7644–7652 Nucleic Acids Research, 2012, Vol. 40, No. 16 Published online 16 June 2012
doi:10.1093/nar/gks519

Computational mapping reveals dramatic effect of Hoogsteen breathing on duplex DNA reactivity with formaldehyde

Tanggis Bohnuud[1], Dmitri Beglov[2], Chi Ho Ngan[2], Brandon Zerbe[2], David R. Hall[2], Ryan Brenke[2], Sandor Vajda[1,2,3,*], Maxim D. Frank-Kamenetskii[1,2,*] and Dima Kozakov[2,*]

[1]Graduate program in Bioinformatics, [2]Department of Biomedical Engineering and [3]Department of Chemistry, Boston University, Boston, MA 02215, USA

Received April 12, 2012; Revised May 8, 2012; Accepted May 9, 2012

ABSTRACT

Formaldehyde has long been recognized as a hazardous environmental agent highly reactive with DNA. Recently, it has been realized that due to the activity of histone demethylation enzymes within the cell nucleus, formaldehyde is produced endogenously, in direct vicinity of genomic DNA. Should it lead to extensive DNA damage? We address this question with the aid of a computational mapping method, analogous to X-ray and nuclear magnetic resonance techniques for observing weakly specific interactions of small organic compounds with a macromolecule in order to establish important functional sites. We concentrate on the leading reaction of formaldehyde with free bases: hydroxymethylation of cytosine amino groups. Our results show that in B-DNA, cytosine amino groups are totally inaccessible for the formaldehyde attack. Then, we explore the effect of recently discovered transient flipping of Watson–Crick (WC) pairs into Hoogsteen (HG) pairs (HG breathing). Our results show that the HG base pair formation dramatically affects the accessibility for formaldehyde of cytosine amino nitrogens within WC base pairs adjacent to HG base pairs. The extensive literature on DNA interaction with formaldehyde is analyzed in light of the new findings. The obtained data emphasize the significance of DNA HG breathing.

INTRODUCTION

Although formaldehyde has long been considered as a chemical hazard capable of damaging DNA, only in recent years it has been realized that formaldehyde forms endogenously within the cell nucleus. It has been established that a process of histone demethylation by recently discovered histone demethylation enzymes inevitably leads to the accumulation of formaldehyde within the cell nucleus, in direct vicinity of DNA [reviewed in (1)]. This may lead to the extensive DNA damage. Most recently, the lethal consequences of the endogenous formaldehyde accumulation in the cell nucleus have been experimentally demonstrated for cells deficient in both the formaldehyde digesting enzyme ADH5 and the enzyme that repairs DNA damages inflicted by formaldehyde (2). These findings put into the forefront the issue of DNA reactivity with formaldehyde and constitute a major motivation of the present work.

Our focus here is not the well-known effect of DNA cross-linking by formaldehyde, which is a very important but very rare event, but the formaldehyde's chemical reaction with amino and imino groups of DNA bases. Previous attempts to fully understand the reactivity of duplex DNA with respect to formaldehyde on the basis of extensive experimental and theoretical studies (3–10) left a pivotal question unanswered. Although it was convincingly shown that fluctuational openings of base pairs play an important role in the process, there was also a strong indication that a major reaction, hydoxymethylation of the cytosine amino group, proceeds without full base pair openings (8,9). Does it mean that formaldehyde attacks the cytosine amino group directly at the bottom of the B-DNA major groove? This issue could not be resolved before. Herein, we address this question computationally.

The main tool of our analysis is computational solvent mapping, a method originally developed for identification and characterization of binding hot spots on proteins, i.e. smaller regions of the binding site that are major

*To whom correspondence should be addressed. Tel: +1 617 353 4842; Fax: +1 617 353 6766; Email: midas@bu.edu
Correspondence may also be addressed to Sandor Vajda. Tel: +1 617 353 4757; Fax: +1 617 353 6766; Email: vajda@bu.edu
Correspondence may also be addressed to Maxim D. Frank-Kamenetskii. Tel: +1 617 353 8498; Fax: +1 617 353 8501; Email: mfk@bu.edu

contributors to the binding free energy (11). It was shown earlier by X-ray crystallography (12) and nuclear magnetic resonance (NMR) (13) that in proteins the hot spots bind small organic molecules of various sizes and shapes and that the number of bound 'probe' molecules predicts the potential importance of the site for ligand binding. We have developed the FTMap algorithm to identify such probe binding sites computationally (11). FTMap samples probe–protein interactions on a dense grid, finds favorable positions using empirical energy functions, clusters the conformations and ranks the clusters on the basis of the average energy. For each probe, six bound probe clusters with the lowest mean interaction energies are retained. The clusters from the different probe types are then clustered into consensus clusters. The positions of the consensus clusters, termed consensus sites (CSs), define the binding hot spots where multiple probe molecules bind, identifying the regions most likely to bind small ligands. The CSs are ranked on the basis of the number of probe clusters they incorporate. The largest CSs represent the most important hot spots, in good agreement with the results of screening libraries of small molecules using X-ray crystallography (12) or NMR (13) for detection.

In addition to exploring computationally the B-DNA reactivity with respect to formaldehyde, we also check how the reactivity can be affected by a recently found new mode of the DNA double helix breathing (14), which was overlooked before [see commentaries (15,16)]. According to these new data, classical Watson–Crick (WC) base pairs spontaneously flip into unusual Hoogsteen (HG) pairs under most normal ambient conditions (Figure 1). The probability of WC-to-HG transition proved to be about three orders of magnitude higher than that of the previously extensively studied breathing mode of DNA consisting in base pair opening, which occurs with the probability of about 10^{-5} (10,14,16–18). Previously, flipping of WC pairs into HG pairs was demonstrated by X-ray crystallography within some DNA–protein complexes, most notably in case of p53 binding to DNA (19). The HG pair is expected to inflict much milder violation on the canonical DNA structure than base pair openings. Still, although stacking interaction between the HG pair and adjacent WC pairs is preserved (Figure 1B), the shorter distances between glycosidic bonds in HG pairs may entail a significant perturbation in the regular structure of the canonical all-WC double helix (B-DNA).

In this work, we describe an extension of the FTMap algorithm (11) to nucleic acids and determine the binding hot spots of DNA structures with WC and HG base pairs. In contrast to proteins, no experimentally determined hot spots are known for DNA structures, but our results agree well with the binding sites of drug-like molecules in the minor groove. In addition to the hot spots in the minor groove, the mapping also reveals a hot spot in the major groove. Once the existence of a hot spot in the major groove was established, we focused on the number and orientation of carbonyl groups from various probes near the amino nitrogen atom of cytosine and also performed mapping calculations using formaldehyde as the additional probe.

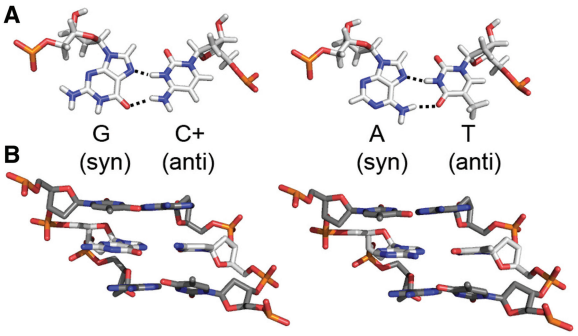

Figure 1. (**A**) HG AT and GC+ base pairs between anti-parallel DNA strands. Note that purines in both pairs must assume the unusual syn conformation. Hydrogen bonds are shown as dashed lines. (**B**) Incorporation of HG pairs into the regular double helix. Only HG pairs (light gray) and two adjacent WC pairs (dark gray) are shown. HG pairs are nicely stacked with neighboring WC pairs.

Our results show that in case of B-DNA, the cytosine amino groups must be totally inaccessible for formaldehyde reaction. HG pairing increases the accessibility of amino groups of cytosines, participating in WC pairing adjacent to HG pairs, to small molecules, changes the local binding properties of the DNA and dramatically increases the exposure of the amino groups to formaldehyde reaction. In case of regular B-DNA with only WC pairs, no clustering of formaldehyde molecules in close vicinity of cytosine amino groups is observed. In contrast, the incorporation of an HG pair into B-DNA results in very substantial clustering of formaldehyde molecules in close vicinity of the amino group nitrogen of a cytosine participating in WC pairing but adjacent to the HG pairs. A substantial portion of the formaldehyde molecules in the cluster have orientation favorable for chemical attack leading to hydroxymethylation of the amino group. Our results indicate that some aspects of DNA reaction with formaldehyde, which have remained a mystery over the past 25 years, may be explained on the basis of recently discovered HG breathings of base pairs (14).

MATERIALS AND METHODS

Computational solvent mapping

The FTMap algorithm (11) was extended for application to nucleic acids and is available as a beta version at http://ftmap.bu.edu/beta. The algorithm includes five computational steps as follows:

Step 1: Soft rigid body docking of probe molecules. DNA structures were downloaded from the Protein Data Bank (20). All bound ligands, ions and water molecules were removed. We used 16 small molecules as probes (ethanol, isopropanol, *tert*-butanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide and *N*,*N*-dimethylformamide). For each probe, millions of docked conformations were sampled by soft rigid body docking based on

fast Fourier transform correlation approach (11). The method performs exhaustive evaluation of an energy function in the discretized 6D space of mutual orientations of the protein (receptor) and a small molecular probe (ligand). The center of mass of the DNA was fixed at the origin of the coordinate system. The translational space was represented as a grid of 0.8 Å displacements of the probe center of mass, and the rotational space is quasi-uniformly sampled using 500 rotations of the probe (11). The energy expression includes a stepwise approximation of the van der Waals energy with attractive and repulsive contributions, a hydrogen bonding term and an electrostatics/solvation term based on nonlinear Poisson–Boltzmann continuum calculation. Details of the last two terms are included in Supplementary Methods. The values of the potential were smoothly interpolated up to the first derivative using a tricubic interpolation algorithm.

Step 2: Minimization and rescoring. For each probe type, the 2000 lowest energy DNA–probe complexes, generated in Step 1, were refined by off-grid energy minimization during which the DNA atoms are held fixed, whereas the atoms of the probe molecules are free to move. The energy function includes the bonded and van der Waals terms of CHARMM force field (21) and an electrostatic interaction term using the Poisson–Boltzmann potential generated in Step 1.

Step 3: Clustering and ranking. The minimized probe conformations from Step 2 are grouped into clusters using a simple greedy algorithm. The lowest energy structure is selected and the structures within 4 Å RMSD are joined in the first cluster. The members of this cluster are removed, and the next lowest energy structure is selected to start the second cluster. This step is repeated until the entire set is exhausted. Clusters with less than 10 members are excluded from consideration. The retained clusters are ranked by their Boltzmann-averaged energies. Six clusters with the lowest average free energies are retained for each probe.

Step 4: Determination of consensus clusters. To determine CSs, i.e. regions on the DNA where clusters of different probes overlap, the probe cluster centers are clustered using the distance between their centers of mass with a 4 Å clustering radius. The resulting consensus clusters are ranked based on the number of their clusters. Duplicate clusters of the same type are considered in the count.

Step 5: Functional clustering of CO groups. To determine the regions favorable for interactions with the carbonyl group, we applied the clustering algorithm described in Step 3 to the CO groups of all acetone, acetaldehyde, acetamide and urea probes in the CS located in the major groove.

Docking assisted by functional clustering

The initial docking was performed using the FTMap algorithm, simply considering formaldehyde as an additional probe. The formaldehyde positions were clustered and ranked on the basis of their Boltzmann-averaged energies, and six clusters with the lowest average energy were retained. The formaldehyde cluster overlapping with the highest population cluster of CO groups was selected as the most likely region with preference for formaldehyde binding and the lowest energy formaldehyde pose within the cluster was used to represent the most likely bound position.

RESULTS

Identification of the binding hot spots in B-DNA

Although the existence of binding hot spots has been well established for proteins both experimentally (12,13) and computationally (11), no such analyses have been performed for DNA. To see whether similar hot spots exist in DNA, we have mapped a ligand-free B-DNA structure of the $d(CGCGAATTCGCG)_2$ dodecamer (known as the Dickerson–Drew dodecamer, PDB code 1BNA), because both unbound and ligand-bound structures are available for this sequence. Figure 2A shows the main hot spot from mapping, defined by the largest CS (CS1), which includes 23 probe clusters (shown in cyan, in mesh representation). In addition to the main hot spot, Figure 2A also shows the drug distamycin A bound in the minor groove of B-DNA (PDB code 267D). Figure 2B shows the nine largest CSs that bind probe clusters ranging from 23 in CS1 to 10 in CS9. CSs CS1 (cyan), CS3 (18 probes clusters, yellow) and CS5 (15 probe clusters, purple) overlap the site that binds a variety of drugs. Sites CS2 (20 clusters), CS4 (16 clusters), CS6 (14 clusters), CS7 and CS8 (both 12 clusters) are also in the minor groove, extending the major drug binding site on both sides and are all shown colored in wheat. Finally, CS9 (with 10 probe clusters, shown in blue) is located in the major groove. Figure 2C is a close-up of CS1, CS3 and CS5 in the minor groove and shows the overlap of these hot spots with three drugs (distamycin A, netropsin and diamidine) bound to DNA. That the main hot spot and the nearby substantial hot spots identified in unbound DNA structures are in the preferred drug binding site supports the hypothesis that the concept of binding hot spots as sites mostly contributing to the binding free energy is transferable to nucleic acids. We note that, based on mapping with 16 different types of probe molecules, a druggable site on a protein comprises a hot spot binding at least 16 probe clusters and one or two adjacent hot spots within reach of a drug-sized molecule (22). The site in the minor groove clearly satisfies these conditions, in agreement with its ability to bind drugs with high affinity.

Note that although the $d(CGCGAATTCGCG)_2$ dodecamer is a palindrome, i.e. its sequence obeys a dyad symmetry, the X-ray structure of the molecule does not obey the dyad symmetry and, as a result, the distribution of ligands around the dodecamer is asymmetrical (Figure 2). This reflects the well-known effect of the DNA crystal packing on the X-ray structure. So the fact that our computations do not yield the same cluster as CS9 in a position symmetrical to the CS9 spot is not unexpected. A more detailed FTMap analysis reveals a weaker cluster in the vicinity of the spot connected by

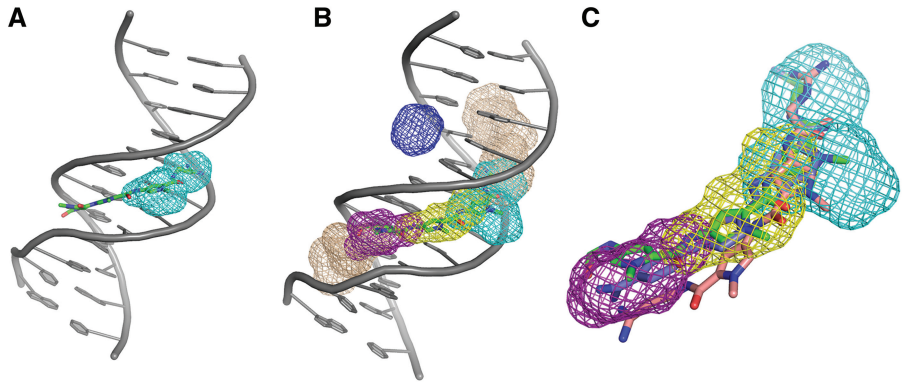

Figure 2. Mapping of ligand-free B-DNA (PDB code 1BNA). (**A**) The largest CS1 (with 23 probe clusters, shown as cyan mesh) overlaps with a fragment of the drug distamycin A (shown as sticks in green) in the minor groove. (**B**) CS1 (cyan), CS3 (yellow) and CS5 (purple) cover the entire drug binding site. CS2, CS4 and CS6–CS8 (all shown in wheat) extend the site in both directions. CS9 (with 10 probe clusters, shown in blue) is the largest hot spot located in the major groove. (**C**) Close-up of CS1, CS3 and CS5 overlapping the bound structures of three drugs shown as sticks. The names and color codes of the drugs and the PDB codes of the structures are as follows: distamycin A, green, 267D; netropsin, salmon, 101D and diamidine, blue, 1VZK.

the dyad symmetry with the CS9 spot (data not shown). Thus, in this article, in relation to formaldehyde reactivity, we focus on the CS9 hot spot in the major groove.

Influence of HG pairing on surface accessibility of the cytosine amino group

Al-Hashimi and coworkers (14) observed transient sequence-specific excursions away from WC base pairing at CA steps inside canonical duplex DNA toward low-populated, short-lived HG base pairs. As will be described in the 'Discussion', reaction with the cytosine amino group is of primary interest for exploring possible effects of HG pairing on formaldehyde reaction with DNA. Increased reactivity most likely requires increased surface accessibility (SA) of the cytosine's amino nitrogen. We have calculated the accessible surface area of the nitrogen for all possible triplets of bases that include the CA step. For these calculations, we have built structures with CA in the center based on structure of the double-stranded nanomer 5′-ATGTAA$_H$TTC-3′ (PDB code 1K61, chain E), where the subscript H indicates HG pairing. In Table 1, we underline the cytosine considered in the accessibility calculation. The results show that the accessibility of the amino nitrogen increases when C is at the 5′ side of a HG A•T. Furthermore, the SA of the amino nitrogen reaches its maximum when a purine (A or G) is at the 5′ side of the cytosine. As HG pairing most significantly increases SA of the cytosine's amino nitrogen in the triplets G$\underline{C}$A$_H$ and A$\underline{C}$A$_H$, we created structures including these triplets in both WC and HG conformations for mapping and docking studies.

Mapping of DNA structures containing ACA or GCA triplets with HG or WC pairing

Mapping results for HG and WC CA steps are similar for the minor groove, but qualitatively different in the major groove. Flipping to HG pairing does not affect the largest consensus cluster, which remains in the minor groove (Figure 2). Herein, we focus on CS9 in the major groove. As shown in the left panels of Figure 3, there are consensus clusters near the cytosine's amino nitrogen atom with both HG pairs (Figure 3A) and WC pairs (Figure 3C). Although these figures show only the lowest energy structures for each cluster of the bound probes, even visual inspection reveals that the probes, and particularly their polar functional groups, shift toward the amino nitrogen atom (shown as a blue sphere) in the HG structure (Figure 3A). This difference is clearly seen in Figure 3E, which shows the number of probes in spherical intervals around the nitrogen atom. For the HG pair, maximum probe density occurs at 3.6 Å from the nitrogen atom, whereas for the WC pair the maximum shifts to 4.4 Å. Thus, for the HG pair, the entire consensus cluster shifts by about 1 Å toward the reactive group. The number of probe clusters in the consensus clusters also changes, albeit less visibly than the probe density: we observe 15 probe clusters for HG and only 13 for WC pairing. Thus, we conclude that HG pairing not only increases the solvent accessibility of the cytosine amino group but also creates a more preferential site for the binding of small molecules.

Since we are specifically interested in the formaldehyde reaction, we selected the probes with a carbonyl functional group to see whether there is a strong pattern of the carbonyl (CO) moiety binding in the vicinity of the amino nitrogen. Among the probes clustering in this region, four molecules contained a CO moiety (acetone, acetaldehyde, acetamide and urea). We have selected these probes contained in the consensus cluster in the major groove and clustered the positions of their carbonyl groups using a 1.5 Å clustering radius (see 'Materials and Methods' section). Figure 3B and D shows the CO group closest to the geometric center of the highest population cluster formed by the CO groups for HG and WC pairing, respectively. As with total probe density, the CO distribution shifts toward the amino nitrogen atom in the structure with HG pairing. Figure 3F shows the number of CO groups in spherical intervals around the amino nitrogen atom for both HG and WC pairings. For the

Table 1. SA of C-N4 atom in DNA structures with HG and WC pairing

Sequence[a,b]	SA of C-N4 (Å^2)		
	HG	WC	HG–WC
G$\underline{C}$A$_H$	**19.71**	**14.93**	**4.78**
A$\underline{C}$A$_H$	**19.38**	**15.05**	**4.32**
C$\underline{C}$A$_H$	15.60	12.27	3.33
T$\underline{C}$A$_H$	14.50	8.59	5.90
T$\underline{C}_H$A	7.63	8.46	−0.83
A$\underline{C}_H$A	12.39	11.75	0.63
G$\underline{C}_H$A	12.24	11.63	0.61
C$\underline{C}_H$A	9.92	9.23	0.69
$\underline{C}\underline{C}_H$A	14.87	16.51	−1.63

[a]Subscript H indicates base with HG pairing.
[b]SA is calculated for C underlined.
C-N4, atom N4 of Cytosine; HG–WC, difference in SA from switching WC to HG pairing. The figures for two triplets with largest SA values in case of HG pairing are emboldened.

HG pairing, the CO groups form a compact cluster at 3.6 Å from the nitrogen atom. In contrast, for DNA with WC pairing, the density distribution is very diffuse, with a minor peak at 4.4 Å, but a large fraction of probes with CO groups is >5 Å away from the amino nitrogen. The same statistics have also been calculated for a structure with the GCA triplet (Supplementary Figures S1 and S2). The results are qualitatively similar to the ones for the ACA triplet. For HG pairing, both the peak probe density and peak CO group density are at 3.6 Å, about 0.8 Å closer to the amino nitrogen atom than for WC pairing. The maximum values of both the probe and the CO group density for GCA triplet with the HG pairing are slightly lower than the corresponding values observed in DNA with an ACA triplet. Thus, the transition to HG pairing provides better access for small molecules to the cytosine amino nitrogen atom in DNA with an ACA triplet than in a DNA with GCA triplet, but changing from WC to HG pairing has a strong effect in both cases.

Distribution of formaldehyde positions around the amino nitrogen of cytosine

In order to study the interactions between formaldehyde and DNA with WC or HG pairing, we considered DNA structures that include the ACA triplet, docked formaldehyde to the structures without any constraints on the binding site and clustered the low-energy poses using a 2.5 Å clustering radius (see 'Materials and Methods' section). As expected for small ligands (18), low-energy formaldehyde clusters were located in all hot spots shown in Figure 2, including the minor and major grooves. To find the most likely formaldehyde position in the major groove and in a region with preference for a carbonyl group, we applied a recently introduced filtering algorithm (22) and selected the formaldehyde pose that overlaps with the highest population cluster of CO groups identified by the mapping as described in the previous section. Since we used a small (1.5 Å) radius, the CO cluster is small, and this approach selects a single formaldehyde pose, discarding potentially spurious poses that arise from docking. Figure 4 shows the density of low-energy formaldehyde poses (in a transparent surface representation) and the selected formaldehyde pose for HG (purple) and WC (green) pairing, indicating that both the density of poses and the lowest energy formaldehyde position are shifted closer to the amino nitrogen by ∼1 Å upon the WC to HG flipping. Figure 4B shows the distributions of formaldehyde positions in the hot spot around the amino nitrogen for DNA with HG and WC pairing. According to these results, for the ACA triplet the distribution shifts toward the amino nitrogen in the HG pair by >1 Å. Thus, the formation of an HG pair next to the C•G pair entails the formation of a formaldehyde cluster located in the major groove in close proximity to the cytosine amino group's nitrogen. As a result, a significant fraction of molecules in the cluster become very well positioned to attack the nitrogen lone pair with formaldehyde's carbon atom, exactly the mechanism for the amino group hydroxymethylation, according to *ab initio* quantum mechanical data (23,24). For the GCA triplet, the shift of the formaldehyde cluster toward the amino nitrogen atom is somewhat smaller, but it is still substantial, confirming that the increased access of formaldehyde to the amino nitrogen atom remains valid, independently of the nature of the HG pair adjacent to the cytosine under consideration.

To further prove that only amino nitrogens next to an HG pair are accessible for formaldehyde molecules, we have docked formaldehyde to four additional B-DNA crystal structures of good resolution, all in WC pairing (Supplementary Table S1). The results show that formaldehyde molecules cluster predominantly in the minor groove of the DNA double helix, whereas the cytosine amino group is located in the major groove. In all cases, either no cluster was observed in the major groove at all or a cluster similar to one in Figure 4A was formed, always at least 4 Å apart from the cytosine amino nitrogen (data not shown). This makes it impossible for formaldehyde molecules within the cluster to attack the nitrogen lone pair, which is a necessary condition for the amino group hydroxymethylation (23,24). This demonstrates that the reactive CO group interacts with the amino nitrogen of the adjacent cytosine only in DNA structures with HG pairing. This result extends to other aldehydes. We have performed CO density assisted docking of acetaldehyde to HG pairs, which revealed that the results are almost identical with the ones we have obtained for formaldehyde (Supplementary Figure S3).

DISCUSSION

The substantial increase in the availability of DNA structures in the Protein Data Bank (20), coupled with improved modeling and visualization tools (25–27), demonstrated that interactions between DNA and other molecules, including proteins and small ligands, frequently occur in non-canonical regions of the DNA (28–31). In particular, we already mentioned that flipping of WC pairs into HG pairs was observed in a complex of DNA

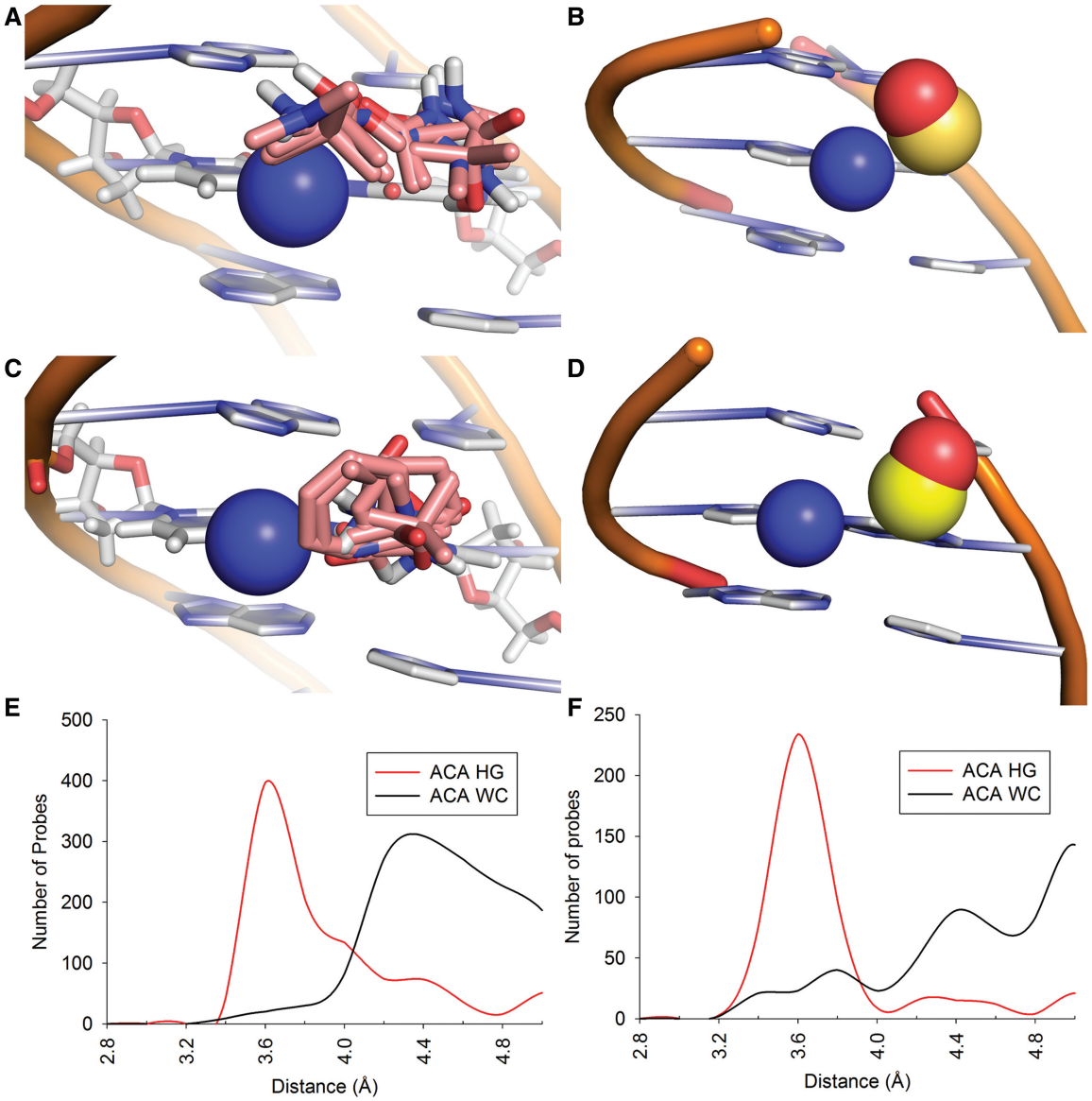

Figure 3. Consensus clusters of 16 probes and clustering of carbonyl group around the amino nitrogen of cytosine of DNA with an ACA triplet. The blue sphere indicates the cytosine's amino nitrogen atom. For each probe cluster, only one representative probe pose is shown. (**A**) Consensus cluster of all probes in the major grove of DNA with HG pairing. (**B**) Carbonyl moiety closest to the geometric center of the cluster formed by all CO groups in probes acetone, acetaldehyde, acetamide and urea for DNA with HG pairing (**C**) Consensus cluster in the major grove of DNA with WC pairing. (**D**) Carbonyl moiety closest to the geometric center of the cluster formed by all CO groups for DNA with WC pairing. (**E**) Number of probe atoms in spherical intervals around the amino nitrogen atom. (**F**) Number of CO groups in spherical intervals around the amino nitrogen atom.

with the p53 protein (19). Thus, it is expected that the spontaneous transition to HG pairing, as demonstrated by Al-Hashimi and coworkers (14), will influence the binding properties. Herein, we show, for the first time, that the HG pair formation indeed affects binding as well as the reactivity of DNA. We have chosen formaldehyde as a chemical agent and the cytosine amino group as the target. Our data show a dramatic effect of HG base pairs on the accessibility for formaldehyde attack of the amino group of cytosines adjacent to HG pairs. Indeed, in the case of the all-WC duplex, either no substantial cluster is formed in the major groove or the closest formaldehyde cluster is located in the major groove farther than 4 Å from the amino group's nitrogen. By contrast, when an HG base pair is next to the WC GC pair under consideration, the cluster forms near the cytosine amino group and a significant portion of formaldehyde molecules in the cluster have orientation favorable for attacking the nitrogen lone pair, which is the initial step of hydroxymethylation. Thus, our hot spot mapping data indicate that while the all-WC duplex must be totally non-reactive for formaldehyde with respect to the cytosine amino groups, the formation of HG pairs makes these groups' hydroxymethylation very much possible.

How do our results, together with the HG breathing concept (14), fit the extensive literature on DNA reaction with formaldehyde? To answer this question, let us first briefly summarize the picture of DNA reaction with

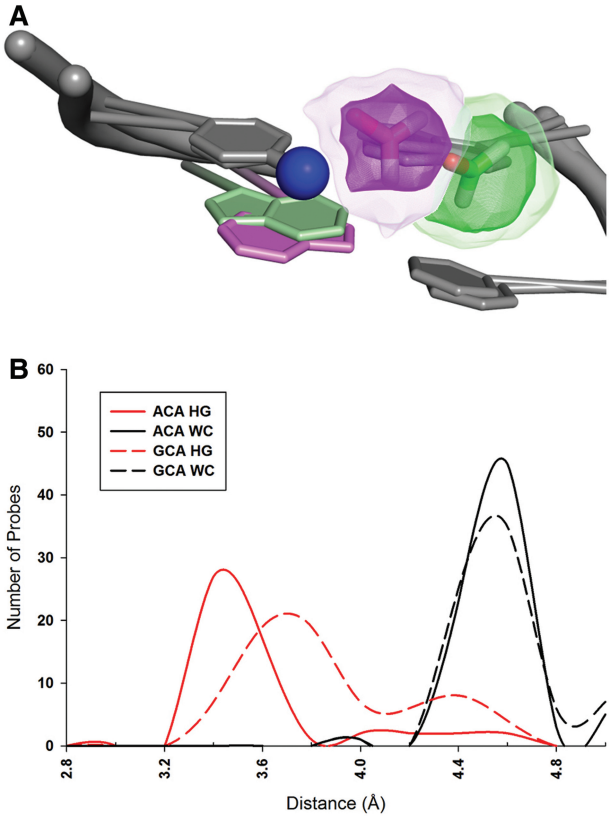

Figure 4. Formaldehyde positioning near the amino nitrogen atom of cytosine. (**A**) Superimposed formaldehyde (sticks) positioning near the amino nitrogen atom (blue sphere) of cytosine in DNA with HG pairing (purple) versus WC pairing (green) at the adjacent base pair. The density map shows distribution of low-energy formaldehyde positions. (**B**) Distributions of formaldehyde positions around the amino nitrogen atom for DNA with HG and WC pairing.

formaldehyde as it emerged on the basis of extensive experimental and theoretical studies made over two decades ago (3–10).

Formaldehyde hydroxymethylates amino and imino groups of DNA bases. However, the rate constants and the equilibrium constants of the reaction significantly differ for different groups (5,6):

(i) The guanine imino group is not hydroxymethylated.
(ii) The thymine imino group's reaction with formaldehyde is characterized by very high forward and reverse kinetic constants, which both strongly depend on pH, while their ratio, the equilibrium constant, is pH-independent and is much smaller than corresponding constants for amino groups.
(iii) The cytosine amino group has the largest forward kinetic constant and the largest equilibrium constant of reaction thus playing the leading role among amino groups.
(iv) Although reactions with guanine and adenine amino groups have to be allowed for in a quantitative theoretical analysis, they are much less important than the reaction with cytosine amino group and can be neglected within the first approximation.

In WC pairs, imino groups are totally inaccessible for reaction. Therefore, the imino reaction may proceed exclusively due to base pair breathing. Traditionally, it has been assumed that only base pair openings make imino group's hydroxymethylation possible. Discovery of the HG breathing does not change the situation. Indeed, the thymine imino group remains inaccessible in the HG A•T pair (Figure 1). Although in case of the HG G•C^+ pair, the guanine imino group does not participate in the inter-base hydrogen bonding, this is immaterial as guanine imino group does not react with formaldehyde anyway (see (i) above). Thus, only full breathing, consisting of base pair openings, can allow imino group hydroxymethylation. And, obviously, the corresponding A•T base pair must remain open as long as the imino group remains hydroxymethylated. We conclude that with respect to reaction with imino groups, the traditional picture based on only one breathing mode (base pair openings (7)) remains unchanged when accounting for HG breathing.

Let us turn now to the amino group reaction. Due to item (iii) above, we will consider only the reaction with cytosine. In contrast with imino groups, reaction with amino groups does not prevent the reformation of WC base pairs, although such pairs containing hydroxymethylated amino group are weakened (32). Now, we approach the central point of our analysis. Does reaction with amino groups necessarily require 'deep' breathing (base pair openings) or it is possible otherwise? Since, in contrast to imino groups, cytosine amino groups are not buried deep inside the double helix but are located at the bottom of the major groove, one can assume an 'outside' reaction of formaldehyde with the cytosine amino group, which somehow proceeds in the all-WC B-DNA although much slower than in open pairs (9,10,33). But the results presented here make this reaction route very unlikely: they do not leave any possibility for formaldehyde molecules to attack the cytosine amino group in all-WC B-DNA. At the same time, detailed computer simulations of the kinetics of DNA reaction with formaldehyde showed that the overall reaction was significantly slower than in experiment if the reaction with amino groups could proceed only via full base pair openings (8,9). That is why the 'outside' reaction was invoked in the first place ((9,10); note that the 'outside' reaction was also detected by Demidov in specially designed DNA melting experiments (33)).

It was understood that the 'outside reaction' could be a result of some base pair 'breathings' other than full base pair openings (9,10,33), but the nature of these fluctuations has remained a total mystery over the past 25 years. The results presented here suggest that the 'outside' reaction may proceed due to HG breathing, which makes amino groups of cytosines adjacent to transiently formed HG base pairs readily accessible for formaldehyde attack. The slow-down of the 'outside' reaction when compared with unsheltered cytosines (which is about 10^{-2}) (9,10,33) may be roughly considered as an estimate of the probability of HG breathing, which is in reasonable agreement with the HG breathing probability based on NMR data (14). Of course, this is only a very rough

estimate since it is done under several assumptions, first of all that the rate of reaction for cytosine participating in WC pairing is the same as for free cytosine.

Based on our current findings and on extensive computer simulations of the kinetics of DNA unwinding by formaldehyde (8–10), we arrive at the following picture of the process of DNA interaction with formaldehyde, which quantitatively explains all its features. When duplex DNA is exposed to formaldehyde, no hydroxymethylation is possible for the 'ground' state of DNA, which is the all-WC B-form. The process of hydroxymethylation proceeds via two types of fluctuations: transient base pair openings and transient flipping of WC pairs into HG pairs. The former events lead to a very fast formation of a stationary fraction of open A•T base pairs with hydroxymethylated thymine imino groups (see item (ii) above and computer simulations in (8)), whereas the latter events lead to accumulation of substantially weakened G•hmC pairs, where hmC are cytosines hydroxymethylated by their amino groups. Adenines in open A•T pairs as well as cytosines and adenines in the adjacent base pairs become easily accessible to hydroxymethylation by their amino groups because they are not protected any more by stacking with adjacent base pairs. In parallel, accumulation of weak G•hmC pairs results in local melting of weakened regions in DNA making all bases in the region fully accessible for reaction with formaldehyde. Eventually, accumulation of hydroxymethylated amino groups, of both cytosine and adenine, throughout DNA results in full separation of the DNA strands.

As shown by our results, computational mapping provides a highly sensitive tool for determining the effects of DNA conformational changes on the binding of small molecules. The method has been used in similar fashion for the comparison of protein binding sites in different environments (22,34). Mapping exhaustively explores the rotational/translational space of each probe molecule on a dense grid, and thus in principle provides detailed information on the energy landscape. Since the number of conformations generated is close to a million, we can retain and analyze only a number of the lowest energy clusters, which describe the low-energy regions of the landscape for each probe. However, these are the regions that are most likely to bind the probes, and the distributions of low-energy probe positions provide more exhaustive information on the energy landscape than methods generating a finite number of trajectories (e.g. Monte Carlo, molecular dynamics or Brownian dynamics).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figures 1–3, Supplementary Methods and Supplementary Reference [35].

FUNDING

National Institute of General Medical Sciences [GM064700]. Funding for open access charge: National Institute of General Medical Sciences.

Conflict of interest statement. None declared.

REFERENCES

1. Mosammaparast,N. and Shi,Y. (2010) Reversal of histone methylation: biochemical and molecular mechanisms of histone demethylases. *Annu. Rev. Biochem.*, **79**, 155–179.
2. Rosado,I.V., Langevin,F., Crossan,G.P., Takata,M. and Patel,K.J. (2011) Formaldehyde catabolism is essential in cells deficient for the Fanconi anemia DNA-repair pathway. *Nat. Struct. Mol. Biol.*, **18**, 1432–1434.
3. Lazurkin,Y.S., Frank-Kamenetskii,M.D. and Trifonov,E.N. (1970) Melting of DNA: its study and application as a research method. *Biopolymers*, **9**, 1253–1306.
4. Utiyama,H. and Doty,P. (1971) Kinetic studies of denaturation and reaction with formaldehyde on polydeoxyribonucleotides. *Biochemistry*, **10**, 1254–1264.
5. McGhee,J.D. and von Hippel,P.H. (1975) Formaldehyde as a probe of DNA structure. 1. Reaction with exocyclic amino groups of DNA bases. *Biochemistry*, **14**, 1281–1296.
6. McGhee,J.D. and von Hippel,P.H. (1975) Formaldehyde as a probe of DNA structure. 2. Reaction with endocyclic imino groups of DNA bases. *Biochemistry*, **14**, 1297–1303.
7. Lukashin,A.V., Vologodskii,A.V., Frank-Kamenetskii,M.D. and Lyubchenko,Y.L. (1976) Fluctuational opening of the double helix as revealed by theoretical and experimental study of DNA interaction with formaldehyde. *J. Mol. Biol.*, **108**, 665–682.
8. Frank-Kamenetskii,M.D. (1983) Fluctuational motility of DNA. *Mol. Biol.*, **17**, 639–652.
9. Frank-Kamenetskii,M.D. (1985) Fluctuational motility of DNA. In: Clemeti,E., Corongiu,G., CG, Sarma,M.H. and Sarma,R.H. (eds), *Structure and Motion: Membranes, Nucleic Acids and Proteins*. Adenine Press, NY, pp. 417–432.
10. Frank-Kamenetskii,M. (1987) How the double helix breathes. *Nature*, **328**, 17–18.
11. Brenke,R., Kozakov,D., Chuang,G.-Y., Beglov,D., Hall,D., Landon,M.R., Mattos,C. and Vajda,S. (2009) Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics*, **25**, 621–627.
12. Mattos,C. and Ringe,D. (1996) Locating and characterizing binding sites on proteins. *Nat. Biotechnol.*, **14**, 595–599.
13. Liepinsh,E. and Otting,G. (1997) Organic solvents identify specific ligand binding sites on protein surfaces. *Nat. Biotechnol.*, **15**, 264–268.
14. Nikolova,E.N., Kim,E., Wise,A.A., O'Brien,P.J., Andricioaei,I. and Al-Hashimi,H.M. (2011) Transient Hoogsteen base pairs in canonical duplex DNA. *Nature*, **470**, 498–502.
15. Honig,B. and Rohs,R. (2011) Biophysics: flipping Watson and Crick. *Nature*, **470**, 472–473.
16. Frank-Kamenetskii,M.D. (2011) DNA breathes Hoogsteen. *Artif. DNA*, **2**, 1–3.
17. Gueron,M., Kochoyan,M. and Leroy,J.L. (1987) A single mode of DNA base-pair opening drives imino proton exchange. *Nature*, **328**, 89–92.
18. Krueger,A., Protozanova,E. and Frank-Kamenetskii,M.D. (2006) Sequence-dependent base pair opening in DNA double helix. *Biophys. J.*, **90**, 3091–3099.
19. Kitayner,M., Rozenberg,H., Rohs,R., Suad,O., Rabinovich,D., Honig,B. and Shakked,Z. (2010) Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.*, **17**, 423–429.
20. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
21. Brooks,B.R., Brooks,C.L., Mackerell,A.D., Nilsson,L., Petrella,R.J., Roux,B., Won,Y., Archontis,G., Bartels,C., Boresch,S. *et al.* (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, **30**, 1545–1614.
22. Kozakov,D., Hall,D.R., Chuang,G.-Y., Cencic,R., Brenke,R., Grove,L.E., Beglov,D., Pelletier,J., Whitty,A. and Vajda,S. (2011) Structural conservation of druggable hot spots in protein–protein interfaces. *Proc. Natl Acad. Sci. USA*, **108**, 13528–13533.

23. Hall,N.E. and Smith,B.J. (1998) High-level ab initio molecular orbital calculations of imine formation. *J. Phys. Chem. A*, **102**, 4930–4938.
24. Liao,R.Z., Ding,W.J., Yu,J.G., Fang,W.H. and Liu,R.Z. (2007) Water-assisted transamination of glycine and formaldehyde. *J. Phys. Chem. A*, **111**, 3184–3190.
25. Petrey,D. and Honig,B. (2003) GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol.*, **374**, 492–509.
26. Jo,S., Vargyas,M., Vasko-Szedlar,J., Roux,B. and Im,W. (2008) PBEQ-Solver for online visualization of electrostatic potential of biomolecules. *Nucleic Acids Res.*, **36**, 270–275.
27. Zheng,G., Colasanti,A.V., Lu,X.-J. and Olson,W.K. (2010) 3DNALandscapes: a database for exploring the conformational features of DNA. *Nucleic Acids Res.*, **38**, 267–274.
28. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.
29. Gao,M. and Skolnick,J. (2009) A threading-based method for the prediction of DNA-binding proteins with application to the human genome. *PLoS Comput. Biol*, **5**, e1000567.
30. Ma,D.-L., Lai,T.-S., Chan,F.-Y., Chung,W.-H., Abagyan,R., Leung,Y.-C. and Wong,K.-Y. (2008) Discovery of a drug-like G-quadruplex binding ligand by high-throughput docking. *Chem. Med. Chem.*, **3**, 881–884.
31. Ma,D.-L., Chan,D.S.-H., Lee,P., Kwan,M.H.-T. and Leung,C.-H. (2011) Molecular modeling of drug–DNA interactions: virtual screening to structure-based design. *Biochimie*, **93**, 1252–1266.
32. McGhee,J.D. and von Hippel,P.H. (1977) Formaldehyde as a probe of DNA structure. 3. Equilibrium denaturation of DNA and synthetic polynucleotides. *Biochemistry*, **16**, 3267–3276.
33. Demidov,V.V. (1980) Interaction with completely fluctuationally opened base-pairs is not the only pathway of formaldehyde-DNA reaction. *Dokl. Akad. Nauk SSSR*, **251**, 1268–1270.
34. Hall,D.H., Grove,L.E., Yueh,C., Ngan,C.H., Kozakov,D. and Vajda,S. (2011) Robust identification of binding hot spots using continuum electrostatics: application to hen egg-white lysozyme. *J. Am. Chem. Soc.*, **133**, 20668–20671.
35. Baker,N.A., Sept,D., Joseph,S., Holst,M.J. and McCammon,J.A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci USA*, **98**, 10037–10041.