Assessing the sequence dependence of pyrimidine-pyrimidone (6-4) photoproduct in a duplex double-stranded DNA: a challenge for microsecond range simulation

Natacha Gillet,[1] Alessio Bartocci,[1] and Elise Dumont[1, 2]

[1] *Univ Lyon, ENS de Lyon, CNRS UMR 5182, Université Claude Bernard Lyon 1, Laboratoire de Chimie, F-69342, Lyon, France*[a]

[2] *Institut Universitaire de France, 5 rue Descartes, 75005 Paris, France*[b]

(Dated: 17 December 2020)

Sequence dependence of the (6-4)photoproduct dynamics when embedded in six 25-bp duplexes is evaluated along extensive unbiased and enhanced (replica exchange with solute tempering, REST2) molecular dynamics simulations. The structural reorganization as the central pyrimidines become covalently tethered is traced back in terms of non-covalent interactions, DNA bending and extrusion of adenines of the opposite strands. The close sequence pattern impacts the conformational landscape around the lesion, inducing a different upstream and downstream flexibilities. Moreover, REST2 simulations allow to probe structures possibly important for damaged DNA recognition.

I. INTRODUCTION

Pyrimidine-pyrimidone (6-4) photoproduct and cyclobutane pyrimidine dimers (64-PP and CPD) are the two main DNA UV-induced lesions, which can lead to mutation with carcinogenic properties[1,2] notably in skin cells, widely exposed to UV radiation. CPD represents a large part of the formed lesion and mutation occurrences[3], but 64-PP is associated to a higher degree of mutagenicity.[4,5] In mammals, both damages can be repaired by the nucleotide excision repair (NER) pathway[6,7], but with contrasted rates. Indeed, 64-PP is more tightly bound than CPD by the UV-damage DNA binding protein (UV-DDB),[8–11] which is known to better recognized bulky damages. From a chemical point of view, 64-PP induces a more important distortion of the DNA stack as the C4-C6 covalent bond implies that the pyrimidine and pyrimidone (PYO) rings are roughly perpendicular whereas the two pyrimidines remain mostly stacked in the CPD damage (see Figure 1). Consequently, some Watson-Crick hydrogen bonds can be lost and rearrangements of the DNA B-helix have to occur to accommodate this lesion, leading to an increased DNA flexibility. For instance, X-Ray and NMR experiments suggest a large bending going up to 44°.[12–14] Moreover, molecular dynamics (MD) simulations over several hundreds of nanoseconds have underlined the conformational polymorphism around 64-PP (compare to the more rigid CPD lesion), with a tendency of PYO or adenines facing the damage to seek opportunistic interactions.[15,16]

This flexibility and the deformation of the B-helix structure can help the repair proteins to recognize the damaged nucleobases and to flip them within the active site, as observed for photolyases[17] or the UV-DDB complex for NER.[18] The XPC-RAD23B-CETN2, also belonging to NER pathway, recognizes 64-PP with a relatively good rate whereas UV-DDB is required for CPD repair .[19,20] By using the yeast ortholog of XPC-Rad23B, namely Rad4-Rad23, Paul *et al.* have recently obtained a crystallographic structure of this DNA-protein complex around 64-PP damage.[21] The recognition mechanism benefits from the damaged DNA flexibility as a DNA kink and an untwisting around the 64-PP lesion are observed in the protein-bound DNA structure. More striking is the flip-out of the adenines toward Rad4, instead of the damage itself.

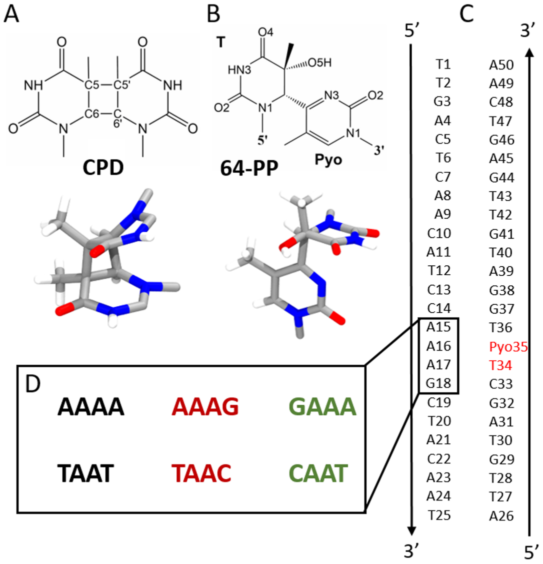
A
B
C
CPD
64-PP
T
Pyo
O4
HN3
O5H
O2
N1
N3
O2
N1
5'
3'
NH
C5
C5'
C6
C6'
O
N
5′
3′
3′
5′
T1 T2 G3 A4 C5 T6 C7 A8 A9 C10 A11 T12 C13 C14 A15 A16 A17 G18 C19 T20 A21 C22 A23 A24 T25
A50 A49 C48 T47 G46 A45 G44 T43 T42 G41 T40 A39 G38 G37 T36 Pyo35 T34 C33 G32 A31 T30 G29 T28 T27 A26
D
AAAA
AAAG
GAAA
TAAT
TAAC
CAAT

FIG. 1. CPD (A) and 64-PP (B) photolesions in chemical and 3-dimensional representation. C: DNA sequence used in this study, similar to the one from ref[21]. D: six variations on the Watson strands of the flanking pairs of the damage.

Consequently, for all repair pathways, the ability of the photolesion-containing DNA to overcome the barrier of base pair opening represents a crucial issue. Free energies calculations and MD simulations have been performed to characterize the transition from the intra-helical to the extra-helical conformation of nucleobases in regular DNA,[22–24] mismatch opening[25–27] or damaged DNA.[28–36] They range from 6 to 15

[a] Electronic mail: natacha.gillet@ens-lyon.fr

[b] Electronic mail: elise.dumont@ens-lyon.fr

kcal/mol, which is too high for a spontaneous flip-out, but experiments suggests a millisecond base pair opening.[37,38] Besides, the sequence impacts the formation, the conformational behavior and the repair of UV-photolesion.[39] For CPD, O'Neil *et al.*[40] have underlined the importance of the nature of the flanking bases around the damage in the energy range required to open DNA. The DNA flexibility is higher when A-T pairs flank the thymine dimer instead of G-C pairs, which are more rigid thanks to an additional hydrogen bond. In undamaged B-DNA, the exhaustive study of tetranucleotide sequences by the ABC consortium[41,42] has also demonstrated that the importance of the nearest-neighbors and next-nearest-neighbors, that must not be neglected and can influence the DNA conformational space. Besides, Lindahl *et al.* have reported for undamaged base pair[24] a sequence dependence of the opening preference and the associated free energy. Such questions about the flip-out mechanism and its sequence dependence also arise for 64-PP; it is even more crucial because of the high lesion flexibility and the propensity of the damaged thymines or facing adenines to look for opportunistic non-covalent interactions. Ma and van der Vaart have presented a two-dimensional free energy surface of 64-PP flipping out for one sequence only, using bending as second reaction coordinate.[33] The flat surface with respect to bend angle highlights the high DNA flexibility around this lesion; however, no basin is visible for the flipped out conformation, even though the opening of the damage appears relatively easy.

The previous studies also confirmed the relevance of MD simulations in the description of the conformational space in normal and damaged DNA, with a specific care brought to dedicated force fields. The dramatic enhancement of computational power, thanks to the advent of GPU implementation, makes microsecond timescales simulations almost routine calculations and support a wider exploration of the biomolecules conformational space. Nevertheless, the free energy surface of the damaged DNA is complex and even microsecond timescale simulations are not sufficient to capture rare events. Biased methods, such as umbrella sampling, represent an efficient alternative to simulate these phenomena of interest but require a preliminary knowledge of the relevant collective variables to describe the considered mechanism. The relative rigidity of CPD lesion allows the use of a pseudo dihedral angle ("extrusion angle"[31]) as a reaction coordinate for the flip-out mechanism, involving the center of mass of the opposite adenines, farther nucleobases, the backbone of the damage and the damage itself. The conformational polymorphism around 64-PP lesions disqualifies such approach as damaged DNA can undergo large distortions. Then, unwanted motions can be captured along the reaction coordinate while the damage or opposite adenines flipping-out is not observed. Another solution lies in the enhancement of the conformational sampling using replica exchange molecular dynamics approach (REMD). Traditional Temperature REMD (T-REMD)[43] consists in running at different temperatures simultaneous and independent MD simulations which periodically attempt an exchange, which becomes effective if the energy cost from one conformation to the other is negligible. The computational cost to reach a number of replica allowing a exchange rate higher than 20 % can make such an approach computationally prohibitive. Consequently, other REMD protocols have been developed where only the molecule of interest is heated and thus the number of required replica decreases. Among them, the REST2 approach (replica exchange with solute tempering 2)[44] has given satisfactory results for inhibitor-protein interaction,[45], protein folding or stability[46,47], peptides in lipid bilayers[48], carbohydrates conformation[49] or RNA folding and force field optimization.[50–52]

In this study, we propose to sample the conformational space of the 64-PP lesion within 25-bp oligonucleotides, differing by the inner motif defined by six different tetranucleotide sequences : these biomolecules were simulated using unbiased microsecond range MD or REST2 approach. The 25-bp oligonucleotide sequence corresponds to the DNA molecule used in the RAD4/XPC-DNA complex (Figure 1): the damage is situated on the Crick strand and the tetranucleotide sequence on the Watson strand is AAAG. To give a first insight into the sequence impact on the 64-PP polymorphism, we inverted the flanking pairs (TAAC) sequence, or mutated them to obtain similar systems as the one we had previously studied, and then also proceed to the flanking pair inversion: TAAT and AAAA,[15] CAAT and GAAA.[16] The conformational samples provided by unbaised MD or REST2 REMD are deeply analysed with respect to the sequence dependence of the DNA behavior.

II. METHODOLOGY

All the DNA sequences were built using the NAB module of the Amber18 package[53] and solvated in a parallelepiped ($60 \times 66 \times 115$ Å^3) box of TIP3P water molecules.[54] 48 potassium cations were added to neutralize the system. All simulations were performed using the Amber18 MD package[53], the parmbsc1[55] force field for the standard nucleoases. The 64-PP lesion was parameterized as described in an earlier work.[15]

The systems were initially minimized for 10000 steps (5000 of steepest descent and 5000 of conjugated gradient). Then, the systems were heated gradually from $0\,K$ to $300\,K$ by using a restraint force constant of 10 kcal/mol/Å^2 around the region of the damage (bases 15 to 18 and 33 to 36). For both systems, the integration time step t_{step} was fixed to 1.0 fs for a total of 30 ps (NVT), and constant temperature was achieved using a Langevin thermostat ($\gamma_{coll} = 1\,ps^{-1}$). An equilibration step of 1 ns was performed in the NPT ensemble using the Langevin thermostat and the Berendsen barostat ($P = 1\,atm$ and $T = 300K$). Finally, 1 μs production simulations were run in the NPT ensemble. For all simulations, a timestep of 2 fs was used, covalent bonds involving a hydrogen were constrained using the SHAKE algorithm, the long-range electrostatic interactions were computed using Particle Mesh Ewald (PME) algorithm[56] with a cutoff of 10 Å for the van der Waals and the real space of the electrostatic interaction.

We used the last structure of the production run as starting point for the 1 μs REST2 simulations in the NVT ensemble using Langevin thermostat at 300 K. The solute consists of the whole oligonucleotide (25 base pairs). The λ values were chosen to range the effective solute temperature (T/λ) between 300 and 500 K. 32 replica were required to ensure an exchange rate higher than 25 %. Exchanges were attempted every 200 fs. The same parameters were used for the 500 ns REST2 trajectories when the solute is limited to the damaged four base pairs. The reduction of the size of the solute allows to involve less replica, going from 32 to 12.

The trajectory analysis was performed on 1 μs for all systems, excluding the first 100 ns for REST2 simulations. Hydrogen bond lengths, distances, angles and RMSF were computed using the cpptraj module of Amber, as well as the clustering analysis. To determine the frequency of hydrogen bonds, we selected the geometries where the distance between hydrogen and proton acceptor is lower than 3.0 Å and the donor-hydrogen-acceptor angle is higher than 135°. The DBSCAN method[57] was employed for the cluster analysis with a minimum number of points to form a cluster equal to 50, and a distance cutoff between points for forming a cluster of 1.3 Å. The RMSD of the nucleobases 15 to 18 and 33 to 36 has been chosen for the clustering. The bend angle of the DNA segment from A11 to C22 (G29 to T40) was calculated using Curves+.[58]

The uncertainties on the distance and bend values were estimated by dividing the corresponding standard deviation by the square root of the statistical sample efficiency ρ multiply by the number of steps. These values where calculating using the CODA library[59] from the R environment.[60]

III. RESULTS

The different conformational spaces visited during the unbiased (MD) or enhanced sampling (REST2) simulations for the different sequences have been firstly condensed by a cluster analysis, focusing on the different geometries of the four base pairs including the damage and its two neighbors (*i.e.* bases 15 to 18 and 33 to 36). The representative conformation of each cluster containing more than 10% of the total trajectory frames are shown in Figure 2 for AAAG and TAAC sequences and Figure S1 and S2 in ESI for the other sequences.

During the unbiased MD simulations, only one conformation is encountered for the four base pairs: it consists in a geometry close to the Watson-Crick arrangement, with conserved hydrogen bonds between the complementary strands, including the ones between adenines 16-17 and the damaged thymines. In addition, the PYO group remains stacked to base 36 while T34 becomes almost perpendicular to the complementary adenines planes. Then, the oxygen O4 of T34 can interact with the N6 amine group of A16. This conformation is common to all sequences but we can notice that another hydrogen bond is formed between the alcohol moiety of T34 and an ketone oxygen of a guanine or a thymine in position 15, so for GAAA, TAAT and TAAC sequences (see also Table I). No additional interactions are observed between the damage and A15 (AAAA and AAAG sequences) but the formation of a hydrogen bond with Gua36 O6 is also occurring for the CAAT sequence. The frequencies of the 34-15 or 34-36 bonds in the unbiased MD trajectories (see Table I) support the idea of a strong interaction that can contribute to maintain this Watson-Crick like conformation. Indeed, a look to the N1-N3 distances between the nucleobases 15-36 (Table II) show a closer and less flexible conformation for GAAA, TAAT, TAAC and CAAT than for AAAA and, in a more dramatic way, AAAG. For the latter, an average distance of 4.54 Å, with a relatively high standard deviation (2.19 Å) suggests that the damage impairs the stability of the Watson-Crick hydrogen bonds between A15 and T36.

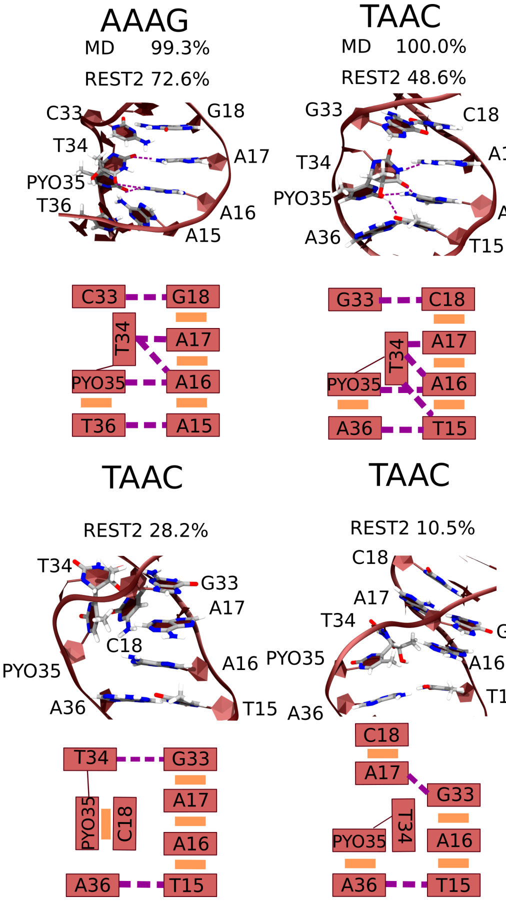

FIG. 2. Cartoon and schematic representations of the centroid conformations of the clusters from the MD and REST2 1 μs trajectories for the AAAG and TAAC sequences. Hydrogen bonds are represented by dashed purple lines and π-stacking by orange rectangles. The conformations for the other sequences are given in Figure S1 and S2 in ESI. Only the representative geometry of the clusters containing more than 10% of the total frames are drawn. All DNA pictures were rendered with vmd.[61]

The clustering of the REST2 trajectories for the six sequences attributes the same conformation to the main cluster. This cluster contains from 30% (AAAA) to 72.5% (AAAG) of the total trajectory snapshots. The frequency of the T34 nucleobase 15 hydrogen bond, with respect to the representative percentage of this cluster, is similar as the unbiased MD frequency (Table I). Except for the AAAG sequence, the clustering analysis describes another structure where PYO35 is perpendicular to the A16 plane, as well as the base in position 18, which comes in a π-stack interaction with the pyrimidone

TABLE I. Frequencies of the hydrogen bond between the alcohol group of T34 and the base in position 15 (GAAA, TAAT or TAAC sequence) or 36 (CAAT sequence) during the 1 μs unbiased MD simulations (MD) or the replica exchange REST2 simulation at 300 K (REST2).

	GAAA	TAAT	TAAC	CAAT
MD	45.6%	77.2%	77.1%	17.7%
REST2	23.0%	39.4%	37.3%	14.3%

TABLE II. Averaged distance (in Å) between the purine N1 - pyrimidine N3 atoms of the 15-36 and 18-33 base pairs during the 1 μs unbiased MD simulations (MD) or the replica exchange REST2 simulation at 300 K (REST2).

		AAAA	AAAG	GAAA	TAAT	TAAC	CAAT
MD	15-36	3.18	4.54	2.94	2.97	2.97	3.00
		± 0.27	± 0.83	± 0.00	± 0.00	± 0.01	± 0.01
	18-33	2.97	2.95	2.947	2.98	2.96	2.99
		± 0.00	± 0.00	± 0.01	± 0.01	± 0.00	± 0.01
REST2	15-36	4.37	5.34	3.78	3.13	3.66	3.07
		± 0.06	± 0.07	± 0.04	± 0.01	± 0.05	± 0.01
	18-33	6.12	3.18	4.83	5.55	4.87	6.32
		± 0.12	± 0.03	± 0.08	± 0.05	± 0.06	± 0.13

TABLE III. Averaged bend angle (in °) for the DNA segment between 11-22 and 29-40 nucleobases during the 1 μs unbiased MD simulations (MD) or the replica exchange REST2 simulation at 300K (REST2).

	AAAA	AAAG	GAAA	TAAT	TAAC	CAAT
MD	24.25	34.20	26.06	25.41	23.49	28.04
	± 1.80	± 0.60	± 0.91	± 0.51	± 0.62	± 2.13
REST2	30.06	36.03	30.27	28.01	24.73 3	23.99
	± 1.16	± 0.90	± 0.37	± 0.45	± 0.83	± 0.54

ring. This important deformation also leads to the stacking of the base in position 33 over A17 (for TAAC) or A16 (for AAAA, GAAA, TAAT and CAAT). In this geometry, a hydrogen bond can be formed between T34 and the nucleobase 33, which become coplanar. The position of A17 depends on the sequence: it stays in the normal strand stacking for the TAAC sequence but twists to lie in the PYO35-18 alignment for AAAA, GAAA,TAAT and CAAT. This conformation represents between 15.8% (CAAT) to 29.8% (AAAA) of the 1μs REST2 trajectory. Consequently, as reported in Table II, the 18-33 N1-N3 distance for these sequences increases, with several peaks in the distribution profile (see Figure S3 ESI), while it remains short for the AAAG sequence.

The analysis of GAAA, TAAT, TAAC and CAAT trajectories describes a third cluster comprising up to 19.5 % (CAAT) of the conformations. For the three sequences with pyrimidines flanking the A16-A17, this third cluster is characterized by a 15-16-33 π-stacking whereas the A16-A17 distance is elongated to 6.16 Å (TAAC), 8.78 Å (TAAT) and 10.73 Å (CAAT). A17 and the nucleobase in position 18 can stay unpaired or form a new mispaired interaction creating an orphan base, few nucleobases further. The GAAA third conformation is a highly bent geometry where the damage strand form a nearly 90° angle between C36 and PYO35. A17 is in a π-stacked interaction with the pyrimidone group, while the T33-A18 interactions are conserved.

The cluster analysis thus draws a quite similar conformational landscape for the pyrimidine-A16-A17-pyrimidine sequences (TAAT, TAAC and CAAT). The presence of a stronger C-G base pair in 5' or 3' of the damage does not seem to have a strong influence on this panorama. AAAA and GAAA sequences also present similarities with the flanking pyrimidine sequences for the two main clusters whereas clustering analysis of AAAG emphasizes only one conformation, the same as in the unbiased MD. These first results may lead to conclude that AAAG is less flexible than the other sequences. However, a look to the average distance within the clusters reveals that the main cluster of AAAG is wider than the other ones, with an averaged distance between points within the cluster of 2.6 Å, significantly higher than for the other sequences (0.6 to 1.3 Å wider). As observed in unbiased MD, the A15-T36 interaction suffers from the presence of the damage more than in other sequences, despite the lack of highly distorted conformations. While G18 and C33 stay close, the wider main cluster of AAAG can be explained by the larger fluctuations of the upstream nucleobases (see also RMSF in Figure S7.).

The 64PP repair by enzymes requires the flip out of the photodamage from DNA to the protein active site.[17,18] On the opposite, Rad4 releases the damage by capturing the complementary adenines in a extrahelical conformation.[21] In any case, the constraint applied to the DNA strands leads to a relatively high bending, which is nevertheless energetically affordable due to the presence of the 64-PP damage.[33] In agreement with the previous studies,[15,33] we observe an average bend angle for the damaged DNA between 24 to 36°(see Table III and figure S5), with standard deviation ranging from 10 to 20° and very high maxima, beyond 90°. One can notice that the bend angles are quite similar for the MD as for the REST2 simulation, suggesting that the different conformational spaces explored by the damaged area does not impact the bend. Indeed, we observe no correlation between the cluster belonging and the bend angle value. The AAAG bend stands out from the other sequences with the only averaged value above 30°. While it fluctuates around one representative conformation, this sequence seems more prone to adopt a suitable geometry for repair enzymes-DNA interactions.

The REST2 approach allowed us to enhance our conformational sampling of the damaged DNA, but it is rather GPU demanding (32 GPU in parallel), and this demand increase with the size of the re-scaled solute. As we only want to focus on the four base pairs comprising the 64-PP damage, one can imagine to select only these eight nucleobases within the "solute" part. Such splitting only requests 12 GPU cores to reach an exchange rate between replica of at least 25 %. Nevertheless, it also raises the question of the frontier between a hot and a cold DNA within the same strand. We analyse 500 ns of REST2 simulations of each sequence using this small defini-

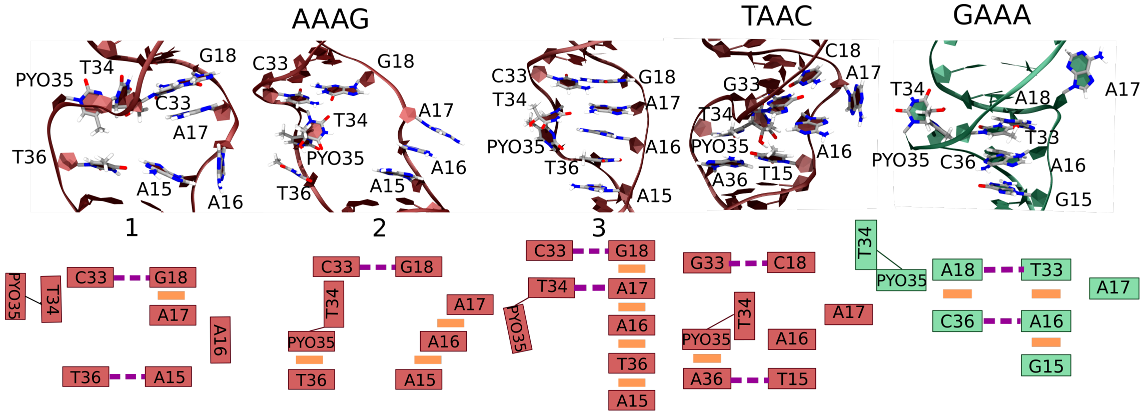

FIG. 3. Cartoon and schematic representation of the main conformations of the 100 ns MD trajectories starting from extrahelical geometries for sequence AAAG **1**, **2** and **3**, TAAC and GAAA. Hydrogen bonds are represented by dashed purple lines and π-stacking by orange rectangles.

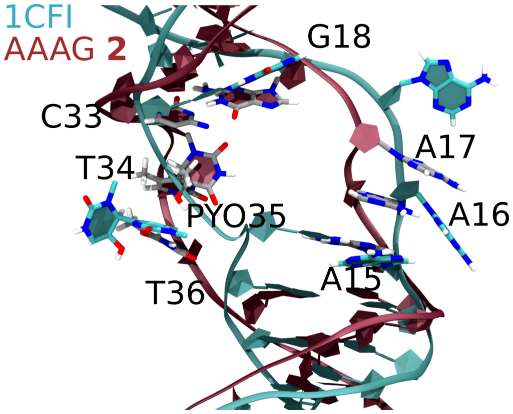

FIG. 4. Cartoon representation of the overlap of the AAAG **2** (red and carbon atoms in gray) conformation and crystallographic structure of damaged DNA in interaction with Rad4 (PDB 6CFI,[21] cyan).

tion of the solute. The clustering analysis (Figure S8 and S9) provides similar results as REST2 simulations with the whole DNA as solute for AAAG and TAAC sequence. For all sequence, the first or the second cluster corresponds to the conformation explored in unbiased MD. The previously described second cluster is found for AAAA, GAAA and TAAC and additional conformations are characterized for AAAA, GAAA and CAAT with relatively high distortions of the DNA strands. Such deformations can be due to the mix of a "hot" tetranucleotide region in a "cold" DNA and justify the use of the full DNA as a solute, despite the encouraging similarities between the two REST2 cluster analysis.

Extrahelicity can also occur in the damage neighborhood without the help of a protein. A common parameter to detect extrahelical conformation is the C1'-C1' distance between the complementary nucleobases. For the B-DNA, this distance fluctuates around 11 Å; above 14 Å, one can consider extrahelicity.[62] However, in our MD simulations, the C1'-C1' distances display from 6.61 Å (for A16-PYO35 in AAAA, AAAG and GAAA sequence) to 16.47 Å (for A16-PYO35 in AAAG) values. In REST2 simulations, this distance range increases and we observe extrema values of 3.39 to 21.0 Å for CAAT 17A-34T distance (see also Figure S6). An extrahelical conformation is not necessarily associated to the largest distance which can be due to the Watson-Crick pairing rearrangement. Nevertheless, some conformations encountered during the REST2 simulations correspond to an extrahelical conformation of A16, A17 or the damage. These events are rare in the 300 K trajectory as these conformations easily exchange with the others replica. Consequently, the relevance of such conformations at 300 K can be questioned. MD simulations of 100 ns were performed on some of these extrahelical conformations. The examples provided in Figure 3 for the AAAG, TAAC and GAAA sequences keep the extrahelicity during at least 70 ns. For AAAG, the first geometry, **1**, concerns the A16 extrahelicity; the second one, **2**, presents extrahelical A15, A16 and A17 and an opening of the damage; the last one, **3**, has a partial extrahelicity of the damage and an unusual stack of the upstream nucleobases. For TAAC and GAAA, a relatively compact geometry is obtained with an extrahelical A17. In addition, in GAAA structure, the damage is in a partially extrahelical conformation. Interestingly, the averaged C1'-C1' distance is greater than 14 Å only for A16-PYO35 in AAAG **1** and A17-T34 in GAAA. The averaged bend angle is very high, around 47 °for AAAG **1**, **2** and TAAC while it falls to 20 °for AAAG **3**; for GAAA, it fluctuates around 35°. The conformation of AAAG **2** presents interesting similarities with the Rad4 bound damaged DNA, as shown in Figure 4, suggesting that the exploration of from REST2 trajectories can provide some meaningful extrahelical conformations.

This large variety of structures complicates the definition of a common parameter to measure the extrahelicity of the 64-PP damage and the complementary adenines. Likewise, a collective variable that described the flip-out mechanism seems hard to describe due to the large deformations applied on the strands in the extrahelical states. Moreover, a sequence effect must not be neglected: in our AAAG simulation, the G18-C33 pair is maintained close while the A15-T36 interaction is lost, in agreement with our previous observation of AAAG dynamical behavior; in TAAC, the proximity between T34 and T15 is maintained, even though the hydrogen bond involving the T34 alcohol group has disappeared; in GAAA, the upstream G15-C36 pair is disrupted, which comes with a large rearrangement of pairing (see Figure 3). The different observed extrahelical structures are consequently related to the previously described conformational specificities of each tetranucleotide sequence. Our examples are far from exhaustive but they suggested that one should consider several flip-out mechanisms with regard to the sequence for 64-PP.

IV. CONCLUSIONS

In this study, we have explored the conformational landscape of 25-base pairs long oligonucleotide featuring a 64-PP photolesion, using REST2 replica exchange, with a focus on the sequence context around the lesion. While the system is trapped within one conformation during unbiased 1 μs MD, our enhanced exploration on the same timescale using REST2 provide a larger range of visited structures including rare events such as opening of the damage. The comparison between six tetranucleotide sequences including the lesion and the two flanking pairs suggests a strong dependence of the dynamical behavior of the damaged double strands with respect to the sequences. When purines flank the damaged thymines (sequences called TAAT, TAAC, CAAT), the conformational sampling provided by REST2 simulations presents strong similarities, with a relatively rigid conformation upstream the lesion, favored by a hydrogen bond between these flanking nucleobases (in case of T, C or G) and the alcohol group of the damaged thymine, and more flexibility downstream, with even a 90 °rotation of the adenine flanking nucleobase. The AAAG sequence shows a unique behavior with a high upstream flexibility and a strong downstream G-C pair. The REST2 simulations of this sequence also provide some extrahelical conformations which stay open during tens of nanoseconds, but the opening mechanism remains obscure and would also be sequence dependent.

This study is a first step towards a better computationally driven rationalization of the sequence dependence of the damaged DNA featuring a 64-PP lesion. Enhanced sampling with REST2 provides a sufficiently large sampling without a prohibitive computational cost to observe difference between the considered tetranucleotides whereas unbiased MD simulations visit only one conformation which can highly depend on the starting point. Other sequences have to be tested to enlarge our systematic assessment of the long range structural impact of the 64-PP lesion. This work also represents a keystone to consider damaged DNA within a nucleosome core particle, in order to evaluate the sequence dependence *versus* the structural constraints and the interactions with histone core and tails.

ACKNOWLEDGMENTS

A.B. is grateful for a post-doctoral fellowship by ENS de Lyon ("Fonds de recherche"). The authors thank the SYSPROD project and AXELERA Pôle de Compétitivité for financial support (PSMN Data Center). This work was performed using HPC resources from GENCI-IDRIS (Grant 2020-A0070800609). This work was performed within the framework of the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon. We thank P. Banas for providing us his REST2 code.

[1] J. Cadet and T. Douki, “Formation of UV-induced DNA damage contributing to skin cancer development,” Photochem. Photobiol. Sci. **17**, 1816–1841 (2018).

[2] G. P. Pfeifer and A. Besaratinia, “UV wavelength-dependent DNA damage and human non-melanoma and melanoma skin cancer,” Photochem. Photobiol. Sci. **11**, 90–97 (2012).

[3] Y.-H. You, D.-H. Lee, J.-H. Yoon, S. Nakajima, A. Yasui, and G. P. Pfeifer, “Cyclobutane pyrimidine dimers are responsible for the vast majority of mutations induced by UVB irradiation in mammalian cells,” Journal of Biological Chemistry **276**, 44688–44694 (2001).

[4] M. J. Horsfall and C. W. Lawrence, “Accuracy of replication past the T-C (6-4) adduct,” Journal of Molecular Biology **235**, 465 – 471 (1994).

[5] J. E. LeClerc, A. Borden, and C. W. Lawrence, “The thymine-thymine pyrimidine-pyrimidone(6-4) ultraviolet light photoproduct is highly mutagenic and specifically induces 3’ thymine-to-cytosine transitions in escherichia coli.” Proceedings of the National Academy of Sciences **88**, 9685–9689 (1991).

[6] K. Sugasawa, “Molecular mechanisms of DNA damage recognition for mammalian nucleotide excision repair,” DNA Repair **44**, 110 – 117 (2016).

[7] C. F. Errol, C. W. Graham, S. Wolfram, D. W. Richard, A. S. Roger, and E. Tom, *DNA Repair and Mutagenesis, Second Edition* (American Society of Microbiology, 2006).

[8] D. K. Treiber, Z. Chen, and J. M. Essigmann, “An ultraviolet light-damaged DNA recognition protein absent in xeroderma pigmentosum group E cells binds selectively to pyrimidine (6–4) pyrimidone photoproducts,” Nucleic Acids Research **20**, 5805–5810 (1992).

[9] J. T. Reardon, A. F. Nichols, S. Keeney, C. A. Smith, J. S. Taylor, S. Linn, and A. Sancar, “Comparative analysis of binding of human damaged dna-binding protein (XPE) and escherichia coli damage recognition protein (UvrA) to the major ultraviolet photoproducts: T[c,s]T, T[t,s]T, T[6-4]T, and T[dewar]T.” Journal of Biological Chemistry **268**, 21301–8 (1993).

[10] Y. Fujiwara, C. Masutani, T. Mizukoshi, J. Kondo, F. Hanaoka, and S. Iwai, “Characterization of DNA recognition by the human UV-damaged DNA-binding protein,” Journal of Biological Chemistry **274**, 20027–20033 (1999).

[11] B. Wittschieben, S. Iwai, and R. D. Wood, “DDB1-DDB2 (Xeroderma Pigmentosum group e) protein complex recognizes a cyclobutane pyrimidine dimer, mismatches, apurinic/apyrimidinic sites, and compound lesions in DNA,” Journal of Biological Chemistry **280**, 39982–39989 (2005).

[12] J.-K. Kim, D. Patel, and B.-S. Choi, “Contrasting structural impacts induced by cis-syn cyclobutane dimer and (6–4) adduct in DNA duplex decamers: Implication in mutagenesis and repair activity,” Photochemistry and Photobiology **62**, 44–50 (1995).

[13] J.-K. Kim and B.-S. Choi, “The solution structure of DNA duplex-decamer containing the (6-4) photoproduct of thymidylyl(3$\rightarrow$5)thymidine by NMR and relaxation matrix refinement,” European Journal of Biochemistry **228**, 849–854 (1995).

[14] J.-H. Lee, G.-S. Hwang, and B.-S. Choi, “Solution structure of a DNA decamer duplex containing the stable 3’ T.G base pair of the pyrimidine(6–4)pyrimidone photoproduct [(6–4) adduct]: Implications for the highly specific 3’T $\rightarrow$ C transition of the (6–4) adduct,” Proceedings of the National Academy of Sciences **96**, 6632–6636 (1999).

[15] F. Dehez, H. Gattuso, E. Bignon, C. Morell, E. Dumont, and A. Monari, “Conformational polymorphism or structural invariance in DNA photoinduced lesions: implications for repair rates,” Nucleic Acids Research **45**, 3654–3662 (2017).

[16] E. Matoušková, E. Bignon, V. E. P. Claerbout, T. Dršata, N. Gillet, A. Monari, E. Dumont, and F. Lankaš, “Impact of the nucleosome histone core on the structure and dynamics of DNA-containing pyrimidine–pyrimidone (6–4) photoproduct,” Journal of Chemical Theory and Computation **16**, 5972–5981 (2020).

[17] J. Li, Z. Liu, C. Tan, X. Guo, L. Wang, A. Sancar, and D. Zhong, “Dynamics and mechanism of repair of ultraviolet-induced (6-4) photoproduct by photolyase,” Nature **466**, 887–890 (2010).

[18] A. Scrima, R. Koníčková, B. K. Czyzewski, Y. Kawasaki, P. D. Jeffrey, R. Groisman, Y. Nakatani, S. Iwai, N. P. Pavletich, and N. H. Thomä, “Structural basis of UV DNA-Damage recognition by the DDB1–DDB2 complex,” Cell **135**, 1213–1223 (2008).

[19] D. Batty, V. Rapic’-Otrin, A. S. Levine, and R. D. Wood, “Stable binding of human XPC complex to irradiated DNA confers strong discrimination for damaged sites,” Journal of Molecular Biology **300**, 275 – 290 (2000).

[20] T. Hey, G. Lipps, K. Sugasawa, S. Iwai, F. Hanaoka, and G. Krauss, “The XPCHR23B complex displays high affinity and specificity for damaged DNA in a true-equilibrium fluorescence assay,” Biochemistry **41**, 6583–6587 (2002).

[21] D. Paul, H. Mu, H. Zhao, O. Ouerfelli, P. D. Jeffrey, S. Broyde, and J.-H. Min, “Structure and mechanism of pyrimidine–pyrimidone (6-4) photoproduct recognition by the Rad4/XPC nucleotide excision repair complex,” Nucleic Acids Research **47**, 6015–6028 (2019).

[22] N. K. Banavali and A. D. MacKerell, “Free energy and structural pathways of base flipping in a DNA GCGC containing sequence,” Journal of Molecular Biology **319**, 141 – 160 (2002).

[23] E. Giudice, P. Várnai, and R. Lavery, “Base pair opening within B-DNA: free energy pathways for GC and AT pairs from umbrella sampling simulations,” Nucleic Acids Research **31**, 1434–1443 (2003).

[24] V. Lindahl, A. Villa, and B. Hess, “Sequence dependency of canonical base pair opening in the DNA double helix,” PLOS Computational Biology **13**, e1005463 (2017).

[25] P. Várnai, M. Canalia, and J.-L. Leroy, “Opening mechanism of G·T/U pairs in DNA and RNA duplexes: A combined study of imino proton exchange and molecular dynamics simulation,” Journal of the American Chemical Society **126**, 14659–14667 (2004).

[26] S. M. Law and M. Feig, “Base-flipping mechanism in postmismatch recognition by MutS,” Biophysical Journal **101**, 2223–2231 (2011).

[27] L.-T. Da and J. Yu, “Base-flipping dynamics from an intrahelical to an extrahelical state exerted by thymine DNA glycosylase during DNA repair process,” , 16.

[28] N. Huang, N. K. Banavali, and A. D. MacKerell, “Protein-facilitated base flipping in DNA by cytosine-5-methyltransferase,” Proceedings of the National Academy of Sciences **100**, 68–73 (2003).

[29] S. D. Bruner, D. P. G. Norman, and G. L. Verdine, “Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA,” Nature **403**, 859–866 (2000).

[30] K. Nam, G. L. Verdine, and M. Karplus, “Analysis of an anomalous mutant of MutM DNA Glycosylase leads to new insights into the catalytic mechanism,” Journal of the American Chemical Society **131**, 18208–18209 (2009).

[31] A. Knips and M. Zacharias, “Both DNA global deformation and repair enzyme contacts mediate flipping of thymine dimer damage,” Scientific Reports **7**, 41324 (2017).

[32] Y. Qi, M. C. Spong, K. Nam, A. Banerjee, S. Jiralerspong, M. Karplus, and G. L. Verdine, “Encounter and extrusion of an intrahelical lesion by a DNA repair enzyme,” Nature **462**, 762–766 (2009).

[33] N. Ma and A. van der Vaart, “Free energy coupling between DNA bending and base flipping,” Journal of Chemical Information and Modeling **57**, 2020–2026 (2017).

[34] L. L. O’Neil, A. Grossfield, and O. Wiest, “Base flipping of the thymine dimer in duplex DNA,” The Journal of Physical Chemistry B **111**, 11843–11849 (2007).

[35] K. Song, A. J. Campbell, C. Bergonzo, C. de los Santos, A. P. Grollman, and C. Simmerling, “An improved reaction coordinate for nucleic acid base flipping studies,” Journal of Chemical Theory and Computation **5**, 3105–3113 (2009).

[36] C. Bergonzo, A. J. Campbell, C. de los Santos, A. P. Grollman, and C. Simmerling, “Energetic preference of 8-oxoG eversion pathways in a DNA Glycosylase,” Journal of the American Chemical Society **133**, 14504–14506 (2011).

[37] D. Coman and I. M. Russu, “A Nuclear Magnetic Resonance investigation of the energetics of basepair opening pathways in DNA,” Biophysical Journal **89**, 3285 – 3292 (2005).

[38] Y. Yin, L. Yang, G. Zheng, C. Gu, C. Yi, C. He, Y. Q. Gao, and X. S. Zhao, “Dynamics of spontaneous flipping of a mismatched base in DNA duplex,” Proceedings of the National Academy of Sciences **111**, 8043–8048 (2014).

[39] G. P. Pfeifer, “Formation and processing of UV Photoproducts: Effects of DNA sequence and chromatin environment,” Photochemistry and Photobiology **65**, 270–283 (1997).

[40] L. L. O’Nei and O. Wiest, “Structures and energetics of base flipping of the thymine dimer depend on DNA sequence,” The Journal of Physical Chemistry B **112**, 4113–4122 (2008).

[41] M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham, P. D. Dans, B. Jayaram, F. Lankas, C. Laughton, J. Mitchell, R. Osman, M. Orozco, A. Pérez, D. Petkevičiūtė, N. Spackova, J. Sponer, K. Zakrzewska, and R. Lavery, “ABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA,” Nucleic Acids Research **42**, 12272–12283 (2014).

[42] M. Zgarbova, P. Jureck, and M. Otyepka, “Influence of BII backbone substates on DNA twist: A unified view and comparison of simulation and experiment for all 136 distinct tetranucleotide sequences,” J. Chem. Inf. Model. , 13 (2017).

[43] Y. Sugita and Y. Okamoto, “Replica-exchange molecular dynamics method for protein folding,” Chemical Physics Letters **314**, 141 – 151 (1999).

[44] L. Wang, R. A. Friesner, and B. J. Berne, “Replica exchange with solute scaling: A more efficient version of replica exchange with solute tempering (REST2),” The Journal of Physical Chemistry B **115**, 9431–9438 (2011).

[45] A. P. Bhati, S. Wan, and P. V. Coveney, “Ensemble-based replica exchange alchemical free energy methods: The effect of protein mutations on inhibitor binding,” Journal of Chemical Theory and Computation **15**, 1265–1277 (2019).

[46] M. Katava, M. Kalimeri, G. Stirnemann, and F. Sterpone, “Stability and function at high temperature. What makes a thermophilic GTPase different from its mesophilic homologue,” The Journal of Physical Chemistry B **120**, 2721–2730 (2016).

[47] M. Kamiya and Y. Sugita, “Flexible selection of the solute region in replica exchange with solute tempering: Application to protein-folding simulations,” The Journal of Chemical Physics **149**, 072304 (2018).

[48] N. Parikh and D. K. Klimov, “Inclusion of lipopeptides into the DMPC lipid bilayers prevents Aβ peptide insertion,” Physical Chemistry Chemical Physics **19**, 10087–10098 (2017).

[49] S. Jo, D. Myatt, Y. Qi, J. Doutch, L. A. Clifton, W. Im, and G. Widmalm, “Multiple conformational states contribute to the 3D structure of a glucan decasaccharide: A combined SAXS and MD simulation study,” The Journal of Physical Chemistry B **122**, 1169–1175 (2018).

[50] P. Kührová, R. B. Best, S. Bottaro, G. Bussi, J. Šponer, M. Otyepka, and P. Banáš, “Computer folding of RNA Tetraloops: Identification of key Force Field deficiencies,” Journal of Chemical Theory and Computation **12**, 4534–4548 (2016).

[51] P. Kührová, V. Mlýnský, M. Zgarbová, M. Krepl, G. Bussi, R. B. Best, M. Otyepka, J. Šponer, and P. Banáš, “Improving the performance of the Amber RNA Force Field by tuning the Hydrogen-bonding interactions,” Journal of Chemical Theory and Computation **15**, 3288–3305 (2019).

[52] M. Havrila, P. Stadlbauer, P. Kührová, P. Banáš, J.-L. Mergny, M. Otyepka, and J. Šponer, “Structural dynamics of propeller loop: towards folding of RNA G-quadruplex,” Nucleic Acids Research **46**, 8754–8771 (2018).

[53] D. Case, “amber molecular dynamics package,” et al. **AMBER 2018**, University of California San Francisco (2018).

[54] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Comparison of simple potential functions for simulating liquid water,” The Journal of chemical physics **79**, 926–935 (1983).

[55] I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, *et al.*, “Parmbsc1: a refined force field for DNA simulations,” Nature methods **13**, 55 (2016).

[56] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, “A smooth particle mesh Ewald method,” The Journal of chemical physics **103**, 8577–8593 (1995).

[57] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD’96 (AAAI Press, 1996) p. 226–231.

[58] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska, “Conformational analysis of nucleic acids revisited: Curves+,” Nucleic Acids Research **37**, 5917–5929 (2009).

[59] M. Plummer, N. Best, K. Cowles, and K. Vines, “Coda: Convergence diagnosis and output analysis for mcmc,” R News **6**, 7–11 (2006).

[60] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2017).

[61] W. Humphrey, A. Dalke, and K. Schulten, “VMD – Visual Molecular Dynamics,” Journal of Molecular Graphics **14**, 33–38 (1996).

[62] L. Ayadi, C. Coulombeau, and R. Lavery, “Abasic sites in duplex DNA: Molecular modeling of sequence-dependent effects on conformation,” Biophysical Journal **77**, 3218 – 3226 (1999).

[63] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general amber force field,” Journal of computational chemistry **25**, 1157–1174 (2004).