An Application of 3D-QSAR to the Analysis of the Sequence Specificity of DNA Alkylation by Uracil Mustard

Arthur M. Doweyko* and William B. Mattes

CIBA-GEIGY Corporation, Environmental Health Center, Farmington, Connecticut 06032

Received May 5, 1992; Revised Manuscript Received July 7, 1992

ABSTRACT: The sequence specificity of DNA alkylation by uracil mustard was examined using a novel three-dimensional QSAR method known as HASL, or the hypothetical active site lattice. The structures of a variety of 4-mer sequences obtained from pBR322 and SV40 were related to their degree of guanine-N7 alkylation by uracil mustard. The resulting correlations were found to point to a significant contribution from bases on the 3′ side of the target guanine nucleotide. The HASL models derived from the analysis of 52 guanine-containing 4-mer sequences were used to highlight those atomic features in the favored TGCC sequence that were found most important in determining specificity. It was found that the NH_2–O systems present in the two GC base pairs on the 3′ side of the target guanine were significantly correlated to the degree of alkylation by uracil mustard. This finding is consistent with a prealkylation binding event occurring between these sites along the major groove and the uracil mustard O2/O4 system.

The sequence recognition of double-stranded DNA has been a subject of much interest and extensive research since such interactions can be used by proteins to regulate gene expression (Pabo & Sauer, 1984) and play a central role in the action of a number of ligands having antibiotic and antitumor properties (Zakrzewska & Lavery, 1989). While a typical binding motif for proteins appears to involve the major groove, the majority of small-molecule, noncovalent ligands have been observed to prefer minor groove binding at AT-rich sequences of B-DNA (Zimmer & Wahnert, 1986). Such minor groove interactions are largely explained by the electrostatic properties of DNA. The uracil mustard alkylation of DNA represents an opportunity to study the behavior of a small molecule in the major groove wherein a significant prealkylation hydrogen bonding interaction may be occurring.

5-[Bis(2-chloroethyl)amino]uracil (uracil mustard, UM) exhibits a sequence specificity in DNA alkylation unique for nitrogen mustards (Mattes et al., 1986a). Nitrogen mustards are known to alkylate, almost exclusively, the guanine-N7 position in double-stranded DNA (Lawley, 1966; Singer, 1975) and to preferentially alkylate guanines in oligoguanine sequences (Mattes et al., 1986a). This observation has been explained by the influence of the nearest neighbor base pairs on the molecular electrostatic potential in the vicinity of guanine-N7 positions in B-DNA (Kohn et al., 1987; Mattes et al., 1986a; Pullman & Pullman, 1981). The sequence specificity of DNA alkylation by uracil mustard exhibits a further detail (5′-YGC-3′) which cannot be explained by electrostatics alone. Computer modeling studies have suggested an interaction between the uracil-O4 and the NH of the 3′-C as a possible explanation for the observed specificity (Kohn et al., 1987).

The present investigation was undertaken in an attempt to further understand the key structural features in the DNA/UM interaction responsible for sequence specificity by applying a novel three-dimensional QSAR technique, the hypothetical active site lattice (HASL) (Doweyko, 1988, 1989, 1990). HASL model building typically entails the superpositioning of a number of small molecules and results in a "hypermolecule" model embodying elements (structural, electronic, etc.) necessary to explain the macromolecular binding of each member of the series. In the case of enzymatic activity, the model would potentially reflect active site structure. Our analysis of DNA sequences involved a similar procedure, but with a twist: the structures of a series of DNA sequences were superposed to create a model of uracil mustard binding. Thus, a series of macromolecules (representing receptors) were combined to develop an understanding of the features common to all regarding the binding of one small molecule. We were able to obtain uracil mustard alkylation intensity data by quantitating the gel electrophoretic analysis of alkylated/hydrolyzed DNA segments. These data were then correlated to structures by applying HASL methodology.

MATERIALS AND METHODS

Chemicals. Mechlorethamine dihydrochloride (HN2) was obtained from the National Cancer Institute Chemical Carcinogen Reference Standard Repository via Midwest Research Institute. Uracil mustard was a gift of the Upjohn Co. Dimethyl sulfate was obtained from Aldrich Chemical Co. Triethanolamine and piperidine were obtained from Fisher Chemical Co. Acrylamide, *N,N,N′,N′*-tetramethylethylenediamine (TEMED), and ammonium persulfate were obtained from Bio-Rad Laboratories. Urea and redistilled phenol were obtained from International Biotechnologies, Inc. [α-^{32}P]dGTP (3200 Ci/mmol) and [α-^{32}P]ATP (7000 Ci/mmol) were obtained from New England Nuclear.

Enzymes. T4 polynucleotide kinase and pBR322 DNA were obtained from Pharmacia P-L Biochemicals. Asp 718, calf alkaline phosphatase, glycogen, and yeast transfer RNA were obtained from Boehringer-Mannheim. SV40 DNA, *Hin*dIII, *Bam*HI, *Sal*I, *Taq*I, and *Escherichia coli* DNA polymerase large fragment were obtained from Bethesda Research Laboratories.

Preparation of End-Labeled DNA. DNA was digested with the indicated restriction enzyme according to the manufacturer's recommendations. If the DNA was to be labeled with T4 polynucleotide kinase, 0.3 unit of alkaline phosphatase was added in the last 30 min of incubation at 37 °C. Enzyme digests were terminated by the addition of EDTA to a final concentration of 20 mM and phenol–chloroform extraction as described by Maniatis et al. (1982). DNA was recovered by ethanol precipitation.

0006-2960/92/0431-9388$03.00/0

Dephosphorylated DNA (1–10 μg) was labeled at its 5′ ends as described by Maxam and Gilbert (1980) except that added radiolabel (50–300 μCi [α-^{32}P]ATP at 7000 Ci/mmol in a final volume of 20–40 μL) was the sole source of ATP. DNA was labeled at its 3′ ends with Klenow fragment of *E. coli* DNA polymerase I and [α-^{32}P]dGTP as described by Maniatis (1982). Labeled DNA was digested with a second restriction enzyme and electrophoresed on a preparative agarose gel. Labeled fragments were located by autoradiography and electroeluted from the gel.

Preparation of Damaged DNA. Labeled DNA was incubated with alkylating agents in a buffer of 1 mM EDTA and 50 mM triethanolamine hydrochloride, pH 7.2, in a total volume of 50 μL. After incubation for 60 min at 22 °C (unless otherwise noted), 50 μL of an ice-cold solution containing 0.6 M sodium acetate, 20 mM EDTA, and 200 μg/mL glycogen was added and the DNA recovered by precipitation with 3 volumes of ethanol. DNA with a low level of depurination sites was created by incubating labeled DNA in 65% formic acid for 5 min at 22 °C and precipitating the DNA in the presence of sodium acetate and glycogen. For all treatments the precipitated DNA was resuspended in 0.3 M sodium acetate–1 mM EDTA and ethanol precipitated a second time, and the resulting pellet was washed with cold ethanol before drying under vacuum. Breaks at sites of guanine-N7 alkylation were created by resuspending the salt-free DNA pellet in freshly diluted 1 M piperidine and incubating at 90 °C for 15 min.

Three DNA segments were alkylated by uracil mustard: set 1, positions 413–551 on the 275 bp *Bam*HI–*Sal*I fragment of pBR322, 5′ labeled at the *Bam*HI site; set 2, positions 4056–4358 (complementary strand) on the 3741 bp *Hin*dIII–*Sal*I fragment of pBR322, 5′ labeled at the *Hin*dIII site; and set 3, positions 47–238 on the 798 bp Asp 718–*Taq*I fragment of SV40, 3′ labeled at the Asp 718 site (Mattes et al., 1986a).

Gel Electrophoresis. The intensities of uracil mustard alkylation at different guanine sites were determined experimentally by a modification of Maxam–Gilbert DNA sequencing methodology (Maxam & Gilbert, 1980; Mattes et al., 1986b). Gels, 17 × 90 × 0.04 cm, 6% polyacrylamide [5.7% acrylamide–0.3% *N*,*N*′-methylenebis(acrylamide)], using 7 M urea and a Tris–boric acid–EDTA buffer system, are essentially those described by Maxam and Gilbert (1980). All samples were heated at 90 °C for 60 s, chilled in an ice bath before loading, and electrophoresed at 4000 V for 3 h. The polyacrylamide gels were dried onto Whatman DE 81 paper and exposed to Kodak XAR-5 film at –70 °C without the use of intensifying screens. A representative sequencing gel is shown in Figure 1A. Lane 2 of the autoradiogram was scanned using a densitometer.

Densitometric Analysis. Autoradiograms were scanned with a Pharmacia UltroScan XL laser densitometer. After conversion to the appropriate format, the data file was analyzed using the GS365W software (Hoeffer Scientific Instruments–Paedia Software). After background subtraction, peaks were integrated using a Gaussian fit algorithm, yielding peak area estimates for each alkylated guanine site (Figure 1B). Shaded areas represent peak areas not included in the fitted Gaussian peak data. The peak area values for the three DNA segments were normalized to the same scale using common 4-mer sequences as points of reference. In addition, end effects were accounted for by the method of Hazeltine (1980).

Modeling. Molecular modeling was conducted using MacroModel (V3.0, VAX version, Clark Still, Columbia University, 1990) and PC-CHEMMOD (V5.0, PC version, Frazer-Williams Ltd., Cheshire, U.K., 1990). The 52 4-mers (tetranucleotide helices) examined in this study were constructed by first using the B-DNA building option in MacroModel, and then removing coordinates corresponding to the sugar–phosphate backbone. Preliminary modeling investigations indicated that the backbone represented a negligible contribution to the sequence-recognition phenomenon. The XGYZ structures were then each superposed on atoms of the heterocyclic bases through the common guanine base using the least squares superpose option in PC-CHEMMOD. The resulting set of structures were then subjected to

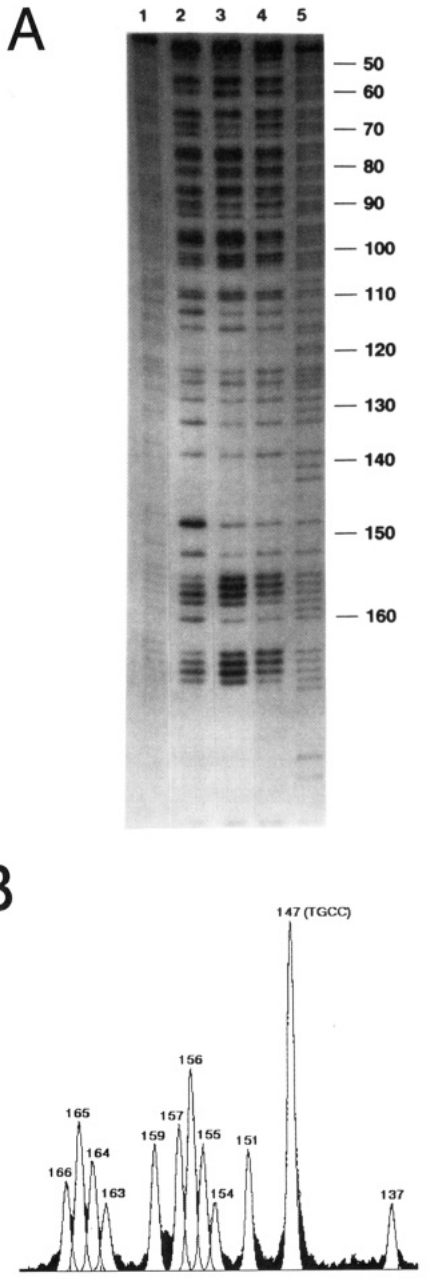

FIGURE 1: Determination of DNA sequence specificity for guanine-N7 alkylation of the enhancer region of SV40 DNA by uracil mustard. (A) Autoradiogram of a denaturing 6% polyacrylamide gel showing sites of guanine alkylation. DNA was alkylated with 10 μM uracil mustard (lane 2), 20 μM mechlorethamine (lane 3), or 100 μM dimethyl sulfate (lane 4: Maxam–Gilbert guanine reaction). Lane 1 is control DNA, and lane 5 is DNA depurinated with mild formic acid treatment (Maxam–Gilbert guanine and adenine reaction). The substrate was the 798 bp *Bam*I–*Taq*I fragment of SV40, 3′ labeled at the *Bam*I site. (B) Densitometric analysis of alkylation intensities illustrating Gaussian peak fitting.

HASL analysis (HASL V3.22, PC version; PC/VAX versions written by A.M.D.).

HASL Analysis. Although details of HASL methodology have been discussed elsewhere (Doweyko, 1988, 1989, 1991), it may be useful to briefly summarize the technique, particularly as it applies to the present investigation. The basic assumption underlying this approach is that the space occupied by a series of molecules can be related to macromolecular ligand binding. This molecular space is mathematically converted to a lattice of points, each retaining information regarding the atomic character present at that point (Figure 2A). Each point lies within the van der Waals radius of an atom. Once merged with lattice points obtained from other sequence structures, these points make up the HASL by accounting for the presence of an atom type at a specific location in space. In addition, each point also serves as the basis for binding estimates. Using the uracil mustard (UM) binding data for 52 4-mer sequences, it was possible to construct a series of HASL models incorporating the structural features of each tetranucleotide and associating it with the degree of alkylation by UM. Once a series of molecules is converted to such a lattice, the activity (e.g., ligand binding, pK_i) of each molecule is distributed among its points in a manner such that all molecules are accurately represented. The distribution of partial binding values is a critical operation which relies on the lattice resolution and the number of molecules. The current implementation of HASL makes use of an iterative solution to this distribution problem which is akin to the solution of a set of simultaneous equations, wherein the activity of each molecule represents an equation, and each lattice point, a coefficient. In an effort to limit the analysis to potentially relevant atom types, HASL models were constructed using only atoms defined as electron-rich ($H = 1$) or electron-poor ($H = -1$).

RESULTS AND DISCUSSION

Quantitated densitometry of the gel scans described earlier yielded the data sets shown in Table I for the three DNA segments (pBR322 [413–551], pBR322 [4056–4358], and SV40 [47–238]). Log peak area values (LogArea) were used to seek out the putative linear relationship with the free energy of binding of UM to DNA prior to alkylation. The averaged LogArea values were used in the HASL analysis.

Resolution. Since HASL model building relies on sampling 3D space, it was important to assess the role of point spacing, or resolution, on the effectiveness of such model building. The effect of resolution on the construction of a predictive HASL model is shown in Figure 2B. In an attempt to identify those resolution values most likely to provide meaningful models, 26 4-mer sequences were used as a learning set (odd-numbered sequences, i.e., 1, 3, 5...). Spanning resolutions of 2.00–3.00 Å (in 0.05-Å increments), each model was used to predict binding (LogArea) for a test set comprised of the remaining 26 4-mers (even-numbered sequences, i.e., 2, 4, 6...). Each 26 4-mer model was iteratively solved to provide an internal error (actual vs predicted LogArea) of less than 0.0001 LogArea unit. The predictivity of each model was measured as the correlation coefficient (r^2) found between the test set actual and predicted binding values. The 26 4-mer learning set predictions are shown in Figure 2C and indicate that predictivity was best ($r^2 = 0.547$) at 2.45 Å. Such a dependency upon resolution can be visualized as the effect of having key atoms drift in and out of detection by the regularly-spaced set of lattice points. A full 52 4-mer model was constructed at 2.45 Å based on the optimal predictivity

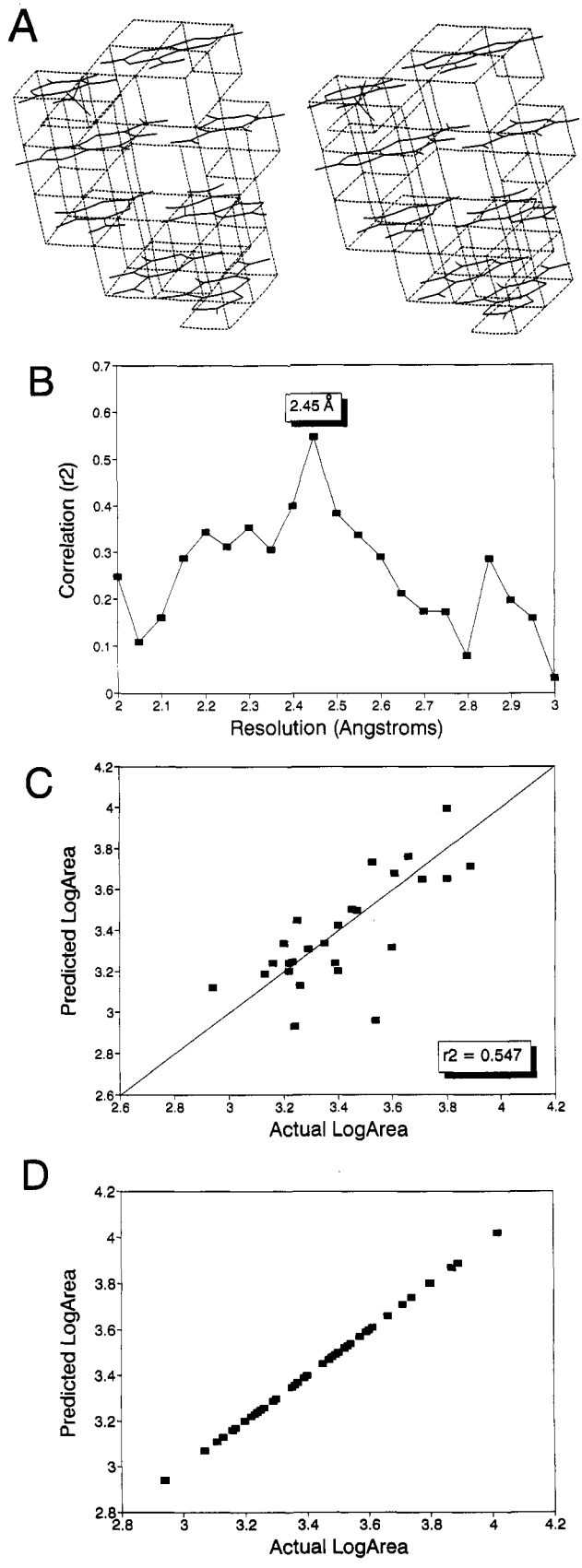

FIGURE 2: (A) The base pair sequence TGCC is used to illustrate the construction of a lattice of points at a resolution of 2.45 Å. (B) The HASL models constructed over the range of 2.00–3.00 Å using 26 4-mer sequences were evaluated by examining their predictivity, i.e., their ability to accurately predict the degree of alkylation observed for the other 26 4-mer sequences. (C) The 26-member learning set model predicted the 26-member test set with a correlation coefficient (r^2) of 0.547. (D) The complete 52 4-mer DNA/UM HASL model at 2.45 Å was solved to an r^2 of 1.000.

Table I: Normalized Sequence Data[a]

no.	sequence	set 1	set 2	set 3	av
1	AGAA		3.41	3.63	3.52
2	AGAC		3.29		3.29
3	AGAG		3.49		3.49
4	AGAT	3.06		3.20	3.13
5	AGCC			3.50	3.50
6	AGCG	3.39			3.39
7	AGGC	3.52			3.52
8	AGGG		3.80		3.80
9	AGGT	3.40	3.29		3.35
10	AGTA		3.71		3.71
11	AGTT			3.59	3.59
12	CGAA		3.54		3.54
13	CGAC	3.11			3.11
14	CGAT	2.94			2.94
15	CGCC	3.67	3.82		3.74
16	CGCG		3.40		3.40
17	CGCT	3.50	3.40		3.45
18	CGGA		3.21	3.59	3.40
19	CGGC	3.47			3.47
20	CGGG	3.27	3.13		3.20
21	CGGT	3.30			3.30
22	CGTC		3.23		3.23
23	CGTG	3.16	3.35		3.26
24	GGAA	3.39	3.54		3.47
25	GGAC			3.37	3.37
26	GGAG			3.60	3.60
27	GGCA	3.34	3.43		3.39
28	GGCC	3.67	3.93		3.80
29	GGCG	3.34	3.37		3.36
30	GGCT	3.53			3.53
31	GGGA	3.75	3.55	3.68	3.66
32	GGGC	3.75	4.01	3.90	3.89
33	GGGG	3.58	3.54	3.59	3.57
34	GGGT	3.66			3.66
35	GGTA	3.87			3.87
36	GGTG	3.61	3.60		3.61
37	GGTT	3.47	3.52	3.46	3.48
38	TGAA		3.26		3.26
39	TGAC			3.17	3.17
40	TGAG	3.35	3.25	3.13	3.24
41	TGAT		3.07		3.07
42	TGCA			3.22	3.22
43	TGCC		4.08	3.96	4.02
44	TGCG	3.38	3.32		3.35
45	TGCT	3.42	3.67	3.50	3.53
46	TGGC	3.43	3.47		3.45
47	TGGG	3.24		3.33	3.29
48	TGGT	3.19	3.17	3.38	3.25
49	TGTT	3.25	3.24		3.25
50	TGTA		3.16		3.16
51	TGTC		3.52		3.52
52	TGTG		3.22		3.22

[a] Area values expressed as LogArea. Set 1: pBR322 [413–551]; set 2: pBR322 [4056–4358]; set 3: SV40 [47–238].

observed at that resolution. This model was found to be internally consistent, with a prediction error of less than 0.0001 LogArea unit (Figure 2D).

Prediction. The real benefit of modeling is prediction. As was discussed above and illustrated in Figure 2D, it is possible to superpose a molecule onto a HASL and predict binding (LogArea). More interestingly, it is also possible to examine the superposed molecule at the atomic level and detail each atom's contribution to the overall predicted binding. This is done by summing the partial binding values at each lattice point within the van der Waals radius of the atom in question. This procedure allows one to identify those molecular features which the model deems most critical to binding. The sequence specificity observed for DNA alkylation by uracil mustard was examined in this way by superposing the preferred TGCC sequence structure onto the 2.45-Å model. The atomic contributions to binding were then determined and are

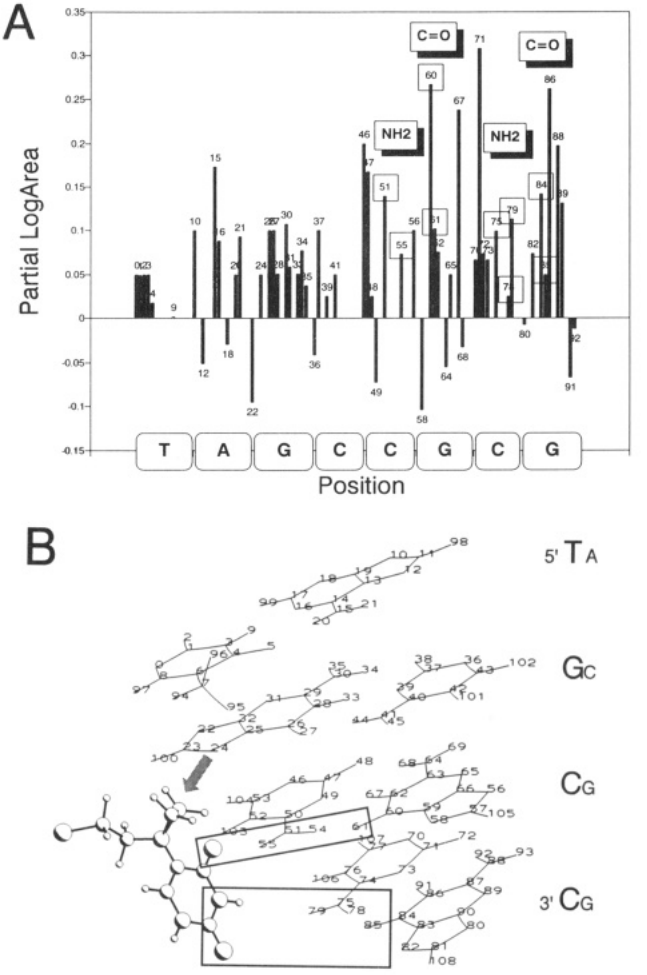

FIGURE 3: (A) Significant contributions to alkylation/binding are identified with atoms along the major groove: cytosine-NH_2 and guanine-C=O in TGCC, and to a lesser extent, the same NH_2/C=O set in TGCC. (B) Illustration of a possible prealkylation event: hydrogen-bonded interactions occur between O4-(UM) and O6(TGCC)NH_2 and, possibly, between O2(UM) and O6(TGCC)NH_2 prior to the nucleophile, N7(TGCC, atom 24), attacking an available methylene in the aziridinium ring.

illustrated in Figure 3A. Although significant contributions to the overall binding of UM aziridinium ion to TGCC were observed for a number of atoms, those atoms that are accessible to the aziridinium ion by way of major groove binding are the cytosine-NH_2 (TGCC: atoms 51, 54, and 55), guanine-C=O (TGCC: atoms 60 and 61), and to a modest extent, the same cytosine-NH_2/guanine-C=O set in the next nucleotide pair (TGCC: atoms 75, 78, and 79; atoms 84 and 85). This CG/CG pair (TGCC) provides the basis for a physically meaningful mechanism to explain the observed UM alkylation preference for the TGCC sequence (illustrated in Figure 3B). The HASL model would suggest that the attack of the target guanine-N7 (atom 24) on the aziridinium ring occurs after the uracil ring interacts with the two CG pairs on the 3′ side. Mechanistically, it is probable that the uracil-O4 hydrogen bonds to the TGCC NH_2–O system, accounting for the atomic binding contributions predicted by HASL and consistent with previous observations (Kohn et al., 1987). In addition, the selectivity observed for TGCC is consistent with a further prealkylation binding event: uracil-O2 (as the lactam or lactim) hydrogen bonding with the NH_2–O system in TGC C. Molecular modeling analysis suggests that both hydrogen bonding events are possible, with H–O or H–N distances averaging 1.5–2.5 Å. This hypothesis is consistent with the lack of TGCC specificity observed for 6-methyl-UM (Kohn

et al., 1987; Mattes, unpublished results). In the case of 6-methyl-UM, there exists a significant steric interaction limiting rotation of the aziridinium ring to a conformation nearly perpendicular to the uracil ring. In contrast, the absence of a 6-methyl group in UM permits the aziridinium ring a greater degree of rotational freedom and an option to adopt the nearly coplanar conformation required to interact with the DNA major groove GC recognition sites illustrated in Figure 3B.

A series of ligands and a single receptor are the necessary elements in any structure–activity study. In the case of the DNA/UM ligand–receptor recognition phenomenon, a set of large molecules (tetranucleotides) served as the ligand series and UM as the receptor. The analysis of this relationship invoked the application of a unique 3D-QSAR strategy (HASL) which provided information about the space occupied by the tetranucleotide and its susceptibility to alkylation by uracil mustard. Presuming a prealkylation equilibrium between UM and DNA near the alkylation site (guanine-N7), it was possible to correlate the structural parameters about that site to the specificity observed. The results of this analysis point to a significant interaction involving potential hydrogen bonding sites on UM and the NH_2–O moieties on the 3′ side of the target guanine-N7 along the major groove and, thus, demonstrate the utility of 3D-QSAR in the study of DNA major groove recognition sites.

REFERENCES

Doweyko, A. M. (1988) *J. Med. Chem. 31*, 1396.

Doweyko, A. M. (1989) in *Probing Bioactive Mechanisms* (Magee, P. S., Henry, D. R., & Block, J. H., Eds.) pp 82–104, ACS Symposium Series 413, Washington, DC.

Doweyko, A. M. (1991) *J. Math. Chem. 7*, 273.

Hazeltine, W. A., Gordon, L. K., Lindan, C. P., Grafstrom, R. H., Shaper, N. L., & Grossman, L. (1980) *Nature 285*, 634.

Kohn, K. W., Hartley, J. A., & Mattes, W. B. (1987) *Nucleic Acids Res. 15*, 10531.

Lawley, P. D. (1966) *Prog. Nucleic Acid Res. Mol. Biol. 5*, 89.

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular Cloning. A Laboratory Manual* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Mattes, W. B., Hartley, J. A., & Kohn, K. W. (1986a) *Nucleic Acids Res. 14*, 2971.

Mattes, W. B., Hartley, J. A., & Kohn, K. W. (1986b) *Biochim. Biophys. Acta 868*, 71.

Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499.

Pabo, C. O., & Sauer, R. T. (1984) *Annu. Rev. Biochem. 53*, 293.

Pullman, A., & Pullman, B. (1981) *Q. Rev. Biophys. 14*, 289.

Singer, B. (1975) *Prog. Nucleic Acid Res. Mol. Biol. 15*, 219.

Zakrzewska, K., & Lavery, R. (1989) in *Computer-Aided Molecular Design* (Richards, W. G., Ed.) pp 129–145, VCH Publishers, Inc., New York.

Zimmer, C., & Wahnert, U. (1986) *Prog. Biophys. Mol. Biol. 47*, 31.