Estimating the vertical ionization potential of single-stranded DNA molecules

Marianne Rooman[†,‡] and Fabrizio Pucci[*,†,‡]

† *Computational Biology and Bioinformatics, Université Libre de Bruxelles, 1050 Brussels, Belgium*

‡*Interuniversity Institute of Bioinformatics in Brussels, 1050 Brussels, Belgium*

E-mail: Marianne.Rooman@ulb.be

Abstract

The electronic properties of DNA molecules, defined by the sequence-dependent ionization potentials of nucleobases, enable long-range charge transport along the DNA stacks. This has been linked to a range of key physiological processes in the cells and to the triggering of nucleobase substitutions, some of which may cause diseases. To gain molecular-level understanding of the sequence dependence of these phenomena, we estimated the vertical ionization potential (vIP) of all possible nucleobase stacks in B-conformation, containing one to four Gua, Ade, Thy, Cyt or methylated Cyt. To do this, we used quantum chemistry calculations and more precisely the second-order Møller-Plesset perturbation theory (MP2) and three double-hybrid density functional theory (DFT) methods, combined with several basis sets for describing atomic orbitals. The calculated vIP of single nucleobases were compared to experimental data and those of nucleobase pairs, triplets and quadruplets, to observed mutability frequencies in the human genome, reported to be correlated with vIP values. This comparison selected MP2 with the 6-31G* basis set as the best of the tested calculation levels. These results were exploited to set up a recursive model, called vIPer, which estimates the vIP of all possible single-stranded DNA sequences of any length based on the calculated vIPs of overlapping quadruplets. vIPer's vIP values correlate well with oxidation potentials measured by cyclic voltammetry and activities obtained through photoinduced DNA cleavage experiments, further validating our approach. vIPer is freely available on the github.com/3BioCompBio/vIPer repository.

Introduction

DNA molecules are not only the repository for genetic information of all living organisms, but they also have unique electronic properties. Indeed, they can act as molecular wires formed by the π-stacking of the aromatic moieties of the nucleobases, along which long-range charge transport takes place.[1,2] Such charge transport even occurs in cells, where exposure to high-energy radiations or to reactive oxygen species generated as by-products of the cellular metabolism can lead to DNA ionization through the creation of electron holes.[3,4] These holes have been shown to migrate along the nucleobase stack until they remain localized in a region of low ionization potential, where they basically have two consequences: they are repaired by specific enzymes[5] or trigger single base substitutions.[6]

Recently, several studies have emphasized the link between DNA mutability and the vertical ionization potential (vIP) of the DNA sequence where the mutation occurs. It has indeed been shown that the base substitution rates depend on the DNA sequence context,[7,8] and that the lower the vIP of the motif formed by a substituted nucleobase and its adjacent DNA sequence, the more likely the substitution, whether for somatic cancer mutations or germline mutations.[6,9,10] The existence of a link of vIP and mutability with pathogenicity has been suggested and discussed.[6,7,11]

The various effects and possible roles of electron holes migrating along cellular DNA are not yet fully elucidated. Too many holes are obviously damaging to the cell if they are not repaired in time; for example, deficiency in repair enzymes are known to cause diseases such as cancer or neurodegeneration.[12–14] But a limited amount of electron holes appears to have a physiological role, as supported by the precise regulation of the amount of reactive oxygen species, which is, e.g., higher during cell differentiation.[15] Moreover, charge transfer seems to be used by the cell as a signaling mechanism, e.g., to cooperatively signal damage to DNA repair proteins.[5,16] The biological role of charge transfer is also supported by the recurrent occurrence of specific DNA conformations consistent with charge migration at the interface with specific proteins, ligands and metal ions.[17]

In this fascinating biological context, it is of utmost importance to have an accurate estimation of the ionization potential of nucleobase π-stacks. The four nucleobases that make up DNA have different propensities to be ionized, with the consequence that all possible nucleobase motifs have also different vIP values. The vIPs of the individual nucleobases have been measured experimentally,[18] but not those of sequences of nucleobase π-stacks. To estimate the vIPs of nucleobase stacks, quantum chemistry calculations have been performed using Hartree-Fock (HF)[19–21] and M06-2X, a simple hybrid density functional theory (DFT).[22,23] However, π-π stacking interactions involve strong electron delocalization and thus large dispersion contributions,[24,25] which are not represented well enough in these theories[26]. These calculations nevertheless demonstrated the high dependence of vIP values on DNA sequence, but also on DNA conformations.[23] Other ways of estimating the vIPs of nucleobase stacks involves using molecular dynamics simulations in conjunction with various levels of quantum chemistry calculations.[27–29] However, these methods are too computer time consuming to obtain the vIPs of all possible nucleobase stacks.

To fill this gap, we present here a user-friendly software called vIPer that estimates the vIP of B-form single-stranded DNA stretches of any length based on quantum chemistry calculations on π-stacks formed of four successive nucleobases combined with mathematical modeling techniques. Since cytosine methylation has crucial functions within cells and is involved in gene expression, development and cancer,[30] we considered 5-methylcytosine (5mC) in addition to the four basic DNA nucleobases Ade, Cyt, Gua and Thy.

Methods

Quantum chemistry calculations

Since dispersive forces have a huge importance in DNA base stacking,[24,25] we used levels of calculation known to account for these forces in a satisfactory manner - one ab initio and three DFT methods.

The ab initio method we used is second-order Møller-Plesset perturbation theory (MP2).[31] Although MP2 has been shown to systematically overestimate π-stacking energy,[32] this effect is rather weak and not very problematic in the present context as we focus on ionization potentials and thus on energy differences, which we moreover compare between different DNA motifs. MP2 also has the advantage of being an ab initio method that, while considering electron correlations, is relatively inexpensive in terms of computation.

We also used three double-hybrid methods that combine exact HF exchange with MP2-like correlation to DFT, i.e. B2-PLYP,[33] mPW2-PLYP[34] and PBE0-DH.[35] Note that the amount of exact exchange is much bigger in these methods than in the simple hybrids and dispersion forces are more accurately computed. The computational cost of this class of DFT methods is similar to that of MP2.

In addition, we tested several basis sets for atomic orbitals. First we considered the split-valence basis sets 6-31G* and 6-31G** with polarization functions on heavy atoms, and on heavy and hydrogen atoms, respectively.[36,37] In these basis sets, the exponent α_d of the Gaussian d-polarization function on the heavy atoms is equal to 0.8. We also considered two modified basis sets with α_d=0.2, which we denote here 6-31G$^{(.2)}$ and 6-31G$^{(.2)}$*. The α_d-exponent modulates the spatial extension of the wave function, with lower values corresponding to more diffuse orbitals. The value of 0.2 has been determined to minimize the MP2 energy of a stair motif involving two stacked Gua bases forming a hydrogen bond and a cation-π interaction with an Arg residue,[38] a common motif found at protein-DNA interfaces.[39] Note that α_d=0.2 is close to the value of 0.25 used by other authors to estimate the stacking energies between aromatic systems.[24]

All quantum chemistry calculations were performed in gas phase and with the Gaussian 16 suite.[40]

Single nucleobase and nucleobase stack geometries

We considered the four nucleobases Ade (A), Cyt (C), Gua (G) and Thy (T) whose combinations form DNA molecules. In addition, we considered 5-methylcytosine (5mC or M) that is the common methylated form of cytosine. In these five nucleobases, the sugar cycle and phosphate group were omitted, and the glycosidic bond was replaced by a hydrogen atom.

The initial geometries of the single nucleobases were taken from the software package x3DNA-DSSR.[41] These geometries were optimized using the different levels of theory and basis sets described in the previous subsection.

In a second stage, we considered single-stranded nucleobase stacks in standard B conformation. The geometries of the stacks were designed using x3DNA-DSSR. This software proposes two generic B-DNA conformations that both have a twist of 36.0°. The first, noted here B1 (B55 in x3DNA-DSSR) has a rise of 3.39 Å ,[42] and the second, noted B2 (B4 in x3DNA-DSSR), a rise of 3.375 Å .[43] We then replaced each of the nucleobases forming the stacks with the same nucleobases but with optimized geometry. This was done by superimposing the optimized bases onto the non-optimized ones by minimizing the root mean square deviation of atomic positions using the U3BEST algorithm.[44] Examples of such stacks are shown in Fig. 1.

Vertical ionization potential (vIP)

We considered two species of each molecule: a neutral species S and a radical cationic species $S^{\bullet +}$ with one missing electron. The vIP is defined as the difference in energy between these two species, both considered to adopt the optimal conformation C of the neutral species,

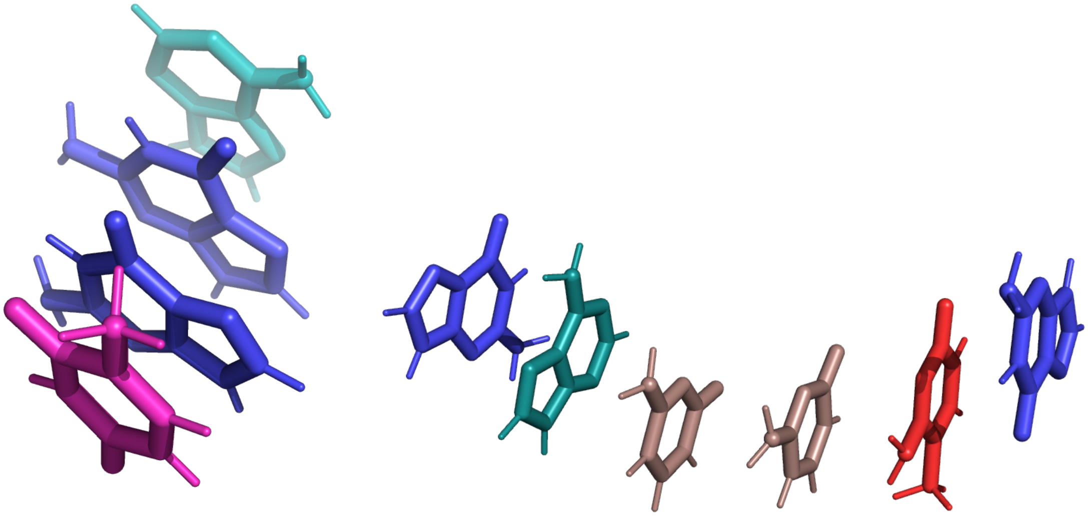

Figure 1: Single-stranded B1 stacking geometry of a TGGA nucleobase quadruplet and a GACCMG sextuplet. Gua are in blue, Ade in teal, Thy in pink, Cyt in darksalmon and 5mC in red.

thus ignoring possible changes in molecular geometry that may result from ionization:

$$\mathrm{vIP}(S, C) = E(S, C) - E(S^{\bullet +}, C) \tag{1}$$

The energy calculations were performed using the levels of theory and basis sets described in the previous subsections. For cationic species, they were performed with restricted open-shell procedures to prevent spin contamination problems.

Results

vIP of single nucleobases

As a first stage, we computed the vIP of the four nucleobases Gua, Ade, Cyt and Thy at MP2, B2-PLYP, mPW2-PLYP and PBE0-DH levels of theory with 6-31G*, 6-31G**, 6-31G$^{(.2)}$ and 6-31G$^{(.2)*}$ basis sets (see Methods), and compared them with experimentally measured vIP values.[18] The results are given in Table 1 and Supplementary Table S1.

Table 1: Calculated and experimental vIP values (in eV) of single nucleobases. The root mean square error (RMSE; in eV) and the Pearson correlation coefficient (r) with the associated p-values are between calculated and experimental[18] vIP values.

Level of theory	MP2		B2-PLYP		Experi
Basis set	6-31G*	6-31G$^{(.2)}$	6-31G*	6-31G$^{(.2)}$	mental[18]
Gua	7.86	7.99	7.71	7.83	8.24
Ade	8.30	8.37	8.04	8.17	8.44
Cyt	8.56	8.56	8.43	8.54	8.94
Thy	8.77	8.88	8.72	8.83	9.14
RMSE	0.33	0.26	0.47	0.35	
r	0.96	0.96	0.99	0.99	
p-value	0.04	0.04	0.009	0.01	
Level of theory	mPW2-PLYP		PBE0-DH		
Basis set	6-31G*	6-31G$^{(.2)}$	6-31G*	6-31G$^{(.2)}$	
Gua	7.75	7.86	7.94	8.00	
Ade	8.08	8.20	8.27	8.35	
Cyt	8.49	8.59	8.73	8.79	
Thy	8.77	8.88	8.98	9.02	
RMSE	0.42	0.31	0.22	0.16	
r	0.99	0.99	1.00	0.99	
p-value	0.008	0.01	0.005	0.007	

We observe that, when considering the nucleobases individually, all the levels of theory and basis sets that we tested give very accurate results, with a slight preference for double hybrid DFT methods, and in particular PBE0-DH/6-31G* that reaches $r = 1$ and PBE0-DH/6-31G$^{(.2)}$ that yields RMSE=0.16 eV. Very low difference is observed between 6-31G* and 6-31G**, and between 6-31G$^{(.2)}$ and 6-31G$^{(.2)}$*. Note, however, that there is an uncertainty of about ±0.03 eV on the experimental values,[18] which means that the observed differences in RMSE and r must be considered with caution.

vIP of nucleobase pairs

We calculated the vIP values of all 16 pairs of successive nucleobases containing Gua, Ade, Thy and Cyt in B1 conformation, using the four levels of theory and the four basis sets considered here (see Methods). The results are given in Table 2 and Supplementary Table

S2. We considered only single-stranded DNA molecules, as electron holes are known to generally migrate along a single strand and to only rarely jump onto the complementary strand.[45]

Here, we do not have experimental vIP values to compare with the calculated vIPs, and we have thus no means to select the best level of theory and basis set. A possibility would be to compare our calculated vIPs with those obtained with gold-standard ab initio quantum chemistry methods such as CCSD(T),[46] but these are much too computer intensive to be applied to calculate the vIPs of all possible nucleobase pairs, triplets and quadruplets.

Another possibility is to correlate the calculated vIP values with mutability data. Indeed, we showed earlier[6] that there is a statistically significant anticorrelation between the frequency of single base substitutions in human genomes and the vIP of the motif consisting of the substituted base and its flanking DNA sequence. In a nutshell, the lower the vIP value, the more probable the substitution. We chose this point of view here and stated that the quantum chemistry method which yields the best anticorrelation is also the most reliable method for calculating the vIP of the considered molecules.

The rationale for supposing a relationship between vIP and mutability originates from the effect of oxidative stress. In the cells, this stress is due to physical or chemical agents, such as exposure to ionizing radiation, long-wave ultraviolet light, and reactive oxygen species. It often causes the extraction of electrons from the DNA molecules. The electron holes so created then migrate along the DNA stack towards regions of low vIP, where they are either repaired by specific enzymes or are likely to mutate.[3–6]

To get mutability data, we used the ClinVar database,[47] which contains essentially germline variants in human genomes and their associated phenotypes. We focused here on synonymous variants in exon regions. The reason of this choice is that missense mutations cause amino acid substitutions in the encoded proteins, of which some are likely to affect protein stability or function too strongly, so that they do not get fixed. The anticorrelation of the vIP values with synonymous variant frequencies are therefore expected to be better than when considering missense mutations, which is indeed what is observed.[6]

Table 2: Calculated vIP values (in eV) of pairs of successive nucleobases in B1 conformation. The Pearson correlation coefficient r and associated p-values are computed between the calculated vIP and $\log F_{mut}^{\mathrm{XN}}$, the logarithm of the mutation frequency.

Level of theory	MP2		B2-PLYP		mPW2-PLYP		PBE0-DH	
Basis set	6-31G*	6-31G$^{(.2)}$	6-31G*	6-31G$^{(.2)}$	6-31G*	6-31G$^{(.2)}$	6-31G*	6-31G$^{(.2)}$
GG	7.47	7.61	7.27	7.42	7.32	7.45	7.47	7.56
GA	7.73	7.78	7.51	7.64	7.56	7.67	7.69	7.80
GC	7.76	7.81	7.58	7.69	7.61	7.72	7.78	7.86
GT	7.74	7.86	7.57	7.69	7.61	7.72	7.79	7.85
AG	7.57	7.71	7.26	7.48	7.34	7.56	7.52	7.62
AA	7.99	7.77	7.50	7.67	7.58	7.74	7.78	7.89
AC	8.25	8.25	7.96	8.08	8.00	8.12	8.17	8.26
AT	8.17	8.24	7.86	8.01	7.92	8.06	8.09	8.18
CG	7.41	7.55	7.23	7.38	7.28	7.40	7.44	7.53
CA	7.85	7.94	7.56	7.70	7.60	7.73	7.78	7.87
CC	8.07	8.04	7.88	8.00	7.99	8.10	8.15	8.24
CT	8.34	8.42	7.89	8.02	7.98	8.11	8.24	8.32
TG	7.73	7.85	7.52	7.67	7.58	7.70	7.75	7.83
TA	8.16	8.21	7.85	7.98	7.90	8.02	8.08	8.16
TC	8.19	8.16	8.14	8.27	8.24	8.36	8.46	8.54
TT	8.11	8.18	8.13	8.25	8.22	8.34	8.48	8.54
r	-0.80	-0.70	-0.73	-0.75	-0.72	-0.74	-0.71	-0.72
p-value	2×10^{-4}	2×10^{-3}	1×10^{-3}	8×10^{-4}	2×10^{-3}	9×10^{-4}	2×10^{-3}	2×10^{-3}

We thus computed from the ClinVar database the frequencies of all possible base pair motifs with the first base mutated, referred to as F_{mut}^{XN}, where X denotes the mutated position and N any base. The F_{mut}^{XN} values and details about their calculation can be found in a previous article.[6] We then correlated the vIP values for all base pairs with the logarithm of their mutation frequency ($\log F_{mut}^{\mathrm{XN}}$). The Pearson correlation coefficient r is given in Table 2 and Supplementary Table S2. We first observe that the correlation is negative, as expected, as lower vIP values attract electron holes, which triggers mutation mechanisms.

Moreover, we see that r is by far the most negative for MP2 level of theory compared to the considered double-hybrid DFT methods. The result here is thus different than for individual nucleobases, where the DFT methods appeared slightly superior. The reason is that, in pairs of stacked nucleobases, π-orbitals overlap and electron delocalization is an important contribution to stability, whereas it is not the case for individual bases. We conclude that for interactions involving aromatic stacking between nucleobases, MP2 appears as a better level of theory than the tested double-hybrid DFT methods.

Moreover, we observe that the basis sets 6-31G$^{(.2)}$ and 6-31G$^{(.2)}$*, which represent more diffuse orbitals than the usual 6-31G* and 6-31G** sets (see Methods), lead to slightly better correlation coefficients when associated with DFT calculations, but to worse correlations when associated with MP2. This contrasts with the results obtained for interaction energies at MP2 level.[24,38] Furthermore, adding polarization on hydrogen atoms, i.e. considering the 6-31G** set rather than 6-31G*, and 6-31G$^{(.2)}$* rather than 6-31G$^{(.2)}$, does not improve the results on the average. As the inclusion of these polarization functions is computationally expensive, we choose not to consider them in what follows.

We also calculated the vIP values of all the nucleobase pairs in B2 conformation at MP2 level of theory using the four different basis sets considered (Supplementary Table S3). Clearly, the vIP values calculated for nucleobase stretches adopting this type of B-conformation are less well correlated with the mutation frequency than those in B1 conformation. This corroborates the high sensitivity of vIP values to DNA conformation previously observed,[23] and explains the lower vIP-frequency correlations that we found in a previous study.[6] This result suggests that the B1 conformation is closer to the average DNA conformation adopted in the cell than the B2 conformation.

In summary, on the basis of the correlation between vIP and mutability values, we conclude that the best of the tested quantum chemistry methods for DNA stacks is MP2/6-31G*, and that the most realistic of the tested single-stranded DNA stack conformations is B1. We thus limit ourselves to this level of theory and this DNA conformation in what follows.

vIP of triplets and quadruplets

We used the MP2/6-31G* quantum chemistry theory to estimate the vIP of all 64 base triplet motifs and all 256 base quadruplet motifs. The results are shown in Table 3 and Supplementary Table S4. The range of vIP values increases with the number of bases in the motif and the average vIP value decreases, as shown in Fig. 2.

In the same way as we did for pair motifs in the previous section, we calculated the correlation between the vIP values of base triplets and quadruplets with $\log F_{mut}^{\mathrm{NXN}}$ and $\log F_{mut}^{\mathrm{NXNN}}$, respectively, i.e. the logarithm of the frequencies of all possible base triplet and quadruplet motifs with the middle base mutated, as obtained from the ClinVar database.[47] The correlation coefficient decreases somewhat, from -0.80 for pairs, -0.60 for triplets and -0.50 for quadruplets, but the statistical significance increases, as shown by the p-values of 2×10^{-4}, 1×10^{-7} and 7×10^{-18}. These correlations are thus highly statistically significant.

Table 3: Calculated vIP values of triplets of successive nucleobases in B1 conformation (in eV). The linear correlation coefficient r between the calculated vIP and $\log F_{mut}^{\mathrm{NXN}}$, the logarithm of the mutation frequency, is equal to -0.60 (p-value 1×10^{-7}).

		MP2		6-31G*			
GGG	7.09	AGG	7.19	CGG	7.03	TGG	7.34
GGA	7.44	AGA	7.44	CGA	7.29	TGA	7.60
GGC	7.45	AGC	7.47	CGC	7.32	TGC	7.63
GGT	7.43	AGT	7.45	CGT	7.30	TGT	7.61
GAG	7.58	AAG	7.45	CAG	7.38	TAG	7.51
GAA	7.71	AAA	7.53	CAA	7.44	TAA	7.66
GAC	7.73	AAC	7.75	CAC	7.81	TAC	8.11
GAT	7.69	AAT	7.71	CAT	7.74	TAT	8.04
GCG	7.48	ACG	7.31	CCG	7.25	TCG	7.36
GCA	7.76	ACA	7.75	CCA	7.68	TCA	7.80
GCC	7.78	ACC	8.27	CCC	7.45	TCC	8.02
GCT	7.73	ACT	8.22	CCT	7.51	TCT	7.73
GTG	7.79	ATG	7.62	CTG	7.56	TTG	7.68
GTA	7.72	ATA	8.04	CTA	7.99	TTA	8.11
GTC	7.75	ATC	8.19	CTC	8.42	TTC	8.17
GTT	7.70	ATT	8.13	CTT	8.20	TTT	8.47

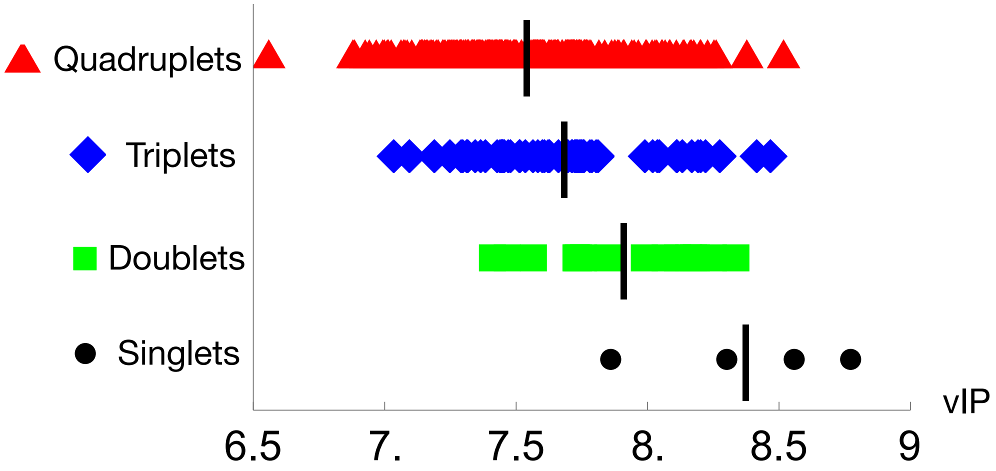

Figure 2: Calculated vIP values at MP2/6-31G* level of theory (in eV) for all singlet, pair, triplet and quadruplet base pair motifs. The mean values are indicated as vertical bars.

Methylated Cyt

The epigenetic process of DNA methylation involves adding methyl groups to specific nucleobases and thus modifies the activity of the DNA without altering its sequence. The most common methylation process targets cytosine and modifies it into 5-methylcytosine (5mC or M). Cytosine methylation and demethylation are highly regulated processes, which occur basically across all species where they play key roles in many biological processes.[48]

Because of its importance, we computed the vIP of 5mC using MP2/6-31G* theory. As seen in Table 4, it is equal to 8.2 eV, and thus lower than the vIP of Cyt (8.6 eV) in agreement with earlier studies[49,50], and even slightly lower than the vIP of Ade (8.3 eV) (Table 1). The ranking from lowest to highest vIPs is thus: Gua, 5mC, Ade, Cyt, Thy. Gua thus remains the best electron hole trap, but 5mC appears as the second best. Cytosine methylation is therefore expected to impact long-range charge transport through DNA as well as mutability properties.

We also calculated the vIPs of all pairs, triplets and quadruplets containing at least one 5mC (Table 4 and Supplementary Table S5). By comparing these results with Tables 2, 3 and S4, we see that Cyt methylation tends to lower the vIP of the DNA stacks. As lower vIPs are associated with higher mutability, we expect that DNA regions in which cytosines are methylated are more subject to mutation,

To check this, we correlated the vIP values of all pairs, triplets and quadruplets in which we considered all cytosines to be methylated, with the logarithm of the frequency of synonymous variants in ClinVar; note that it is not known whether the cytosines are methylated or not. By doing that, we found better correlation coefficients when assuming all Cyt bases to be methylated rather than unmethylated. As shown in Fig. 3, we found indeed that r improves from -0.80 to -0.90 for pairs (p-value 2×10^{-4} and 2×10^{-6}), from -0.60 to -0.64 for triplets (p-value 1×10^{-7} and 1×10^{-8}), and -0.50 to -0.54 for quadruplets (p-value 7×10^{-18} and 2×10^{-20}). This result suggests that many of the cytosines in the genome are methylated, and that the mutability of motifs with methylated cytosines is on the average higher than with unmethylated cytosines, which is indeed observed.[51] It also supports the hypothesis that the higher mutability of 5mC compared to Cyt is due to its lower vIP value.

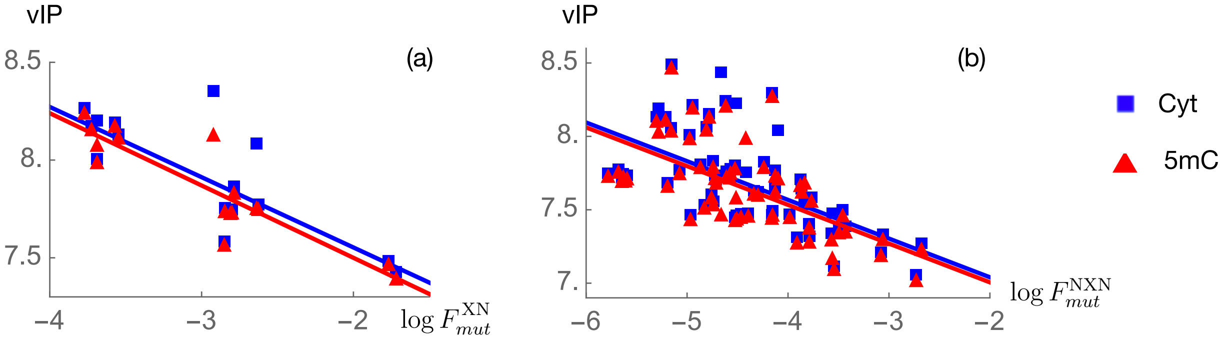

Figure 3: Calculated vIP values at MP2/6-31G* level of theory (in eV) as a function of the logarithm of the mutation frequency, for : (a) all nucleobase pairs of Gua, Ade, Thy and Cyt (blue) and of Gua, Ade, Thy and 5mC (red) ; (b) all nucleobase triplets of Gua, Ade, Thy and Cyt (blue) and of Gua, Ade, Thy and 5mC (red). The linear regression lines are drawn.

Table 4: Calculated vIP values (in eV) of 5mC and of pairs and triplets of successive nucleobases in B1 conformation containing at least one 5mC (M).

		MP2		6-31G*			
M	8.23	GMM	7.78	CMM	7.58	MAT	7.72
GM	7.75	AGM	7.46	TGM	7.62	MAM	7.79
AM	8.24	AAM	7.75	TAM	8.1	MCG	7.24
CM	7.78	ACM	8.28	TCM	7.73	MCA	7.68
TM	8.08	ATM	8.19	TTM	8.03	MCC	7.52
MG	7.39	AMG	7.3	TMG	7.35	MCT	7.60
MA	7.83	AMA	7.73	TMA	7.79	MCM	7.60
MC	8.28	AMC	8.27	TMC	8.12	MTG	7.56
MT	8.13	AMT	8.21	TMT	7.99	MTA	7.99
MM	7.75	AMM	8.27	TMM	7.71	MTC	8.18
GGM	7.46	CGM	7.31	MGG	7.02	MTT	7.58
GAM	7.73	CAM	7.81	MGA	7.27	MTM	7.47
GCM	7.78	CCM	7.6	MGC	7.3	MMG	7.23
GTM	7.76	CTM	7.9	MGT	7.28	MMA	7.67
GMG	7.46	CMG	7.24	MGM	7.3	MMC	7.8
GMA	7.75	CMA	7.67	MAG	7.38	MMT	7.68
GMC	7.78	CMC	7.82	MAA	7.43	MMM	7.17
GMT	7.72	CMT	7.71	MAC	7.79		

vIPer for long single-stranded DNA stacks

We constructed a simple mathematical model based on the vIP values of single bases, doublets, triplets and quadruplets, calculated at MP2/6-31G* level of theory, which were combined to predict the vIP of single-stranded nucleobase sequences of any length, containing arbirary combinations of Gua, Ade, Thy, Cyt and 5mC. To test the model, we also calculated at MP2/6-31G* level the vIP values of 249 randomly chosen quintuplets and 42 sextuplets; we call these our test set.

We started by noticing that the vIP values of nucleobase stacks generally decrease as their length L increases (see Figs 2 and 4.a). This is expected, as π-π stacks of increased length allow for more efficient charge delocalization and are easier to ionize. We chose an exponential decay function to model the average stack length dependence of the vIPs:

$$\mathcal{D}(L) = S \exp^{-a(L^b-1)} \tag{2}$$

where S is the average of the vIP values of the five single bases Gua, Ade, Thy, Cyt and 5mC. The parameters a and b were identified in order to minimize the RMSE between $\mathcal{D}(L)$ and the average vIP of all nucleobase stack motifs of given length $L = 1, \ldots 4$. The resulting $\mathcal{D}(L)$ function is plotted in Fig. 4.a. For $L = 1$, we have $\mathcal{D} = S = 7.86$; for $L > 1$, $\mathcal{D}$ decays exponentially; and for very large L, $\mathcal{D}$ is around 0.6.

The vIP of a given nucleobase sequence $(x_1....x_L)$ of length L, where each x_i is one of the five nucleobases Gua, Ade, Thy, Cyt and 5mC, is then expressed as a recursive function of $\mathcal{D}(L)$ and of the vIP values of the two overlapping sequences of length $L-1$, i.e. $(x_1....x_{L-1})$ and $(x_2....x_L)$, as:

$$\text{vIP}(x_1...x_L) = \mathcal{D}(L) + \frac{\alpha\ L^2}{L^2-2}[\text{vIP}(x_1..x_{L-1}) - \mathcal{D}(L-1)] + \frac{(1-\alpha\,)L^2}{L^2-2}[\text{vIP}(x_2..x_L) - \mathcal{D}(L-1)] \tag{3}$$

where $L > 1$. The parameter $\alpha = 0.4$ was identified by minimizing the RMSE between the vIP values of all possible quadruplet stacks either calculated at MP2/6-31G* level or estimated via Eq. (3) for $L = 4$. The factor $L^2/(L^2-2)$ was introduced to avoid the shrinking of the standard deviation of the distribution of vIP values for long DNA stacks.

Note that we also tried other recursive models, which contained additional free parameters to optimize or were based on more complex model structures. However, we found that the selected model defined in Eq. (3), which has a very simple structure and only α as free parameter, in which the vIP of an L-tuplet is basically derived from the vIP of the two overlapping (L-1)-tuplets, is less prone to biases and reaches the best score.

An important observation from our model is that the value of the α parameter is lower than 0.5, which means that the vIP value of the (L-1)-tuplet near the 3' end contributes more than the vIP of the (L-1)-tuplet near the 5' end to the vIP value of the L-tuplet. This asymmetry can be related to the 5'→3' directional preference of the hole transport through the DNA stack.

To check the performance of the model, we applied it to calculate the vIP values of the quintuplet and sextuplet motifs of our test set. As shown in Fig. 4.b-c, the results are very good with Pearson correlation coefficients between our model predictions and our quantum chemistry calculations of 0.87 (p-value 9×10^{-79}) and 0.81 (p-value 2×10^{-11}) for the quintuplets and sextuplets, respectively, and a RMSE of 0.15 and 0.17 eV.

To make the results of our prediction model available to the scientific community, we developed a python package called vIPer, which can be freely dowloaded from our GitHub repository (github.com/3BioCompBio/vIPer). For nucleobase sequences of length between one and four, vIPer outputs the vIP value calculated at MP2/6-31G*. For sequences longer than four bases, vIPer uses the recursive model of Eq. (3). vIPer is simple to use: the user submits a nucleobase sequence and gets the vIP values predicted by the model for the input sequence as well as the average over the two complementary strands. It has the advantage of avoiding intense quantum chemistry calculations to estimate the vIP of long nucleobase stacks, while maintaining very good precision. Examples of predicted vIP values for DNA sequences of interest are shown in Table 5.

Table 5: vIP values predicted by vIPer (in eV) of examples of long single-base DNA stacks in B1 conformation: telomers,[52] pericentromeric satellite repeats[53] and methylated or unmethylated CpG islands.[54] The vIP column contains the predicted vIP of the motif and the <vIP> column, the predicted vIP averaged over the two complementary strands.

Type	organisms	DNA motif	vIP	<vIP>
Telomers	vertebrates	$(TTAGGG)_{100}$	5.50	5.82
Telomers	plants	$(TTTAGGG)_{100}$	5.67	5.91
pericentromeric satellite	*Drosophila*	$(AATAT)_{35}$	6.55	6.73
pericentromeric satellite	*Drosophila*	$(AAGAGAG)_{25}$	6.20	6.64
pericentromeric satellite	*Drosophila*	$(AATAACATAG)_{8}$	6.48	6.69
CpG island	vertebrates	$(CG)_{500}$	5.59	5.59
methylated CpG	vertebrates	$(MG)_{500}$	5.56	5.56

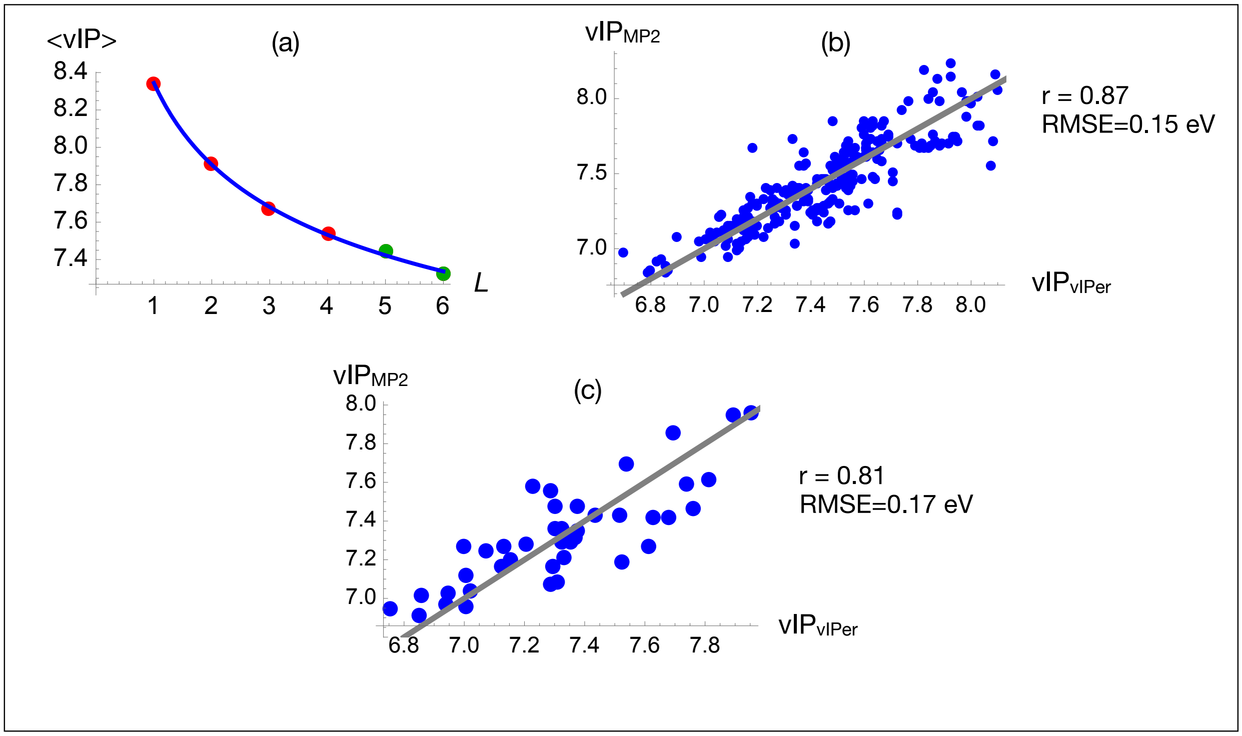

Figure 4: Prediction of vIP values for long single-stranded DNA sequences (a) Exponential decay of the average vIP value of nucleobase sequences of given length L as a function of L. The points correspond to the vIP values calculated at MP2/6-31G* level; for $L = 1, \dots 4$ (red points) the average is over all possible nucleobase motifs and and for $L = 4, 5$ (green points), it is over all randomly chosen motifs included in the test set. The blue curve is the exponential decay curve $\mathcal{D}(L)$ defined in Eq. (2) fitted on the red points $L = 1, \dots 4$. The green points that correspond to $L = 5, 6$ are not used for fitting. (b)-(c) The vIP values calculated at MP2/6-31G* level as a function of the vIP values calculated by our model defined by Eq. (3) for the test set of quintuplet motifs (b) and sextuplet motifs (c). The Pearson correlation coefficient r and RMSE (in eV) are given in the plots. The lines in (b) and (c) are the bisectors of quadrants I and III and correspond to perfect predictions.

Validation of vIPer on experimental data

To further validate the vIPer model, we compared its predictions with experimental data. We first considered photoinduced DNA cleavage experiments.[20] In these experiments, one-electron oxidation of G- and GG-containing DNA segments were examined as a function of the flanking sequence through the measurement of their relative reactivity (k_{rel}). The nucleobase motif that is varied is the 5-tuplet TNNNT, where N denotes any nucleobase; the full 30-tuplet sequence is specified in Table 6. We compared k^{rel} with vIPer's vIP values of the 5-tuplet and 30-tuplet sequences and found excellent Pearson correlation coefficients of -0.91 and -0.93, respectively; these values are statistically significant (p-values 0.00004 and 9×10^{-6}).

As an additional validation test, we compared vIPer's predictions with experiments in which the effects of the DNA sequence on charge transport were analyzed using cyclic voltammetry of daunomycin cross-linked with different palindromic DNA duplexes.[55] Based on the empirical linear relationship between the ionization potential and the peak potential for oxidation V_{ox},[56,57] we compared the latter with the calculated vIP values for the considered duplexes, as shown in Table 7. The Pearson correlation coefficient between V_{ox} and vIP values is equal to 0.89 and statistically significant (p-value 0.003). This result further demonstrates the quality of vIPer.

Conclusion

We performed in this paper a huge number of quantum chemistry calculations at different levels of theory and using various basis sets in view of estimating the vIP values of all possible DNA stacks of arbitrary lengths. This is as such an important achievement. Indeed, vIPer is the very first automatic vIP calculator of single stranded DNA molecules of any sequence and any length, and moreover, it is very fast as it computes the vIP values in just a few seconds. We would like to underline that we found very good correlations between vIPer

Table 6: Comparison between the measured relative reactivity k_{rel} in photoinduced DNA cleavage experiments of 30-tuplet sequences[20] and their vIP value (in eV) calculated using vIPer. The DNA sequences are CGTACTCTTTGGTGGG**TNNNT**TCTTTCTAT, where N denotes any base. The tested TNNNT quintuplet motifs are given in column 1. The Pearson correlation coefficient r is between k_{rel} and vIP values.

Motifs	k_{rel}[20]	vIP	
		5-tuplet	30-tuplet
TGGGT	2.7	6.692	6.120
TAGGT	1.0	7.004	6.346
TTGGT	1.0	7.174	6.464
TCGGT	2.0	6.840	6.236
TGGAT	0.8	7.257	6.496
TAGAT	0.4	7.283	6.566
TTGAT	0.3	7.455	6.684
TCGAT	0.4	7.121	6.457
TGGTT	0.9	7.243	6.531
TAGTT	0.2	7.289	6.609
TGGCT	0.7	7.276	6.523
TAGCT	0.1	7.316	6.599
r		-0.91	-0.93
p-value		0.00004	9×10^{-6}

Table 7: Comparison between the vIP (in eV) calculated by vIPer and the peak potential for oxidation V_{ox} (in V) measured by cyclic voltammetry on the palindromic DNA duplexes of column 1.[55] The Pearson correlation coefficient r is between vIP and V_{ox}; the p-value is in parentheses.

	V_{ox} [55]	vIP
AGCT	-0.655	7.444
$(AA)_2AGCT(TT)_2$	-0.660	7.151
$(AT)_2AGCT(AT)_2$	-0.658	7.248
$(AA)_3AGCT(TT)_3$	-0.662	7.109
$(AT)_3AGCT(AT)_3$	-0.660	7.204
$(A)_2(AT)_2AGCT(AT)_2(T)_2$	-0.658	7.205
$(AA)_5AGCT(TT)_5$	-0.660	7.070
$(AT)_5AGCT(AT)_5$	-0.662	7.129
r		0.89 (0.003)

predictions and experimental cyclic voltammetry and photoinduced one-electron oxidation measurements.

Another point we would like to emphasize is the importance of considering the π-orbital overlap between stacked nucleobases and the associated dispersion energy contributions when performing vIP calculations. Indeed, we compared the MP2/6-31G* vIP values of all pairs, triplets and quadruplets obtained by calculating them either from the complete nucleobase stacks or as the average of the vIP of their constituting nucleobases. The former method is clearly superior. Indeed, we found that the Pearson correlation coefficients between the vIP and the logarithm of the mutations frequency ($\log F_{mut}$) deteriorate when passing from the former to the latter method from -0.80 to -0.70 for pairs, from -0.61 to -0.32 for triplets and from -0.50 to -0.23 for quadruplets.

vIPer has a whole series of applications. Indeed, the estimation of the electronic properties of DNA can be crucial in biomedical investigations aiming, for example, to predict and control the mutability of specific genome regions or other vIP-dependent biological processes. It is also important in biotechnological applications in which DNA molecules are used for their unique properties such as the construction of molecular wires or data storage systems made from DNA.[58]

We would like to stress that understanding the charge transport properties of DNA is still elusive and that even experimental studies have so far not provided definitive answers. Computational analyses are therefore of primary importance in this context for suggesting new hypotheses and guiding new research.

Even though vIPer is certainly a first important step towards accurate vIP estimations of DNA molecules, it is based on a number of approximations that are important to discuss:

- The use of gold-standard quantum chemistry methods such as CCSD(T)[46] rather than MP2 would be a clear improvement, but they are currently too time- and memory-consuming to be applied to all possible pair to quadruplet motifs.

- We omitted the sugar-phosphate backbone in all our calculations, as their inclusion would require too much additional computer time and memory. Another justification of this choice is that earlier calculations on Cyt and Thy motifs indicated that the vIP is strongly affected by the presence of the sugar and phosphate moieties in gas phase, but much less upon bulk hydration, due to the screening by the aqueous solvent.[59] Moreover, quantum chemistry calculations and experimental data showed that the lowest ionization pathway comes from the nucleobases rather than the sugar-phosphate backbone.[59,60]

- DNA environment makes the vIP estimation very challenging. Here we disregarded the solvent and performed all calculations in gas phase, although solvation is known as impacting the electronic properties of DNA molecules.[61] This is of course an approximation, but appears to be justified when comparing vIP values of various nucleobase stack sequences. Indeed, calculations on individual nucleobases highlighted the effect of the solvent in lowering the vIP values while maintaining the relative ordering between the bases.[62,63] Moreover, as discussed above, dropping both the sugar-phosphate backbone and the solvent have opposite effects that tend to cancel out.[59]

Finally note that the vIP values depend significantly on the type of DNA conformation.

Indeed, we clearly showed here that the vIPs differ according to the type of B-conformation we considered (B1 and B2), and it was previously shown that it is even more so when considering A- or Z-conformations instead of B-conformations.[23] We will in the future extend the vIPer algorithm to make it able to predict the vIP of other DNA conformations and introduce additional parameters in the model that take into account the high flexibility of DNA.

Acknowledgments

We thank Emilie Cauët and Nicolas Callebaut for interesting discussions at the beginning of the project. We acknowledge financial support from the FNRS - Fund for Scientific Research through a PDR Research Project; MR is FNRS Research Director.

Data and Software Availability

All quantum chemistry calculations were performed using the Gaussian 16 suite.[40] The geometry of the single nucleobases and nucleobase stacks were constructed using the software package x3DNA-DSSR,[41] and the nucleobases with optimized geometry were superimposed onto the DNA stacks using the U3BEST algorithm.[44] The python code of the vIPer model that estimates the vIP of single-stranded DNA molecules of any length in B conformation is freely available on our github repository github.com/3BioCompBio/vIPer.

References

(1) Giese, B. Long-distance electron transfer through DNA. *Annu. Rev. Biochem.* **2002**, *71*, 51.

(2) Chakraborty, T. *Charge migration in DNA: perspectives from physics, chemistry, and biology*; Springer Science & Business Media, 2007.

(3) Boon, E. M.; Barton, J. K. Charge transport in DNA. *Curr. Opin. Struct. Biol.* **2002**, *12*, 320–329.

(4) Grodick, M. A.; Muren, N. B.; Barton, J. K. DNA charge transport within the cell. *Biochemistry* **2015**, *54*, 962–973.

(5) Boal, A. K.; Genereux, J. C.; Sontz, P. A.; Gralnick, J. A.; Newman, D. K.; Barton, J. K. Redox signaling between DNA repair proteins for efficient lesion detection. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 15237–15242.

(6) Pucci, F.; Rooman, M. Relation between DNA ionization potentials, single base substitutions and pathogenic variants. *BMC genomics* **2019**, *20*, 1–11.

(7) Suárez-Villagrán, M. Y.; Miller Jr, J. H. Computational DNA hole spectroscopy: A new tool to predict mutation hotspots, critical base pairs and disease 'driver' mutations. *Sci. Rep.* **2015**, *5*, 1–12.

(8) Suárez-Villagrán, M. Y.; Azevedo, R. B.; Miller Jr, J. H. Influence of electron–holes on DNA sequence-specific mutation rates. *Genome Biol. Evol.* **2018**, *10*, 1039–1047.

(9) Bacolla, A. et al. Guanine holes are prominent targets for mutation in cancer and inherited disease. *PLoS Genet.* **2013**, *9*, e1003816.

(10) Bacolla, A.; Cooper, D. N.; Vasquez, K. M. Mechanisms of base substitution mutagenesis in cancer genomes. *Genes* **2014**, *5*, 108–146.

(11) Shih, C.-T.; Wells, S. A.; Hsu, C.-L.; Cheng, Y.-Y.; Römer, R. A. The interplay of mutations and electronic properties in disease-related genes. *Sci. Rep.* **2012**, *2*, 1–9.

(12) Katyal, S.; McKinnon, P. J. DNA repair deficiency and neurodegeneration. *Cell Cycle* **2007**, *6*, 2360–2365.

(13) Bermúdez-Guzmán, L.; Leal, A. DNA repair deficiency in neuropathogenesis: when all roads lead to mitochondria. *Transl. Neurodegener.* **2019**, *8*, 1–14.

(14) Wimmer, K.; Etzler, J. Constitutional mismatch repair-deficiency syndrome: have we so far seen only the tip of an iceberg? *Hum. Genet.* **2008**, *124*, 105–122.

(15) Aguirre, J.; Ríos-Momberg, M.; Hewitt, D.; Hansberg, W. Reactive oxygen species and development in microbial eukaryotes. *Trends Microbiol.* **2005**, *13*, 111–118.

(16) Zwang, T. J.; Tse, E. C.; Barton, J. K. Sensing DNA through DNA charge transport. *ACS Chem. Biol.* **2018**, *13*, 1799–1809.

(17) Rooman, M.; Cauët, E.; Liévin, J.; Wintjens, R. Conformations consistent with charge migration observed in DNA and RNA X-ray structures. *J. Biomol. Struct. Dyn.* **2011**, *28*, 949–954.

(18) Hush, N.; Cheung, A. S. Ionization potentials and donor properties of nucleic acid bases and related compounds. *Chem. Phys. Lett.* **1975**, *34*, 11–13.

(19) Saito, I.; Takayama, M.; Sugiyama, H.; Nakatani, K.; Tsuchida, A.; Yamamoto, M. Photoinduced DNA cleavage via electron transfer: demonstration that guanine residues located 5'to guanine are the most electron-donating sites. *J. Am. Chem. Soc.* **1995**, *117*, 6406–6407.

(20) Saito, I.; Nakamura, T.; Nakatani, K.; Yoshioka, Y.; Yamaguchi, K.; Sugiyama, H. Mapping of the hot spots for DNA damage by one-electron oxidation: efficacy of GG doublets and GGG triplets as a trap in long-range hole migration. *J. Am. Chem. Soc.* **1998**, *120*, 12686–12687.

(21) Sugiyama, H.; Saito, I. Theoretical studies of GG-specific photocleavage of DNA via electron transfer: significant lowering of ionization potential and 5‘-localization of HOMO of stacked GG bases in B-Form DNA. *J. Am. Chem. Soc.* **1996**, *118*, 7063–7068.

(22) Kumar, A.; Sevilla, M. D. Density functional theory studies of the extent of hole delo-

calization in one-electron oxidized adenine and guanine base stacks. *J. Phys. Chem. B* **2011**, *115*, 4990–5000.

(23) Rooman, M.; Wintjens, R. Sequence and conformation effects on ionization potential and charge distribution of homo-nucleobase stacks using M06-2X hybrid density functional theory calculations. *J. Biomol. Struct. Dyn.* **2014**, *32*, 532–545.

(24) Šponer, J.; Leszczyński, J.; Hobza, P. Nature of nucleic acid-base stacking: nonempirical ab initio and empirical potential characterization of 10 stacked base dimers. Comparison of stacked and H-bonded base pairs. *J. Phys. Chem.* **1996**, *100*, 5590–5596.

(25) Cabaleiro-Lago, E. M.; Rodríguez-Otero, J. On the Nature of σ–σ, σ–π, and π–π Stacking in Extended Systems. *ACS Omega* **2018**, *3*, 9348–9359.

(26) Josa, D.; Rodríguez-Otero, J.; Cabaleiro-Lago, E. M.; Rellán-Piñeiro, M. Analysis of the performance of DFT-D, M05-2X and M06-2X functionals for studying π π interactions. *Chem. Phys. Lett.* **2013**, *557*, 170–175.

(27) Barnett, R. N.; Cleveland, C. L.; Joy, A.; Landman, U.; Schuster, G. B. Charge migration in DNA: ion-gated transport. *Science* **2001**, *294*, 567–571.

(28) Bongiorno, A. Energy landscape of an electron hole in hydrated DNA. *J. Phys. Chem. B* **2008**, *112*, 13945–13950.

(29) Roca-Sanjuan, D.; Olaso-Gonzalez, G.; Rubio, M.; Coto, P. B.; Merchan, M.; Ferre, N.; Ludwig, V.; Serrano-Andres, L. DNA nucleobase properties and photoreactivity: Modeling environmental effects. *Pure Appl. Chem.* **2009**, *81*, 743–754.

(30) Poulos, R. C.; Olivier, J.; Wong, J. W. The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. *Nucleic Acids Res.* **2017**, *45*, 7786–7795.

(31) Møller, C.; Plesset, M. S. Note on an approximation treatment for many-electron systems. *Phys. Rev.* **1934**, *46*, 618.

(32) Sedlak, R.; Riley, K. E.; Řezáč, J.; Pitoňák, M.; Hobza, P. MP2. 5 and MP2. X: approaching CCSD (T) quality description of noncovalent interaction at the cost of a singfle CCSD iteration. *ChemPhysChem* **2013**, *14*, 698–707.

(33) Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* **2006**, *124*, 034108.

(34) Schwabe, T.; Grimme, S. Towards chemical accuracy for the thermodynamics of large molecules: new hybrid density functionals including non-local correlation effects. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398–4401.

(35) Brémond, E.; Adamo, C. Seeking for parameter-free double-hybrid functionals: The PBE0-DH model. *J. Chem. Phys.* **2011**, *135*, 024106.

(36) Dithcfield, R.; Hehre, W.; Pople, J. Self-consistent molecular-orbital methods. 9. Extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* **1971**, *54*, 724–728.

(37) Hariharan, P. C.; Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta* **1973**, *28*, 213–222.

(38) Wintjens, R.; Biot, C.; Rooman, M.; Liévin, J. Basis set and electron correlation effects on ab initio calculations of cation-π/H-bond stair motifs. *J. Phys. Chem. A* **2003**, *107*, 6249–6258.

(39) Rooman, M.; Liévin, J.; Buisine, E.; Wintjens, R. Cation–π/H-bond stair motifs at protein–DNA interfaces. *J. Mol. Biol.* **2002**, *319*, 67–76.

(40) Frisch, M. J. et al. Gaussian 16. 2016.

(41) Lu, X.-J.; Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **2003**, *31*, 5108–5121.

(42) Premilat, S.; Albiser, G. A new D-DNA form of poly (dA-dT). poly (dA-dT): an A-DNA type structure with reversed Hoogsteen pairing. *Eur. Biophys. J.* **2001**, *30*, 404–410.

(43) Arnott, S. Polynucleotide secondary structures: an historical perspective. 1999.

(44) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32*, 922–923.

(45) O'Neill, P.; Parker, A. W.; Plumb, M. A.; Siebbeles, L. D. Guanine modifications following ionization of DNA occurs predominantly via intra-and not interstrand charge migration: an experimental and theoretical study. *J. Phys. Chem. B* **2001**, *105*, 5283–5290.

(46) Donchev, A. G.; Taube, A. G.; Decolvenaere, E.; Hargus, C.; McGibbon, R. T.; Law, K.-H.; Gregersen, B. A.; Li, J.-L.; Palmo, K.; Siva, K.; Bergdorf, M.; Klepeis, J. L.; Shaw, D. E. Quantum chemical benchmark databases of gold-standard dimer interaction energies. *Sci. Data* **2021**, *8*, 1–9.

(47) Landrum, M. J.; Lee, J. M.; Riley, G. R.; Jang, W.; Rubinstein, W. S.; Church, D. M.; Maglott, D. R. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **2014**, *42*, D980–D985.

(48) Mattei, A. L.; Bailly, N.; Meissner, A. DNA methylation: a historical perspective. *Trends Genet.* **2022**,

(49) Close, D. M. Oxidative damage to cytosine: Implication for the study of radiation-induced damage to DNA. *J. Phys. Chem. B* **2003**, *107*, 864–867.

(50) Yermolina, M. V.; Papadantonakis, G. A. A computational investigation of cytosine and 5-methyl cytosine reactivity by means of ionization potentials and one specific methylation pathway. *Chem. Phys. Lett.* **2020**, *752*, 137544.

(51) Zhou, Y.; He, F.; Pu, W.; Gu, X.; Wang, J.; Su, Z. The impact of DNA methylation dynamics on the mutation rate during human germline development. *G3: Genes, Genomes, Genet.* **2020**, *10*, 3337–3346.

(52) Oeseburg, H.; de Boer, R. A.; van Gilst, W. H.; van der Harst, P. Telomere biology in healthy aging and disease. *Pflügers Arch.-European Journal of Physiology* **2010**, *459*, 259–268.

(53) Lohe, A. R.; Hilliker, A.; Roberts, P. Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. *Genetics* **1993**, *134*, 1149–1174.

(54) Deaton, A. M.; Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **2011**, *25*, 1010–1022.

(55) Li, X.; Peng, Y.; Ren, J.; Qu, X. Effect of DNA flanking sequence on charge transport in short DNA duplexes. *Biochemistry* **2006**, *45*, 13543–13550.

(56) Bredas, J.; Silbey, R.; Boudreaux, D.; Chance, R. Chain-length dependence of electronic and electrochemical properties of conjugated systems: polyacetylene, polyphenylene, polythiophene, and polypyrrole. *J. Am. Chem. Soc.* **1983**, *105*, 6555–6559.

(57) Janietz, S.; Bradley, D.; Grell, M.; Giebeler, C.; Inbasekaran, M.; Woo, E. Electrochemical determination of the ionization potential and electron affinity of poly (9, 9-dioctylfluorene). *Appl. Phys. Lett.* **1998**, *73*, 2453–2455.

(58) Tomek, K. J.; Volkel, K.; Indermaur, E. W.; Tuck, J. M.; Keung, A. J. Promiscuous molecules for smarter file operations in DNA-based data storage. *Nat. Commun.* **2021**, *12*, 1–10.

(59) Slavicek, P.; Winter, B.; Faubel, M.; Bradforth, S. E.; Jungwirth, P. Ionization energies of aqueous nucleic acids: Photoelectron spectroscopy of pyrimidine nucleosides and ab initio calculations. *J. Am. Chem. Soc.* **2009**, *131*, 6460–6467.

(60) Fernando, H.; Papadantonakis, G. A.; Kim, N. S.; LeBreton, P. R. Conduction-band-edge ionization thresholds of DNA components in aqueous solution. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 5550–5555.

(61) Cerón-Carrasco, J.; Requena, A.; Perpète, E.; Michaux, C.; Jacquemin, D. Theoretical study of the tautomerism in the one-electron oxidized guanine-cytosine base pair. *J. Phys. Chem. B* **2010**, *114*, 13439–13445.

(62) Cauët, E.; Valiev, M.; Weare, J. H. Vertical ionization potentials of nucleobases in a fully solvated DNA environment. *J. Phys. Chem. B* **2010**, *114*, 5886–5894.

(63) Close, D. M. Calculation of the ionization potentials of the DNA bases in aqueous medium. *J. Phys. Chem. A* **2004**, *108*, 10376–10379.