superior choice over WTMetaD - it provides the same results whilst requiring substantially less runtime in simulations. For completeness however, both flavours where still utilised for our CAG_{20} simulations; the motivation behind this decision was to determine if more substantial discrepancies would emerge under the increased system size (63 base-pairs for CAG_{20}, up from 48bps in CAG_{10}).

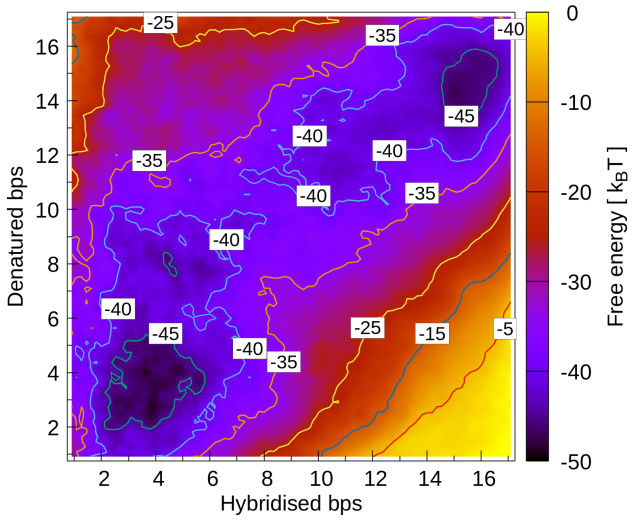

FIG. 4. The TTMetaD Alternative to Figure 2, showing the Free-Energy Landscape Sampled Across Simulations of the CAG_{10} Hairpin. In comparison to WTMetaD on CAG_{10} (See Figure 2), there is little difference in the free-energy landscape. However, the transient period in reaching post-slip convergence is much reduced (See Figure 3).

The extension of the hairpin length also inspired the addition of a new and unbiased CV definition. Highlighted in Figures 5 and 6, these provide increased monitoring detail for the evolution of states.

With this extended hairpin length, little detail can be obtained by only observing each biased CV and subsequent contact map. Each biasing CM is formed of 28 CV-pairs, and plotting these CMs on their own does not distinguish between the branchpoint and hairpin: information on these are provided collectively. However, the biasing CM for hybridising post-slip base-pairs can be split into two new monitoring CVs, one for the branchpoint and the other for the hairpin (See Figure 5). This now provides separated CMs for each, without altering the performance of the applied metadynamics.

The post-processing benefit of this extra CM detailing is apparent in Figure 6. Timestep "A" shows a well slipped state with both the branchpoint and hairpin CMs near their maximum values (9 and 19 respectively). Timestep "B" shows a slight dip in the Branchpoint CM and a more substantial drop in the Hairpin CM, signifying that the branchpoint is opening up only a little, though the hairpin itself has undergone greater denaturing. Timestep "C" shows that the Hairpin CM has returned very close to its minimum value though the Branchpoint CM still has some contribution, this signifies a system state close to the pre-slip (initial) conformation with only the branchpoint in a partially slipped state.

As for the thermodynamic landscape of CAG_{20}, the free-energy heatmaps of Figures 7 and 8 begin to show signs of deviation between the WTMetaD and TTMetaD flavours.

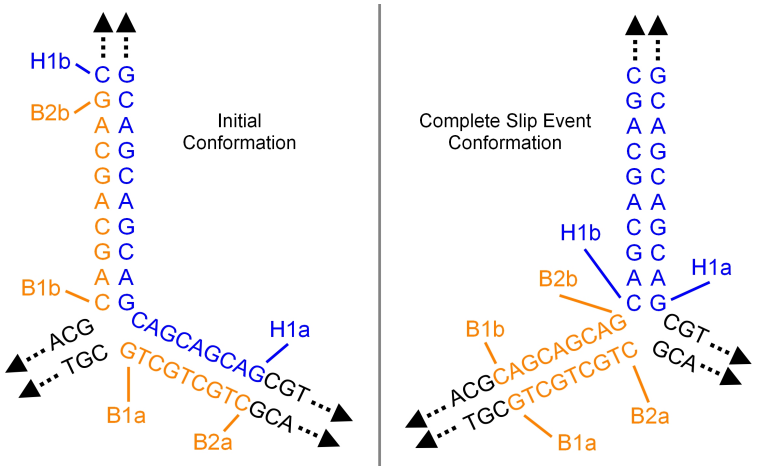

FIG. 5. The Conceptual Definition of Unbiased Monitoring Collective Variables (CV). Nucleotide CV-pairing is illustrated by pairs: B1a/b, B2a/b, and H1a/b. "B" represents "Branchpoint", and "H" the "Hairpin". In the initial conformation (left), there is relatively large spatial separation of CV-pairs and so these CVs result in a negligible switching function output. When the system completes a full slip event under this CV definition (right), the conformation is such that CV-pairs are hybridised and the switching function output is at its maximum. With two distinct unbiased CV definitions, one for the branchpoint (orange) and one for the hairpin itself (blue), fast and reliable information is provided as to whether or not branchpoint migration has occurred *and if* the hairpin has fully migrated also.

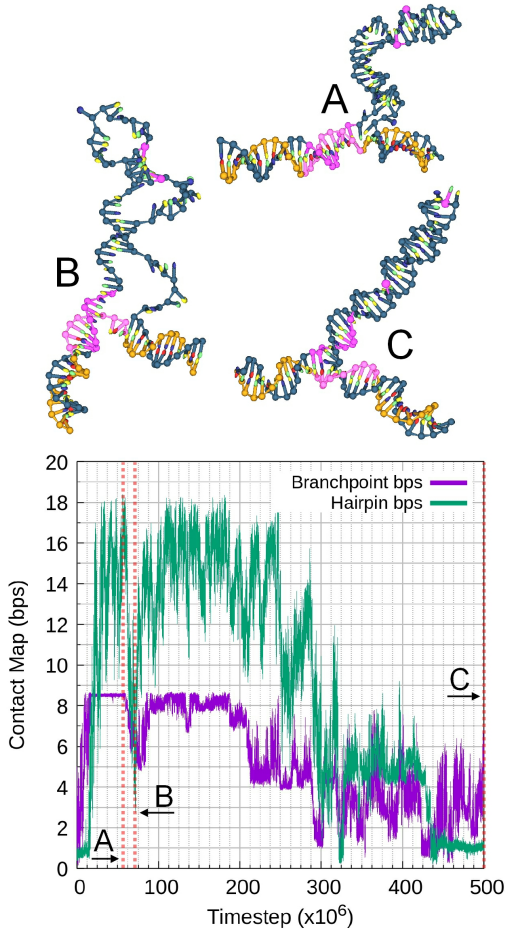

FIG. 6. An Example Utilisation of Contact Maps for Approximating dsDNA-Hairpin Conformation. Some examples of the CV base-pairs from these CMs are highlighted by purple nucleotides within the three visualised dsDNA-hairpin snapshots (top). The three snapshots shown (top) highlight that the branchpoint and hairpin CMs (bottom) provide a good indication of the system's conformational form, with distinct CM value combinations representing various base-pair arrangements.

The average of runs governed by WTMetaD (Figure 7) highlights an underdevelopment of post-slip exploration; the upper right basin is shallower than that of the initial state. Given that initial and fully post-slipped conformations contain the same sequence of nucleotides and the same number of hybridised base-pairs, it is reasonable to expect both wells should have similar free-energies. This suggests these runs have not reached convergence. Two options are readily available here to rectify this problem, though both of which are not without their drawbacks. The obvious solution is to increase the duration of the runs whilst retaining the same WTMetaD parameters, though this brings into question the practical suitability of increased simulation times. The second option is to further increase the bias factor, and whilst from a global perspective this helps reach the post-slipped conforma-

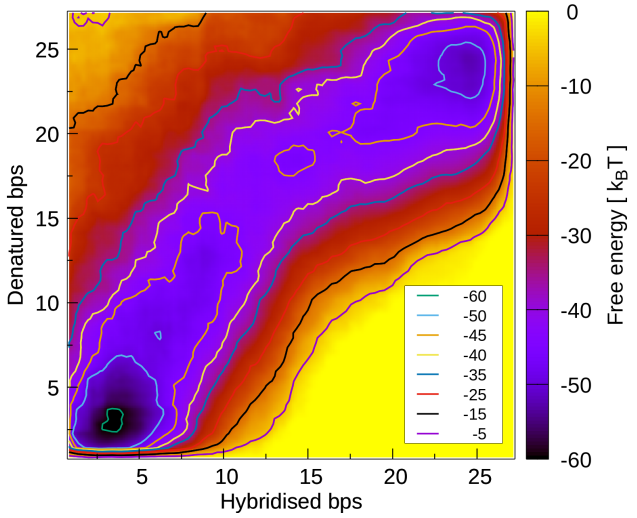

FIG. 7. The Free-Energy Landscape Sampled and Averaged Across WTMetaD Simulations of the CAG_{20} Hairpin. Despite an increased bias factor of 12, the underdeveloped post-slip basin (upper right $-50k_bT$ contour) indicates 500M timesteps falls a little short of the idealistic point of convergence.

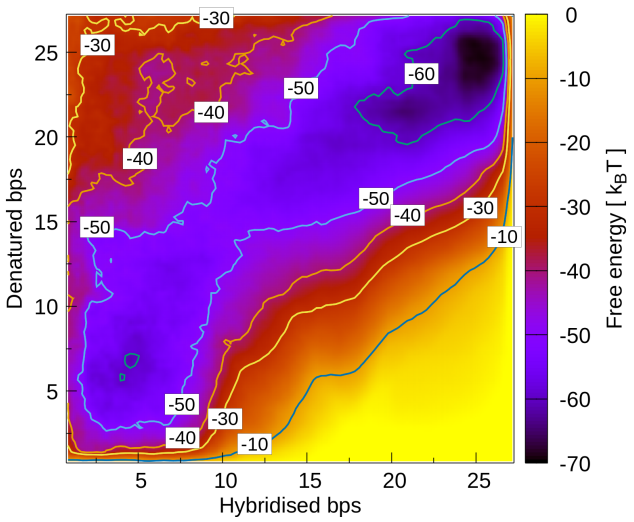

FIG. 8. The Free-Energy Landscape Sampled and Averaged Across TTMetaD Simulations of the CAG_{20} Hairpin. Here, the tempering effects of TTMetaD appear to show sufficient exploration in the post-slip region, however the algorithm is perhaps a little too aggressive in overcoming the energy well of the first initial basin (bottom left $-60k_bT$ green contour).

tional region, it can result in localised overcompensation of hills.

Inversely, the average across CAG_{20} TTMetaD runs (Figure 8) shows a reversed asymmetry in the two aforementioned basins - the post-slip region is well explored, however an over-aggressive exit from the initial state is visible. The upper left-hand corner also appears to have a lower free-energy barrier in comparison to Figure 7 - contour lines have a notably steeper gradient in the latter. The free-energy scales generated in TTMetaD runs also appear to accumulate slightly deeper energy wells over the WTMetaD method. All these factors combined lead to an obscured thermodynamic picture for CAG_{20}, and the free-energy results here should be taken as a general approximation rather than an accurate and conclusive free-energy representation.

New Insights into the Heterogeneity of Larger CAG_{20} Hairpins.– Despite the apparently imprecise CAG_{20} free-energy heatmaps (Figures 7 and 8), some remarkable observations of structural heterogeneity where observed across multiple runs and in both the WTMetaD and TTMetaD test suite. The increased hairpin length unlocks more conformational options to the system and secondary structures emerge within the CAG_{20} hairpin.

It appears, rather unsurprisingly, that the larger CAG_{20} hairpin does not fully denature and re-hybridise in one direct transition - this would imply the path of highest free-energy resistance and lowest thermodynamic probability is taken. Instead, more energetically favourable metastable states are explored. Such conformations can be seen from an exemplar run in Figure 9, with an accompanying CM plot and free-energy heatmap in Figures 10 and 11, respectively. The most dominant secondary hairpin structure to form is one in which the hairpin essentially rearranges into two smaller hairpins of reduced length, illustrated most notably at timesteps 163M, 270M, 390M, and 500M in Figure 9. Similar intramolecular structures have previously been observed in CTG_{12-25} oligomers; Amrane et al refer to these as

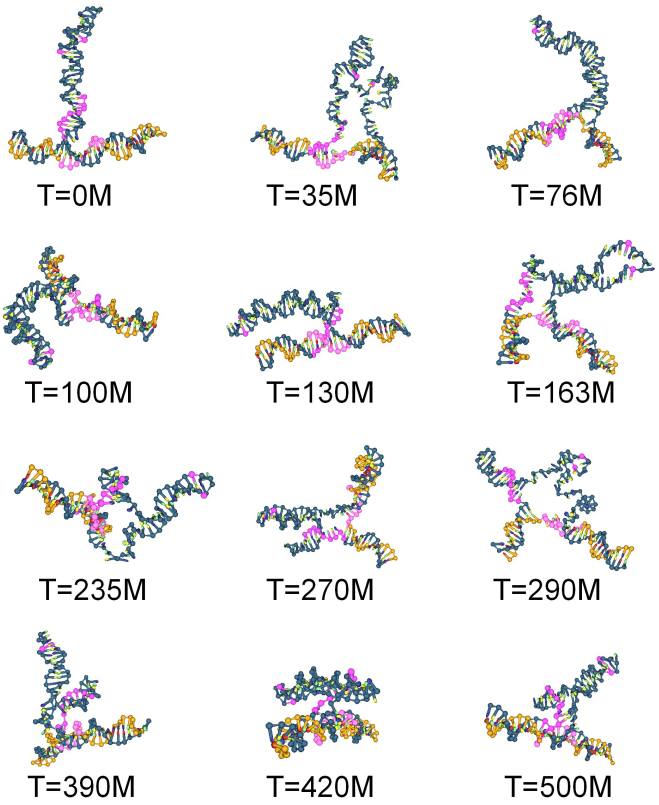

FIG. 9. Visualisation for a Single WTMetaD Thermostat Displaying Heterogeneous CAG_{20} Conformation.

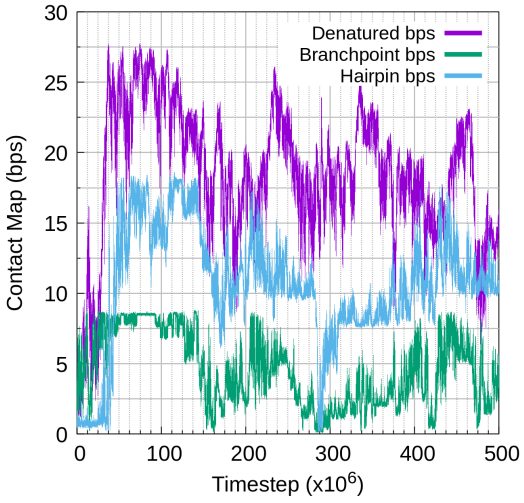

FIG. 10. Contact Maps for a Single WTMetaD Thermostat Displaying Heterogeneous CAG_{20} Conformation.

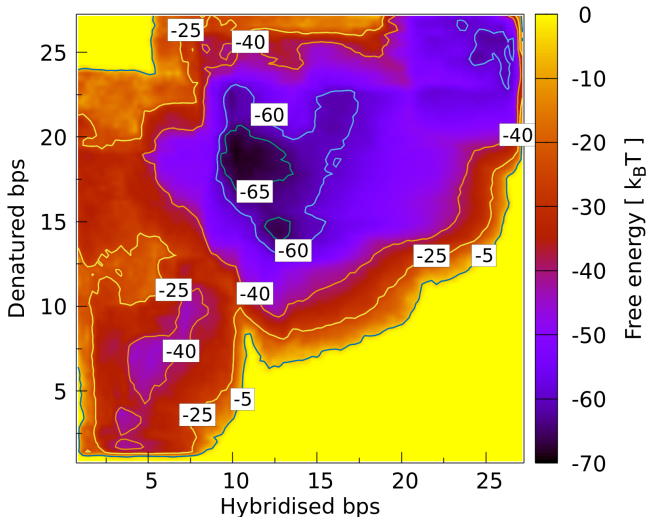

FIG. 11. Free-Energy-Landscape for a Single WTMetaD Thermostat Displaying Heterogeneous CAG_{20} Conformation.

"bis-hairpins" [29]. This structure is analogous to the stacked bulges described for CAG_{10}, although with the increased abundance of nucleotides for CAG_{20}, there are a sufficient number of denatured base-pairs conducive to re-hybridisation of an additional hairpin-like structure.

The results also suggest that the branchpoint and hairpin may be capable of forming an alternative H-like junction, as is evidently present at the T=420M timestep in Figure 9. This conformation corresponds to the CM conditions (Figure 10) of: Hairpin∼11, Branchpoint∼0, and Denature∼17. Interestingly, this CM position is in direct alignment with the significant energy well at (11,17) within Figure 11, and the conspicuous $-70k_bT$ metastability of this state is indicative of the strong hydrogen-bonded pairing evident from the visualisation. The prerequisite dynamics of this H-like junction can also be ascertained - it appears they form through evolution of the preceding split double-hairpin when its branchpoint is in a tightly packed X-like structure (T=390M).

Despite the apparent visual discrepancy between snapshots at T=390M and T=420M, surprisingly little base-pairing rearrangement occurs; only a couple of pairs within the branchpoint denature here. In both of these snapshots, the DNA strand comprising the hairpins takes the same defining form: the strand comes up out from the main dsDNA sequence, curves into the first hairpin, doubles back in direction right along to the end of the second hairpin, doubles back once more, and curves back down to continue into the main dsDNA sequence.

Conclusions and Outlook.– Our metadynamics methodology on CAG_{10} proved successful: system evolution was in agreement with the dynamics observed by Bianco et al [9], and the free-energy landscapes from both WTMetaD and TTMetaD where consistent. TTMetaD was tested on this dsDNA-Hairpin configuration for the first time, and given its superior convergence times whilst producing relatable results to WTMetaD, we suggest TTMetaD should be the metadynamics flavour for choice for such a system. However, the emerging free-energy discrepancies in CAG_{20} modelling suggests the need for improved methodology going forward. Whilst we have demonstrated the existence of structural heterogeneity and branchpoint migration within our CAG_{10} and CAG_{20} hairpin modelling, the next milestone is to achieve effective state sampling with 40 or more trinucleotide repeats ($CAG/CTG_{\geq 40}$). Hairpins below this threshold length are considered to be healthy, with the symptomatic onset of CAG/CTG REDs generally occurring around or above 40 repeats in length [1, 2, 4–6, 9]. Increasing effective state sampling within our modelling is necessary for the future exploration of this symptomatic TNR region; we hereby postulate two available improvements to the modelling.

One notable disadvantage of our current configuration lies within the definition of CVs. Metadynamics capability was introduced to the modelling via PLUMED [27, 28], although the package only allows for the input of centre of mass co-ordinates for its CV definition. The CVs defined within our metadynamics implementation are intended to represent the base-pairing of nucleotides, and so the accurate co-ordinates for this should instead be provided by the hydrogen-bonding sites. Currently, there is no functionality within PLUMED that allows access to this information.

It is therefore possible that PLUMED can falsely interpret hybridisation of nucleotides within particularly close proximity - it cannot ascertain the difference between two nucleotides facing towards or away from one another. As a result, it is likely our implementation of PLUMED is reporting a small number of factually inaccurate CVs under the aforementioned close proximity condition; and in turn, this can result in the premature reduction of hill-heights. With sufficient development time, the source code for the oxDNA model could be modified to create a virtual atom for each of the hydrogen-bonding interaction sites. This virtual atom can then be interpreted within PLUMED by default, and hence allow for the superior CV definition accurately representative of the base-pairing sites.

Additionally, and for a less time-consuming solution to improve state sampling, we suggest that the dsDNA-Hairpin simulations can be ran from a reversed initial state. Referring back to Figure 8, it was observed that TTMetaD was effective in exploring the post-slip conformational region though pre-maturely exited the system's initial state. Given the free-energy symmetry present between the current initial and post-slip states, a complete free-energy landscape can be stitched from two sets of runs: one set containing the set-up described in this Report, and the other taking instead the post-slip conformation as the initial state. Given the observed behaviour

of our TTMetaD CAG_{20} runs, the reversed set-up would begin in the upper right hand corner of the heatmap (Figure 8) and transition into the lower left-hand corner within a short transient period. From here, it would comprehensively explore this region due to its reduction in hill-heights. By combining the data from both sets of runs, a more rigorous exploration of the free-energy landscape can be ascertained; each set would account for underdeveloped regions from the other.

Supplementary Remarks.– Regarding temperatures, Bianco et al [9] conducted their smFRET at a temperature of 20°C± 1°C, and their modelling at a comparable T=300K. All modelling in this Report mimics these temperatures using T=300K.

For software, all modelling was performed using a custom LAMMPS compilation from 7 Jan 2022. This contained pre-released updates to the CG-DNA package, although these are now included in the LAMMPS Stable Release 23 Jun 2022 for reference. Metadynamics capability was introduced to the modelling via the PLUMED library [27, 28], release 2.7.3.

Acknowledgements.– Many thanks is given to my supervisor, Dr. Oliver Henrich, who provided helpful discussion and guidance throughout the complete duration of the project. Particularity when it was unclear how to interpret and proceed with some of the metadynamics configuration, joint consideration on these matters was invaluable. The Department's HPC server was also utilised in this project, operated under the CNQO Group and maintained by Timothy Briggs; this was a necessary resource to run the complete suite of all relevant computational data.

[1] A. N. Khristich and S. M. Mirkin, J. Biol. Chem. Reviews **13**, 4134 (2020).
[2] T. H. D. C. R. Group, Cell **72**, 971 (1993).
[3] S. H. Subramony, Ann Clin Transl Neurol. **4**, 53 (2016).
[4] S. E. Holmes, E. O'Hearn, A. Rosenblatt, *et al.*, Nature Genetics **29**, 377 (2001).
[5] H. T. Orr, M.-y. Chung, S. Banfi, *et al.*, Nature Genetics **4**, 221 (1993).
[6] H. Harley, S. Rundle, W. Reardon, *et al.*, The Lancet **339**, 1125 (1992).
[7] H. Budworth and C. T. McMurray, Methods Mol Biol **1010**, 3 (2013).
[8] C. T. McMurray, Proceedings of the National Academy of Sciences **96**, 1823 (1999).
[9] S. Bianco, T. Hu, O. Henrich, and S. W. Magennis, Biophysical Reports **2**, 100070 (2022).
[10] T. Hu, M. J. Morten, and S. W. Magennis, Nature Communications **12**, https://doi.org/10.1038/s41467-020-20426-3 (2021).
[11] N. J. Veitch, M. Ennis, J. P. McAbney, *et al.*, DNA Repair **6**, 789 (2007).
[12] A. M. Gacy, G. Goellner, N. Juranic, S. Macura, and C. T. McMurray, Cell **81**, 533 (1995).
[13] C. E. Pearson, M. Tam, Y. Wang, S. E. Montgomery, A. C. Dar, J. D. Cleary, and K. Nichol, Nucleic Acids Research **30**, 4534 (2002).
[14] A. A. Polyzos and C. T. McMurray, DNA Repair **56**, 144 (2017).
[15] T. E. Ouldridge, A. A. Louis, and J. P. Doye, J. Chem. Phys. **134**, 10.1063/1.3552946 (2011), arXiv:1009.4480.
[16] B. E. Snodin, F. Randisi, M. Mosayebi, P. Šulc, J. S. Schreck, F. Romano, T. E. Ouldridge, R. Tsukanov, E. Nir, A. A. Louis, and J. P. Doye, J. Chem. Phys. **142**, 1 (2015), arXiv:1504.00821.
[17] H. O, G. F. YA, C. T, and O. TE, The European Physical Journal. E, Soft Matter **41** (2018).
[18] S. A, O. TE, H. O, R. L, and S. P, Frontiers in Molecular Biosciences **8** (2021).
[19] E. Roth, A. Glick Azaria, O. Girshevitz, A. Bitler, and Y. Garini, Nano Letters **18**, 6703 (2018).
[20] P. Gross, N. Laurens, L. B. Oddershede, U. Bockelmann, *et al.*, Nature Physics **7**, 731 (2011).
[21] D. PD, W. J, G. H, and O. M., Current Opinion in Structural Biology **37**, 29 (2016).
[22] A. Barducci, G. Bussi, and M. Parrinello, Phys. Rev. Lett. **100**, 1 (2008), arXiv:0803.3861.
[23] G. Bussi, A. Laio, and P. Tiwary, *Handb. Mater. Model.* (2018) pp. 1–31.
[24] G. Bussi and A. Laio, Nat. Rev. Phys. **2**, 200 (2020).
[25] A. Laio and F. L. Gervasio, Reports Prog. Phys. **71**, 10.1088/0034-4885/71/12/126601 (2008).
[26] J. F. Dama, G. Rotskoff, M. Parrinello, and G. A. Voth, J. Chem. Theory Comput. **10**, 3626 (2014).
[27] B. Massimiliano, B. Giovanni, C. Carlo, T. Gareth, and B. Pavel, Nat Methods **16**, 670 (2019).
[28] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, Computer Physics Communications **185**, 604 (2014).
[29] S. Amrane, B. Saccà, M. Mills, M. Chauhan, H. H. Klump, and J.-L. Mergny, Nucleic Acids Research **33**, 4065 (2005).