Structural Characterization of Formaldehyde-Induced Cross-Links Between Amino Acids and Deoxynucleosides and Their Oligomers

Kun Lu,[†] Wenjie Ye,[†] Li Zhou,[‡] Leonard B. Collins,[†] Xian Chen,[‡] Avram Gold,*,[†]
Louise M Ball,*,[†] and James A. Swenberg*,[†]

Department of Environmental Sciences and Engineering and Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

Received October 6, 2009; E-mail: golda@email.unc.edu; lmball@email.unc.edu; jswenber@email.unc.edu

Abstract: Exposure to formaldehyde results in the formation of DNA−protein cross-links (DPCs) as a primary genotoxic effect. Although DPCs are biologically important and eight amino acids have been reported to form stable adducts with formaldehyde, the structures of these cross-links have not yet been elucidated. We have characterized formaldehyde-induced cross-links of Lys, Cys, His, and Trp with dG, dA, and dC. dT formed no cross-links, nor did Arg, Gln, Tyr, or Asn. Reaction of formaldehyde with Lys and dG gave the highest yield of cross-linked products, followed by reaction with Cys and dG. Yields from the other coupling reactions were lower by a factor of 10 or more. Detailed structural examination by NMR and mass spectrometry established that the cross-links between amino acids and single nucleosides involve a formaldehyde-derived methylene bridge. Lys yielded two additional products with dG in which the linking structure is a $1,N^2$-fused triazino ring. The Lys cross-linked products were unstable at ambient temperature. Reactions between the reactive N^α-Boc-protected amino acids and the trinucleotides $d(T_1B_2T_3)$ where B_2 is the target base G, A, or C and reactions between dG, dA and dC and 8-mer peptides containing a single reactive target residue at position 5 yielded cross-linked products with structures inferred from high resolution mass spectrometry and fragmentation patterns that are consistent with those between N^α-Boc-protected amino acids and single nucleotides rigorously determined by NMR studies. These structures will provide a basis for investigation of the characteristics and properties of DPCs formed in vivo and will be helpful in identifying biomarkers for the evaluation of formaldehyde exposure both at the site of contact and at distant sites.

Introduction

Formaldehyde is a ubiquitous environmental contaminant, with human exposures occurring during use in industrial processes such as the manufacture of resins, particle board, plywood, leather goods, paper, and pharmaceuticals and through emission as a vapor over the lifetimes of these products.[1] Formaldehyde also occurs endogenously as a normal metabolic intermediate in human cells as well as under certain pathological conditions such as oxidative stress. Although formaldehyde is a known carcinogen and mutagen and has thus evoked serious health concerns,[2−6] its mode of action is not well understood. Among the interactions of formaldehyde considered to have

biological relevance are the induction of DNA adducts,[7−14] protein modifications,[15,16] interstrand DNA cross-links,[17−19] and DNA−protein cross-links (DPCs),[5,6,20,21] with the formation of DPCs being considered as the primary genotoxic effect follow-

[†] Department of Environmental Sciences and Engineering.
[‡] Department of Biochemistry and Biophysics.

(1) *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*; International Agency for Research on Cancer: Lyon, France, 1995; Vol. 62, pp 217−245.
(2) Swenberg, J. A.; Kerns, W. D.; Mitchell, R. I.; Gralla, E. J.; Pavkov, K. L. *Cancer Res.* **1980**, *40*, 3398−3402.
(3) Kerns, W. D.; Pavkov, K. L.; Donofrio, D. J.; Gralla, E. J.; Swenberg, J. A. *Cancer Res.* **1983**, *43*, 4382−4392.
(4) Monticello, T. M.; Swenberg, J. A.; Gross, E. A.; Leininger, J. R.; Kimbell, J. S.; Seilkop, S.; Starr, T. B.; Gibson, J. E.; Morgan, K. T. *Cancer Res.* **1996**, *56*, 1012−1022.
(5) Speit, G.; Schutz, P; Merk, O. *Mutagenesis* **2000**, *15*, 85−90.
(6) Merk, O.; Speit, G. *Environ. Mol. Mutagen* **1998**, *32*, 260−268.

(7) Cheng, G.; Wang, M.; Upadhyaya, P.; Villalta, P. W.; Hecht, S. S. *Chem. Res. Toxicol.* **2008**, *21*, 746−751.
(8) Lu, K.; Ye, W.; Gold, A.; Ball, L. M.; Swenberg, J. A. *J. Am. Chem. Soc.* **2009**, *131*, 3414−3415.
(9) Siomin, Y. A.; Simonov, V. V.; Poverenny, A. M. *Biochim. Biophys. Acta* **1973**, *331*, 27−32.
(10) McGhee, J. D.; von Hippel, P. H. *Biochemistry* **1975**, *14*, 1281−1296.
(11) McGhee, J. D.; von Hippel, P. H. *Biochemistry* **1975**, *14*, 1297−1303.
(12) McGhee, J. D.; von Hippel, P. H. *Biochemistry* **1977**, *16*, 3267−3276.
(13) McGhee, J. D.; von Hippel, P. H. *Biochemistry* **1977**, *16*, 3276−3293.
(14) Chang, Y. T.; Loew, G. H. *J. Am. Chem. Soc.* **1994**, *116*, 3548−3555.
(15) Metz, B.; Kersten, G. F. A.; Hoogerhout, P.; Brugghe, H. F.; Timmermans, H. A. M.; de Jong, A.; Meiring, H.; Hove, J. T.; Hennink, W. E.; Crommelin, D. J.; Jiskoot, W. *J. Biol. Chem.* **2004**, *279*, 6235−6243.
(16) Lu, K.; Boysen, G.; Gao, L.; Collins, L. B.; Swenberg, J. A. *Chem. Res. Toxicol.* **2008**, *21*, 1586−1593.
(17) Chaw, Y. F. M.; Crane, L. E.; Lange, P.; Shapiro, R. *Biochemistry* **1980**, *19*, 5525−5531.
(18) Huang, H. F.; Solomon, M. S.; Hopkins, P. B. *J. Am. Chem. Soc.* **1992**, *114*, 9240−9241.
(19) Huang, H. F.; Hopkins, P. B. *J. Am. Chem. Soc.* **1993**, *115*, 9402−9408.
(20) Quievryn, G.; Zhitkovich, A. *Carcinogenesis* **2000**, *21*, 1573−1580.
(21) Solomon, M. J.; Varshavsky, A *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 6470−6474.

Scheme 1. Formation of Formaldehyde-Induced DPCs Originating from the Initial Attack of Formaldehyde on Protein Residues (A) and from the Initial Attack of Formaldehyde on DNA (B)
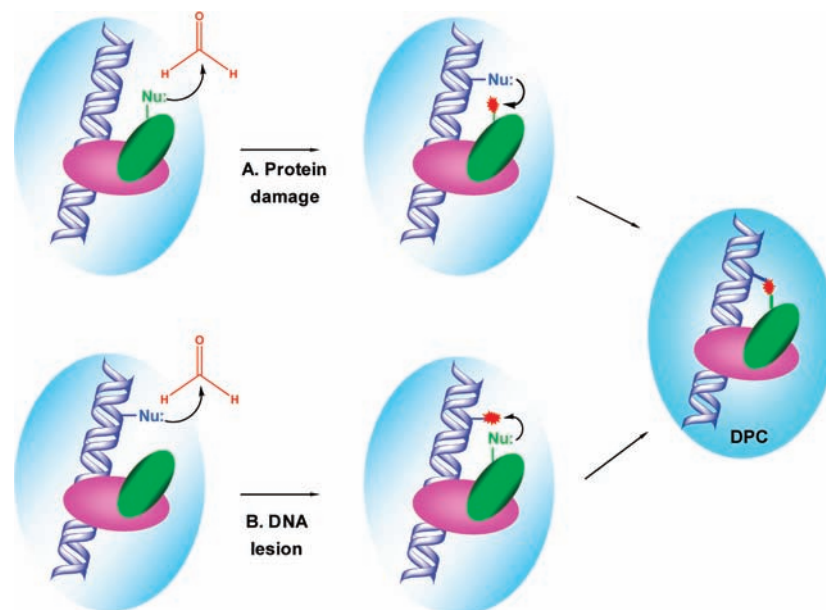
ing exposure to formaldehyde.[5,6] The two routes shown in Scheme 1 may be responsible for the formation of DPCs. In pathway A, formaldehyde addition at a nucleophilic site on a protein is followed by cross-linking between the resulting protein methylol adduct and a nucleophilic site on DNA, whereas in pathway B, addition at a nucleophilic site on a DNA base is followed by cross-linking between the DNA methylol adduct and a protein residue.[22] The formation of DPCs is favored by intimate interactions between DNA and proteins[22] and the lysine-rich DNA-binding histones have been reported to cross-link with DNA[9,20,21] in the presence of formaldehyde.

Key to understanding the toxicity of formaldehyde is the detection and quantitation of the DPCs as biomarkers for formaldehyde exposure, both at sites of contact and at sites removed from contact. Of techniques available for this purpose, those based on liquid chromatography coupled with mass spectrometry have shown great promise in recent years.[23] Selected ion monitoring (SIM), selected reaction monitoring (SRM), or multiple reaction monitoring (MRM) provide ideal sensitivity and specificity for the quantitation of DNA adducts. To take advantage of these techniques, the molecular structures of the DPCs must be characterized and standards generated. In a previous communication, we described a formaldehyde-derived cross-link formed between the Cys sulfhydryl group of glutathione and dG and suggested that this reaction may be relevant to systemic effects of formaldehyde transported as an *S*-hydroxymethylene conjugate of glutathione.[8] In the present work, we have established that cross-linking reactions occur between the amino acids Lys, Cys, His, and Trp and the nucleosides dG, dA, or dC. We have characterized the nature of the cross-links formed in the presence of formaldehyde between the reactive amino acids and trinucleotides $d(T_1B_2T_3)$ where B_2 is the target base G, A, or C. We have also examined the cross-links formed in the presence of formaldehyde between dG, dA, and dC and *N*-terminal protected 8-mer peptides

containing a single target residue at position 5. Finally, we have determined by NMR and mass spectrometric studies the structures of the cross-links formed between potentially cross-linking amino acids and dG, dA, and dC in the presence of formaldehyde as shown in Chart 1. Such structures are anticipated to reflect cross-links in vivo in regions of single-stranded DNA and, implicitly, cross-links formed in double-stranded DNA.

Experimental Section

Chemicals. Potassium phosphate, ammonium bicarbonate, trifluoroacetic acid, formic acid, acetonitrile, methanol, N^α-(*tert*-butoxycarbonyl)-L-lysine, *tert*-butoxycarbonyl-L-cysteine, N^α-(*tert*-butoxycarbonyl)-L-histidine, N^α-(*tert*-butoxycarbonyl)-L-tryptophan, N^α-(*tert*-butoxycarbonyl)-L-arginine, N^α-(*tert*-butoxycarbonyl)-L-asparagine, N^α-(*tert*-butoxycarbonyl)-L-glutamine, N^α-(*tert*-butoxy-carbonyl)-L-tyrosine, deoxyadenosine, deoxythymidine, deoxycytidine, and deoxyguanosine were purchased from Sigma (St. Louis, MO). The *N*-terminal acetylated 8-mer peptides Ac-VEGGRGAA, Ac-VEGGQGAA, Ac-GEGGCGAA, Ac-GEGGYGAA, Ac-VEG-GKGAA, Ac-VEGGNGAA, Ac-GEGGHGAA, and Ac-GEGGW-GAA were synthesized by GenScript Corporation (Piscataway, NJ). Formaldehyde 20% in water was purchased from Tousimis (Rockville, MD). $^{13}CH_2O$ was purchased from Cambridge Isotope Laboratories (Andover, MA). All chemicals were used as received unless otherwise stated.

Instrumental Methods. NMR Analysis. NMR Spectra were recorded on a Varian INOVA NMR spectrometer (Varian, Inc., Palo Alto, CA) at 500 MHz for ^1H NMR and 125 MHz for ^{13}C spectra with the Varian cold probe.

Liquid Chromatography–Mass Spectrometry (LC–MS). LC–MS analyses were performed on a LCQ-Deca ion trap mass spectrometer (Thermo Electron, Waltham, MA), a TSQ Quantum Ultra triple quadrupole mass spectrometer (Thermo Electron, Waltham, MA), or a Q-TOF LC/MS (Model 6240, Agilent Technologies, Santa Clara, CA). The mass spectrometers were equipped with electrospray ionization (ESI) sources. Analytes were separated by reverse phase HPLC using a 250 mm × 2.5 mm C18 analytical column from Grace Vydac (Hesperia, CA) eluted at 200 μL/min with a linear gradient from 2% to 60% solvent A in B over 15 min. Solvent A consisted of 0.1% formic acid in water, and solvent B was methanol. The ion trap mass spectrometer was

(22) Brodolin, K. Protein-DNA crosslinking with formaldehyde in vitro In *DNA-Protein Interaction*; Andrew, T., Malcolm, B., Eds.; Oxford University Press: Oxford, 2000; pp 141−150.
(23) Koc, H.; Swenberg, J. A. *J. Chromatogr. B* **2002**, *778*, 323–343.

Chart 1. Structures of Cross-Linked Adducts between Amino Acids and Nucleosides Identified in This Study[a]
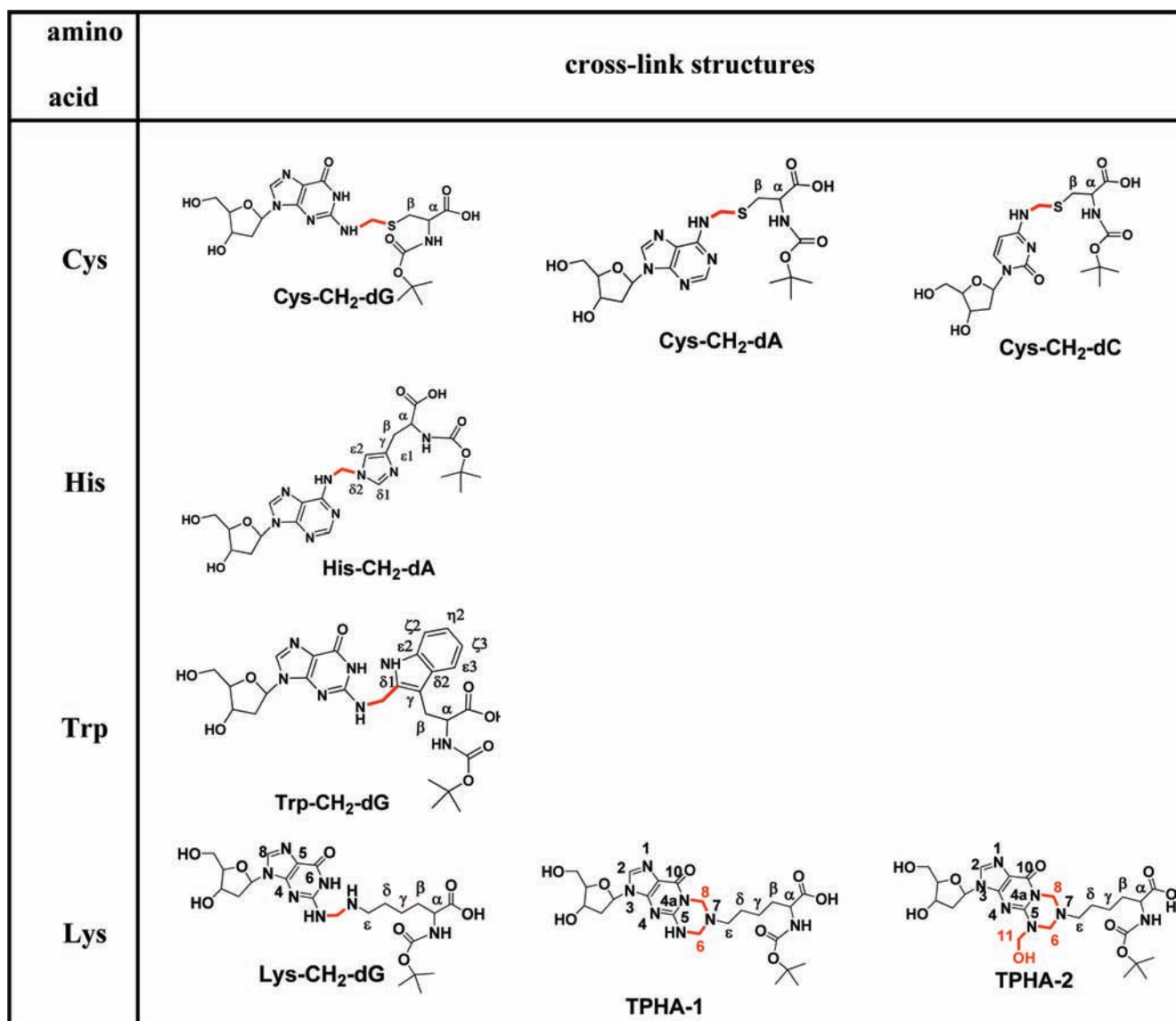
[a] Formaldehyde-derived linkages are shown in red.

operated in full scan as well as dependent scan modes. For the fragmentation of precursor ions, the normalized collision energy of the ion trap mass spectrometer varied from 30% to 35% depending on the adduct. The activation time was set at 30 ms. The triple quadrupole mass spectrometer was operated in full scan, parent mode scan, MS/MS scan, SIM, or SRM mode. The collision energy was set at 17 V for most fragmentation experiments. Q-TOF high resolution mass spectra were obtained on an Agilent 6250 Accurate Mass Q-TOF LC/MS (Agilent Technologies, Santa Clara, CA) equipped with a dual spray ESI source. For liquid chromatography, a Hypersil Gold column (Thermo Scientific, Waltham, MA) (150 mm × 2.1 mm, 3 μm particle size) was used with a linear gradient from 2% acetonitrile in 0.1% formic acid to 80% acetonitrile over 15 min, eluted at 200 μL/min. The ESI source was set as follows: gas temperature, 350 °C; drying gas, 10 L/min; Vcap, 4000 V; nebulizer, 35 psig; fragmentor, 100 V; skimmer, 65 V. Fourier-transform ion cyclotron resonance mass spectra (FTICR-MS) (10 scans) were acquired on a hybrid Qe-Fourier transform ion cyclotron resonance mass spectrometer equipped with a 12 T actively shielded magnet (Apex Qe-FTICR-MS, 12.0 T AS, Bruker Daltonics, Billerica, MA) and an Apollo II microelectrospray (μESI) source. The voltages on the μESI sprayer, interface plate,

heated capillary exit, deflector, ion funnel, and skimmer were set at 4.2 kV, 3.9 kV, 300 V, 250 V, 175 V, and 30 V, respectively. The μESI source temperature was maintained at 180 °C. Desolvation was carried out by using a nebulization gas flow (2 bar) and a countercurrent drying gas flow. Before transfer, ion packets were accumulated inside the collision cell for a duration of 0.02 s. Using a syringe pump (Cole Parmer, Vernon Hills, IL), sample solutions were infused with a 250 μL syringe (Hamilton, Reno, NV) at 90 μL/h.

Formation of Formaldehyde-Induced Cross-Links. Typical reaction mixtures (final volume 50 μL) consisted of amino acid (1 mM) and deoxynucleoside (1 mM) dissolved in potassium phosphate buffer (10 mM, pH 7.2) to which formaldehyde was added to a final concentration of 50 mM. After 48 h, the reaction mixtures were either separated by reverse phase HPLC or analyzed by LC−MS. Lys-dG coupling reactions were carried out with 5, 50, and 100 mM formaldehyde. Reactions using 50 mM formaldehyde were run for 48, 60, 72, and 84 h. The coupling reaction was repeated with 50 mM ^{13}C-formaldehyde for 48 h.

Formaldehyde-induced coupling between peptides and deoxynucleosides was accomplished in the same manner in 50 μL reaction volumes with final concentrations of 50 mM formaldehyde, 1.5 mM

peptide, and 5 mM deoxynucleoside. For coupling between amino acids and trinucleotides in 50 μL reaction volumes, final concentrations were 50 mM formaldehyde, 5 mM amino acid, and 1 mM trinucleotide.

Large scale reactions between amino acids and deoxynucleosides for structural characterization by NMR were run with 20 mg of deoxynucleoside and 40 mg of N^α-Boc-amino acid in 5 mL of 10 mM potassium phosphate buffer (pH 7.2) and 100 mM formaldehyde for 12 h to 1 week, monitored by HPLC until the chromatographic trace remained constant. The reaction mixtures were separated by semipreparative HPLC, and collected products were characterized by mass spectrometry and NMR. Exact masses are given in Table 3. MS/MS data are presented in full as Supporting Information. H and ^{13}C shifts are shown below. For the Lys-dG reactions, HPLC fractions were collected on dry ice, lyophilized, and then stored at −80 °C until analysis.

Kinetics of Cross-Link Formation. Solutions were made up to final concentrations of 4 mM amino acid, 4 mM deoxynucleoside, and 50 mM formaldehyde in 50 mM potassium phosphate buffer (pH 7.2). The reaction mixtures were analyzed at 2, 4, 8, 16, 24, 48, and 72 h by HPLC with UV detection at 254 nm. For the Lys-dG coupling reaction, additional samples were analyzed at 10, 20, 30, and 60 min.

Stability of Cross-Linked Lys-dG Products. To measure the kinetics of cross-link degradation, a reaction mixture containing Lys and dG as described above was treated with 50 mM formaldehyde for 48 h and then separated by HPLC. Fractions eluting at 17.2 and 26.5 min were collected on dry ice, and 50 μL aliquots of each fraction were added to 950 μL of 50 mM phosphate buffer maintained at 37 °C for 1, 5, 10, 20, or 30 min and then analyzed by HPLC with ESI Q-TOF mass spectrometry.

2-Amino-6-(10-oxo-triazino[1,2-*a*]purin-7-yl)hexanoic Acid (TPHA-1). ^{1}H NMR (500 MHz, DMSO-d_6) δ 7.91 (s,1H, H2), 7.87 (s, 1H, N5*H*, J = 2.1 Hz), 6.74 (d, 1H, Boc-$N^\alpha H$-, J = 4.7 Hz), 6.09 (dd, 1H, H1′, J = 7.8, 5.9 Hz), 4.89 (s, 2H, C8*H*$_2$), 4.33 (m, 1H, H3′), 4.24 (bs, 2H, C6*H*$_2$), 3.83 (m,1H, H4′), 3.75 (dt, 1, C$^\alpha H$, J = 8.1, 8.1, 4.7 Hz), 3.54 (dd, 2H, H5′, J = 11.7, 4.5 Hz), 3.47 (dd, 2H, H5″, J = 11.7, 4.5 Hz), ~2.47 (C$^\varepsilon H_2$ (overlaps DMSO-d_6)), ~2.50 (H2′ (overlaps DMSO-d_6)), 2.18 (ddd, 1H, H2″, J = 13.1, 5.9, 3.0 Hz), 1.34−1.48 (m, 2H, C$^\delta H_2$), 1.24−1.37 (m, 2H, C$^\gamma H_2$), 1.44−1.58 (m, 2H, C$^\beta H_2$), 1.35 (s, 9H, C*H*$_3$-Boc). ^{13}C NMR (125 MHz, DMSO-d_6) δ 173.9 *C*OOH, 155.8 C10, 155.2 *C*OO-Boc, 150.2 C4a, 149.0 C3a, 134.9 C2, 115.5 C10a, 87.3 C4′, 82.1 C1′, 77.4 *C*-Boc, 70.5 C3′, 61.4 C5′, 60.5 C8, 59.5 C6, 53.8 C$^\alpha$, 49.2 C$^\varepsilon$, 39.3 C2′, 30.8 C$^\beta$, 27.9 C*H*$_3$-Boc, 26.6 C$^\delta$, 22.8 C$^\gamma$.

2-Amino-6-(5-hydroxymethyl-10-oxo-triazino[1,2-*a*]purin-7-yl)hexanoic Acid (TPHA-2). ^{1}H NMR (500 MHz, DMSO-d_6) δ 7.96 (s,1H, H2), 6.93−6.99 (m,1H, N^αH-Lys), 6.19 (dd, 1H, H1′, J = 7.8, 5.9 Hz), 4.99 (d, 2H, N5C*H*$_2$OH, J = 1.3 Hz), 4.95 (s, 2H, C8*H*$_2$), 4.45 (s, 2H, C6*H*$_2$), 4.35 (m, 1H, H3′), 3.78−3.83 (m, 1H, H4′), 3.77−3.84 (m, 1H, C$^\alpha H$), 3.55 (dd, 1H, H5′, J = 11.6, 4.8 Hz), 3.49 (dd, 1H, H5, J = 11.6, 4.7 Hz), 2.58 (ddd, 1H, H2′, J = 13.2, 7.7, 5.9 Hz), 2.49 (m, 2H, C$^\varepsilon H_2$), 2.21 (ddd,1H, H2″, J = 13.2, 7.8, 3.1 Hz), 1.50−1.69 (m, 2H, C$^\beta H_2$), 1.36−1.50 (m, 2H, C$^\delta H_2$), 1.36 (s, 9H, C*H*$_3$-t-Boc), 1.28−1.34 (m, 2H, C$^\gamma H_2$). ^{13}C NMR (125 MHz, DMSO-d_6) δ 173.9 *C*OOH-Lys, 156.0 C10, 155.4 *C*OO-t-Boc, 149.4 C4a, 148.2 C3a, 136.0 C2, 115.9 C1a, 87.4 C4′, 82.3 C1′, 77.6 *C*-t-Boc, 70.7 C3′, 68.5 N5*C*H$_2$OH, 63.9 C6, 61.5 C5′, 61.2 C8, 53.2 C$^\alpha$, 49.0 C$^\varepsilon$, 39.2 C2′, 30.4 C$^\beta$, 27.9 C*H*$_3$-t-Boc, 26.6 C$^\delta$, 22.9 C$^\gamma$.

Cys−CH$_2$−dG. ^{1}H NMR (500 MHz, DMSO-d_6) δ 12.73 (s, 1H, COO*H*-Cys), 10.82 (s, 1H, N1*H*-dG), 7.99 (s, 1, H8), 7.11 (bs, 1H, $N^2 H$-dG), 6.99−7.08 (m, 1H, N^α-Cys), 6.16 (ψt, 1H, H1′, J = 6.8 Hz), 4.48−4.57 (m, 2H, C*H*$_2$-linker, J = 10 Hz), 4.35 (td, 1, H3′, J = 3.2, 3.2, 6.1 Hz), 4.06−4.12 (m, 1H, C$^\alpha H$-Cys), 3.78−3.82 (m, 1H, H4), 3.55 (dd, 1H, H5′, J = 11.6, 4.8), 3.48 (dd, 1H, H5″, J = 11.62, 4.8 Hz), 3.03 (dd, 1H, C$^\beta H_{2a}$, J = 13.5, 4.47 Hz), 2.84 (dd, 1H, C$^\beta H_{2b}$, J = 13.5, 9.19 Hz), 2.62−2.65 (m, 1H, H2′), 2.21 (ddd, 1H, H2″, J = 13.1, 6.8, 3.2 Hz), 1.36 (s, 9H, C*H*$_3$-t-Boc. ^{13}C

NMR (125 MHz, DMSO-d_6) δ 172.6 *C*OOH-Cys, 156.7 C6, 155.3 *C*OO-t-Boc, 151.9 C2, 150.0 C4, 136.1 C8, 120.0 C5, 87.8 C4′, 82.9 C1′, 78.2 *C*-t-Boc, 70.5 C3′, 61.5 C5′, 53.9 C$^\alpha$, 43.2 *C*H$_2$-linker, 39.3 C2′, 31.8 C$^\beta$, 28.0 C*H*$_3$-t-Boc.

Cys−CH$_2$−dA. ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.42 (bs, 1H, $N^6 H$), 8.37 (s, 1H, H8), 8.26 (s, 1H, H2), 6.35 (Ψt, 1H, H1′, J = 6.8 Hz), 6.01 (d, 1H, $N^\alpha H$-Cys, J = 5.26 Hz), 4.66−4.77 (m, 1H, C*H*$_{2a}$-linker), 4.50−4.61 (m, 1H, C*H*$_{2b}$-linker), 4.40−4.44 (m, 1H, H3′), 3.85−3.89 (m, 1H, H4′), 3.61 (dd, 1H, H5′, J = 11.9, 4.3 Hz), 3.66−3.74 (m, 1H, C$^\alpha H$), 3.53 (dd, 1H, H5″, J = 11.9, 4.3 Hz), 2.96−3.07 (m, 2H, C$^\beta H_2$), 2.66−2.74 (m, 1H, H2′), 2.27 (ddd, 1H, H2″, J = 13.1, 6.8, 3.2 Hz), 1.33 (s, 9H, C*H*$_3$-t-Boc). ^{13}C NMR (125 MHz, DMSO-d_6) δ 172.2 *C*OOH-Cys, 154.6 *C*OO-t-Boc, 153.6 C6, 150.5 C2, 148.5 C4, 139.9 C8, 120.2 C5, 87.3 C4′, 82.7 C1′, 78.0 *C*-t-Boc, 70.4 C3′, 61.5 C5′, 55.2 C$^\alpha$, 43.0 *C*H$_2$-linker, 39.3 C2′, 34.7 C$^\beta$, 27.8 C*H*$_3$-t-Boc.

Cys−CH$_2$−dC. ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.59 (bs, 1H, $N^4 H$), 7.83 (d, 1H, H6, J = 7.5 Hz), 6.11 (Ψt, 1H, H1′ J = 6.9 Hz), 5.99 (d, 1H, $N^\alpha H$-Cys, J = 4.2 Hz), 5.74 (d, 1H, H5, J = 7.4 Hz), 4.48 (dd, 1H, C*H*$_{2a}$-linker, J = 13.4, 6.6 Hz), 4.26 (dd, 1H, C*H*$_{2b}$- linker, J = 13.4 6.5 Hz), 4.20−4.24 (m, 1H, H3′), 3.71−3.74 (m, 1H, H4′), 3.64−3.69 (m, 1H, C$^\alpha H$), 3.53−3.60 (m, 2H, H5′, H5″), 2.98 (ddd, 2H, C$^\beta H_2$, J = 42.10, 13.5 4.3 Hz), 2.11 (ddd, 1H, H2′, J = 12.9, 6.9, 4.4; Hz), 1.93 (td, 1H, H2″, J = 12.9, 6.9, 6.4 Hz), 1.35 (s, 9H, C*H*$_3$-t-Boc). ^{13}C NMR (125 MHz, DMSO-d_6) δ 171.1 *C*OOH-Cys, 162.5 C4, 154.7 C2, 154.3 *C*OO-t-Boc, 140.0 C6, 94.7 C5, 87.1 C4′, 84.5 C1′, 77.2 *C*-t-Boc, 69.5 C3′, 60.8 C5′, 55.2 C$^\alpha$, 42.3 *C*H$_2$-linker, 40.1 C2′, 35.1 C$^\beta$, 27.8 C*H*$_3$-t-Boc.

His−CH$_2$−dA. ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.83 (bs, 1H, $N^\alpha H$-His), 8.42 (s, 1H, H8), 8.34 (s, 1H, H$^{\varepsilon 1}$-His), 7.61 (s, 1H, H$^{\varepsilon 1}$-His), 6.96 (s, 1H, H$^{\delta 2}$-His), 6.4 (bs, 1, $N^6 H$), 6.36 (Ψt, 1H, H1′, J = 6.7 Hz), 5.59 (bs, 2H, C*H*$_2$-linker), 5.32 (bs, 1H, O*H*3′), 5.12 (bs, 1H, O*H*5′), 4.39−4.43 (m, 1H, H3′), 3.84−3.89 (m, 1H, H4′), 3.83−3.87 (m, 1H, C$^\alpha H$), 3.48−3.65 (m, 2H, H5′,H5″), 2.69−2.81 (m, 2H, C$^\beta H_2$-His), 2.70−2.75 (m, 1H, H2′), 2.27 (ddd, 1H, H2″, J = 12.8, 6.7, 2.8 Hz), 1.27 (s, 9H, C*H*$_3$-t-Boc). ^{13}C NMR (125 MHz, DMSO-d_6) δ 173.4 *C*OOH-His, 154.6 *C*OO-t-Boc, 148.7 C4, 140.23 C8, 139.3 C2, 138.2 C$^\gamma$-His, 136.1 C$^{\varepsilon 2}$-His, 119.5 C5, 115.5 C$^{\delta 1}$-His, 87.8 C4′, 83.7 C1′, 77.1 *C*-t-Boc, 70.6 C3′, 61.5 C5′, 54.0 C$^\alpha$-His, 49.9 *C*H$_2$-linker, 39.1 C2′, 30.2 C$^\beta$-His, 27.9 C*H*$_3$-t-Boc.

Trp−CH$_2$−dG. ^{1}H NMR (500 MHz, DMSO-d_6) δ 7.90 (s, 1H, H8), 7.06 (d, 1H, H$^{\varepsilon 3}$-Trp, J = 7.14 Hz), 6.91−6.98 (m, 1H, H$^{\eta 2}$-Trp), 6.50−6.58 (m, 2H, H$^{\zeta 2}$, H$^{\zeta 3}$-Trp), 6.14 (ψt, 1H, H1′, J = 6.8 Hz), 6.24−6.06 (m, 1H, $N^2 H$), 5.30 (bs, 2H, C*H*$_2$-linker), 4.36−4.32 (m, 1H, H3′), 4.29−4.21 (m, 1H, C*H*$^\alpha$), 3.79−3.83 (m, 1H, H4′), 3.68−3.66 (m, 1H, C$^\beta H_{2a}$), 3.58−3.44 (m, 3H, C$^\beta H_{2b}$, + H5′, H5″), 2.51−2.61 (overlapping DMSO-d_6), 2.17−2.25 (m, 1H, H2″), 1.40 (s, 4H, C*H*$_3$-t-Boc1), 1.34 (s, 5H, C*H*$_3$-t-Boc2). ^{13}C NMR (125 MHz, DMSO-d_6) δ 156.3 C6, 152.8 *C*O$_2$H, 150.0 C4, 150.0 C$^{\delta 2}$, 135.5 C8, 129.6 C$^{\varepsilon 2}$, 128.1 C$^{\eta 2}$, 122.8 C$^{\varepsilon 3}$, 116.9 C$^{\zeta 3}$, 116.8 C5, 108.4 C$^{\zeta 2}$, 87.5 C4′, 82.8 C1′, 79.8 *C*H$_2$-linker, 79.0 *C*-t-Boc, 70.8 C3′, 61.5 C5′, 56.4 C$^\alpha$, 45.3 C$^\beta$, 39.3 C2′, 27.9 C*H*$_3$-t-Boc.

Results and Discussion

Eight amino acids previously reported to form stable adducts with formaldehyde[15] were investigated (as N^α-Boc derivatives) in coupling reactions with all four nucleosides to determine which would be of interest for characterization of reactions with oligonucleotides and oligopeptides. No cross-links could be detected with Arg, Gln, Tyr, or Asn, and consistent with previous studies,[10−13] the endocyclic nitrogen of dT did not form a coupling product with any of the amino acids. Figure 1 shows the differences in extent and rate of formation; reactions between Lys and dG, and Cys and dG are essentially complete after 24 h, whereas product increased over 72 h for the less reactive combinations Cys and dA, Cys and dC, His and dA, and Trp and dG.

Table 1. Exact Masses for Reaction Products of Trinucleotides and N^α-Boc-Protected Amino Acids Cross-Linked by Formaldehyde, Determined by Negative Ion ESI-QTOF-MS

cross-links	exptl mass ([M − H]⁻)	calcd mass ([M − H]⁻)	composition	Δ (ppm)
T(G−CH₂−Lys)T	1132.3360	1132.3395	$C_{42}H_{60}N_{11}O_{22}P_2^-$	−3.1
T(TPHA-1)T	1144.3373	1144.3395	$C_{43}H_{60}N_{11}O_{22}P_2^-$	−1.9
T(TPHA-2)T	1174.3494	1174.3501	$C_{44}H_{62}N_{11}O_{23}P_2^-$	−0.6
T(G−CH₂−Trp)T	1190.3264	1190.3239	$C_{47}H_{58}N_{11}O_{22}P_2^-$	2.2
T(G−CH₂−Cys)T	1107.2526	1107.2537	$C_{39}H_{53}N_{10}O_{22}P_2S^-$	−1.0
T(A−CH₂−His)T	1125.3048	1125.3085	$C_{42}H_{55}N_{12}O_{21}P_2^-$	−3.3
T(A−CH₂−Cys)T	1091.2556	1091.2588	$C_{39}H_{53}N_{10}O_{21}P_2S^-$	−2.9
T(C−CH₂−Cys)T	1067.2464	1067.2476	$C_{38}H_{53}N_8O_{22}P_2S^-$	−1.1

The high amount of cross-links formed by Lys is of particular interest because this residue is involved in extensive DNA−protein contacts and may therefore be considered highly likely to form cross-links in vivo. The second most abundant cross-link, between Cys and dG, may also have relevance for the active site of alkylguanine alkyltransferases[24] where a Cys residue can come into proximity with a formaldehyde adduct of guanine.

Trinucleotides Cross-Linked to N^α-Boc-Protected Amino Acids. Since dT did not form adducts, we selected trinucleotides having G, A, or C flanked by T as targets for studies of trinucleotide-amino acid cross-linking. Elemental compositions of the deprotonated molecules [M − H]⁻ of cross-linked products of the trinucleotides and N^α-Boc-protected amino acids were determined by high resolution ESI-QTOF mass spectrometry and are given in Table 1. All trinucleotides yielded deprotonated molecules with summed masses expected for the trinucleotide + N^α-Boc-protected amino acid + 12 mass units, consistent with formation of a methylene link between the trinucleotide and amino acid. The reaction of TGT with Lys and formaldehyde yielded two additional products: [T(TPHA-1)T], having a composition expected for the formation of two methylene linkages and [T(TPHA-2)T], having a composition expected for two methylene linkages and a hydroxymethylene adduct.

As discussed below, NMR analysis of the products with multiple-methylene linkages establishes T(TPHA-1)T as T(10-oxo-triazino[1,2-*a*]purin-7-yl)T-substituted 2-aminohexanoic acid and T(TPHA-2)T as the corresponding *N*5-hydroxymethyl adduct. Formation of triazinane rings has precedent in the
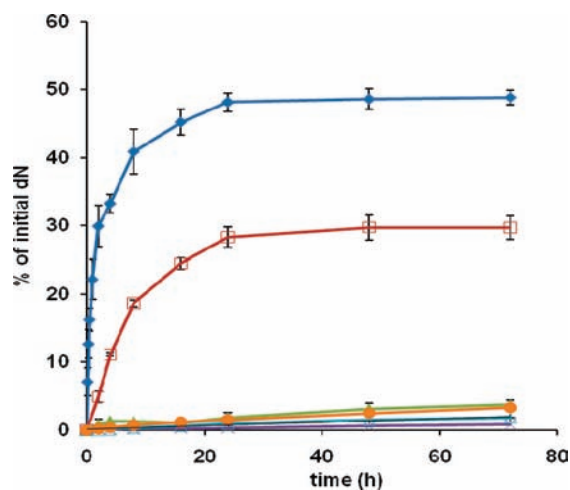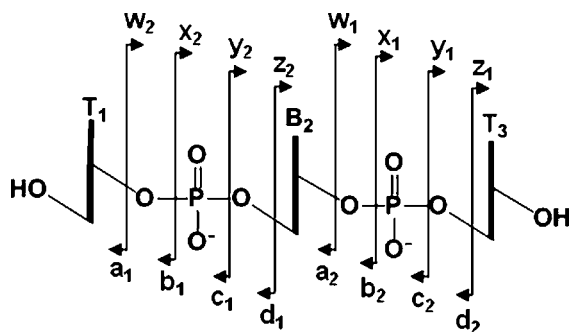

Figure 1. Formation of cross-links between 4 mM amino acid and 4 mM nucleoside treated with 50 mM formaldehyde: ♦, sum of Lys products; □, Cys−CH₂−dG; ▲, Cys−CH₂−dA; ×, Cys−CH₂−dC; △, His−CH₂−dA; ●, Trp−CH₂−dG. Relative yields are based on integration of HPLC peak areas in UV traces recorded at 254 nm.

intramolecular condensation of terminal amino nitrogens from two peptide residues with formaldehyde as well as from glycine-formaldehyde condensation with the guanidino moiety of Arg.[15] Biological significance of these products is unlikely because of the lower concentrations of formaldehyde in vivo.

Fragmentation of the reaction products between trinucleotides and amino acids was investigated by high resolution QTOF MS/MS to confirm that the target base in the second position was indeed the site of the cross-linking reaction. The high resolution data provide elemental compositions to support structural assignments of product ions. The product ion nomenclature applied for describing backbone fragmentation patterns of the trinucleotides follows the widely used convention given in Chart 2.

MS/MS spectra, acquired on an ion trap mass spectrometer, have been reported for all 64 possible unmodified trinucleotides.[25] While extensive sequential decomposition would be more likely in the ion trap than in the QTOF used in our work, the fragmentation patterns observed for the modified trinucleotides show reaction sequences similar to those reported previously.[25] Without exception, the cross-linked trinucleotides formed singly charged anions. The MS/MS spectra of the cross-linked trinucleotides were characterized by initial cleavage of the coupling linkage, with the resulting product ions undergoing backbone fragmentation. With the exception of TAT cross-linked with His, none of the ions containing an intact base-methylene-amino acid linkage was the source of backbone fragmentations. Full MS-MS spectra and assignments of product ions for all cross-linked trinucleotides are presented as Supporting Information (Figures S1−S8).

TGT Derivatives. In addition to the w_1^- ion (dT-5′-P⁻; *m/z* 321), a major ion in the MS/MS of all of the TGT derivatives was observed at nominal mass *m/z* 866 [(TGT − H + 12)⁻]. This ion corresponds in composition to the Schiff base derivative of the trinucleotide at G, which is possible only at the exocyclic N^2 and confirms that the formaldehyde-induced cross-linking reactions of TGT involve the target G. The non sequence ion from loss of neutral TH from the Schiff base adduct of the trinucleotide is also common to all the MS/MS spectra of cross-linked TGT products. In the MS/MS spectrum of T(G−CH₂−Lys)T, bonds on either side of the methylene linkage cleave, leading to a prominent ion at *m/z* 874 resulting from the loss of the Schiff base adduct of Lys in addition to the ion at *m/z* 886. Backbone cleavages of the product ions at *m/z* 874 and 886 give rise to parallel series of sequence ions w_2^-, x_2^-, y_2^-, and z_2^- separated by 12 mass units, as expected for source ions TGT⁻ and its Schiff base adduct at G. The MS/MS of

(24) Daniels, D. S.; Woo, T. T.; Luu, K. X.; Noll, D. M.; Clarke, N. D.; Pegg, A. E.; Tainer, J. A. *Nat. Struct. Mol. Biol.* **2004**, *11*, 714–720.
(25) Vrkic, A. K.; O'Hair, R. A. J.; Foote, S. *Aust. J. Chem.* **2000**, *53*, 307–319.

Chart 2

T(G−CH$_2$−Lys)T in Figure 2 is illustrative of the data obtained from QTOF analysis.

The triazino ring of T(TPHA-1)T fragments to yield the product ions TGT$^-$ and Schiff base adduct of the trinucleotide from which are derived the same parallel series of backbone fragmentations observed in the MS/MS spectrum of the singly bridged TGT-Lys product (Figure S2; formation of sodium adducts of the major ions in the MS/MS spectrum of T(TPHA-1)T should be noted). In the MS/MS of T(TPHA-2)T (Figure S3), the cross-linking structure fragments sequentially to give prominent ions corresponding to [M − H − hydroxymethylene]$^-$ (nominal m/z 1144) and [M − H − N^α-Boc − hydroxymethylene]$^-$ (nominal m/z 1044) in addition to the Schiff base adduct of TGT at m/z 866, which is progenitor of the sole series of observed sequence ions w_2^-, x_2^-, y_2^-, and z_2^-.

The MS/MS spectrum of T(G−CH$_2$−Cys)T (Figure S4) shows a single fragmentation pathway consistent with loss of Cys to give the Schiff base adduct of TGT at m/z 866 and formation of the sequence ions w_2^-, x_2^- y_2^-, and z_2^- from the expected backbone fragmentations. The MS/MS spectrum of T(G−CH$_2$−Trp)T (Figure S5) yielded product ions in low abundance. The Schiff base adduct of TGT and backbone cleavage ions w_2^-, x_2^- y_2^-, and z_2^- were present at the expected nominal mass-to-charge ratios; however, the accuracy of mass measurements was low.

TAT and TCT Derivatives. Loss of N^α-Boc (m/z 1025) and (N^α-Boc + His) (m/z 870) were prominent ions in the MS/MS spectrum of T(A−CH$_2$−His)T (Figure S6). The ion at m/z 1025 gave the non sequence ion (M − H − TH)$^-$ and sequence ions w_2^- and x_2^- from backbone fragmentation in which the cross-link is intact (the only instance of backbone fragmentations with an intact cross-link). The Schiff base derivative from loss of (N^α-Boc + His) (m/z 870) yielded the parallel series of sequence ions w_2^- and x_2^-, fixing the point of cross-link attachment at exocyclic N^6 of adenine. The MS/MS spectrum of the deprotonated molecule T(A−CH$_2$−Cys)T (Figure S7) features (y_2 − B$_2$)$^-$, w_1^- and T$^-$ (base peak) ions along with a prominent ion at nominal mass m/z 545. However, no ion containing the Schiff base adduct of A was detected. The nominal mass m/z 545 corresponds to a 2′,3′-dideoxy-2′,3′-dehydro-AMP or 2′-deoxyadenosine-5′-phosphenate linked through a methylene bridge to N^α-Boc-His. The exact mass differs by +10.8 ppm from that calculated for the composition of the proposed structures, within sufficient tolerance to support the methylene bridge between Cys and the target N^6 of A. However, a backbone fragmentation pathway yielding ions of either of the proposed structures was not reported for any of the 64 possible unmodified trinucleotides or observed for any of the other cross-linked trinucleotides in this study. Thus, support for the proposed structure of the cross-link relies on the NMR studies of the formaldehyde-derived

cross-link between dA and Cys discussed below. Although product ions in MS/MS spectrum of T(C−CH$_2$−Cys)T (Figure S8) were in low abundance, the exact mass of the base peak lies within an acceptable tolerance (4.1 ppm) for a composition corresponding to the Schiff base adduct of trinucleotide which is consistent with a methylene bridge between the C-N^4 and the Cys sulfhydryl group. Sequence ions w_2^-, x_2^-, and w_1^- were also present at the expected nominal masses, in addition to the nonsequence ion (M − H − TH)$^-$.

Peptides Cross-Linked to Deoxynucleosides. We reacted dG, dA, and dC with formaldehyde and N-terminal acetylated 8-mers in which position 5 contained one of the four residues that had been established as targets for cross-linking in the screening reactions. The exact masses of the protonated molecules are given in Table 2. As in the case of the trinucleotides, all of the 8-mers yielded singly charged ions with compositions corresponding to formation of a methylene cross-link, and the Lys-containing 8-mers yielded two additional products consistent with formation of the tricyclic cross-linked structures TPHA-1 and TPHA-2. In accord with a report on the formation of formaldehyde adducts of 7-mer and 9-mer peptides under conditions similar to ours,[15] we found no evidence of peptide−peptide cross-linking by LC−MS analysis over the range of mass-to-charge ratios expected for cross-linked peptides.

By MS/MS (Figures S9−S16 in Supporting Information), the protonated molecules undergo initial fragmentation of the glycosidic bond or the bridging structure, and the resulting product ions undergo backbone fragmentations similar to those observed for unmodified peptides, giving predominantly y_n, a_n and b_n ions.[26] Analysis of MS/MS spectra of the protonated molecules definitively established the site of formaldehyde-induced cross-linking for all of the peptides with the exception of the His-containing 8-mer cross-linked with dA.

Backbone fragmentations in the MS/MS spectra of the Lys- and Trp-containing 8-mers with a single methylene link to dG (Figures S9 and S10) are derived from the intermediate product ions in which the cross-link has broken to yield a peptide sequence containing the Schiff base of Lys (Scheme 2) or the 2-methylene indole derivative of Trp. Figure 3 illustrates the MS/MS of the singly bridged Lys-containing octapeptide.

The mass difference between the b_5 and b_4 ions corresponds to the Schiff base adduct of the target Lys (or 2-methylene-substituted Trp) at position 5, which establishes the point of attachment of the methylene bridge at the predicted target residues. The modification of the Lys and Trp residues is further confirmed by the identification of y_4 ions, and the determination that the mass differences between (b_5 + y_4) and the intermediate product ions yielding the backbone fragmentation series (m/z 742 for the Lys 8-mer and 758 for the Trp 8-mer) correspond to the mass of the modified residue. The product ions representing loss of the Schiff base adduct of dG from the cross-linked peptides establishes the exocyclic N^2 of dG as the second point of attachment of the methylene bridge, as discussed above for the cross-linked trinucleotides.

In the MS/MS spectrum of the Cys-containing 8-mer cross-linked to dC (Figure S11), a series of low-abundance b_n ions originates from the intermediate product ion (m/z 697) after loss of the deoxyribose from the protonated molecule. The difference in mass between the b_5 and b_4 ions of this b_n series corresponds to the mass of the Cys−CH$_2$−C unit, confirming that the cross-link is attached to the peptide at Cys. Additional support for
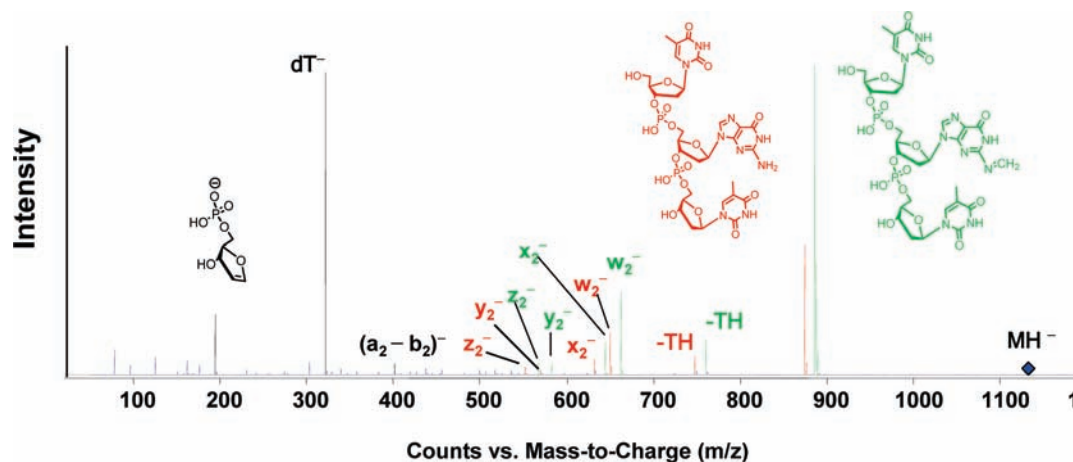
(26) Paizs, B.; Suhai, S. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 103–113.

Figure 2. ESI-QTOF-MS/MS of the deprotonated molecule T(G−CH$_2$−Lys)T: (green) ions containing Schiff base guanine; (red) ions containing unmodified guanine; (black) ions that do not contain guanine.

Table 2. Exact Masses of Protonated 8-Mers Cross-Linked to Deoxynucleosides by Formaldehyde Determined by Positive Ion ESI-QTOF-MS

cross-links	exptl mass ([MH]$^+$)	calcd mass ([MH]$^+$)	composition	Δ (ppm)
acetyl-VEGG(K−CH$_2$−dG)GAA	1009.4687	1009.4697	C$_{41}$H$_{65}$N$_{14}$O$_{16}^+$	−1.0
acetyl-VEGG(TPHA-1) GAA	1021.4702	1021.4697	C$_{42}$H$_{65}$N$_{14}$O$_{16}^+$	0.5
acetyl-VEGG(TPHA-2)GAA	1051.4805	1051.4803	C$_{43}$H$_{67}$N$_{14}$O$_{17}^+$	0.2
acetyl-GEGG(W−CH$_2$−dG)GAA	1025.4048	1025.4071	C$_{43}$H$_{57}$N$_{14}$O$_{16}^+$	−2.2
acetyl-GEGG(C−CH$_2$−dG)GAA	942.3372	942.3370	C$_{35}$H$_{52}$N$_{13}$O$_{16}$S$^+$	0.2
acetyl-VEGG(H−CH$_2$−dA)GAA	1002.4383	1002.4388	C$_{41}$H$_{60}$N$_{15}$O$_{15}^+$	−0.5
acetyl-GEGG(C−CH$_2$−dC)GAA	902.3304	902.3309	C$_{34}$H$_{52}$N$_{11}$O$_{16}$S$^+$	−0.6
acetyl-GEGG(C−CH$_2$−dA)GAA	926.3427	926.3421	C$_{35}$H$_{52}$N$_{13}$O$_{15}$S$^+$	0.6

Scheme 2. Fragmentation of AcVEGG(K−CH$_2$−dG)GAA
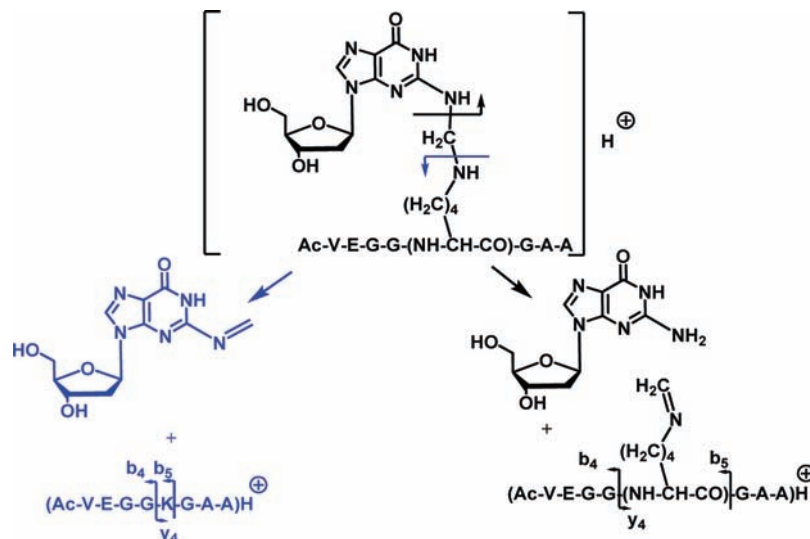
the cross-link to Cys is a second series of low abundance b$_5$, b$_6$, and b$_7$ ions having compositions consistent with backbone fragmentation of the product ion from cleavage of the C−S bond of methylene bridge which transforms Cys to an α-amidoacrylic acid residue. As required by this scheme, the difference in mass between the b$_5$ (*m/z* 412) and b$_4$ (*m/z* 343), which does not contain a modified residue, is 69 mass units, corresponding to α-amidoacrylic acid at position 5. Major backbone fragmentation series for the Cys-containing 8-mers cross-linked with dA (Figure S12) and dG (Figure S13) originate from the product ions subsequent to the loss of the Schiff base derivatives of the nucleosides and therefore are not informative with respect to

the site of attachment of the methylene bridge to the peptide. However, the b$_n$ series from the intermediate product ions containing α-amidoacrylic acid at position 5 are present in the MS/MS spectra of both cross-linked 8-mers, fixing the methylene bridge attachment at Cys.

The MS/MS spectrum of the Lys-containing 8-mer linked by two methylene groups through formation of a fused triazino ring (Figure S14) is consistent with initial fragmentation of the triazino ring followed by proton transfer to give Schiff base adducts of both dG and the 8-mer (nominal mass, *m/z* 742) according to Scheme 3, with the charge residing predominantly on the 8-mer fragment. The Schiff base adduct of dG loses
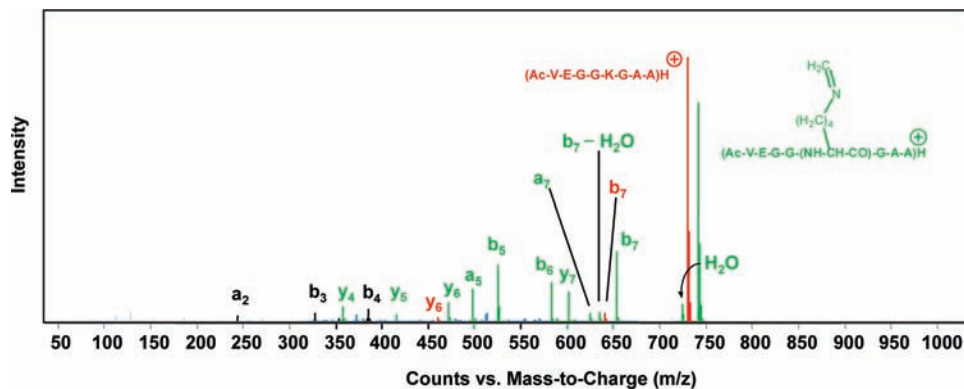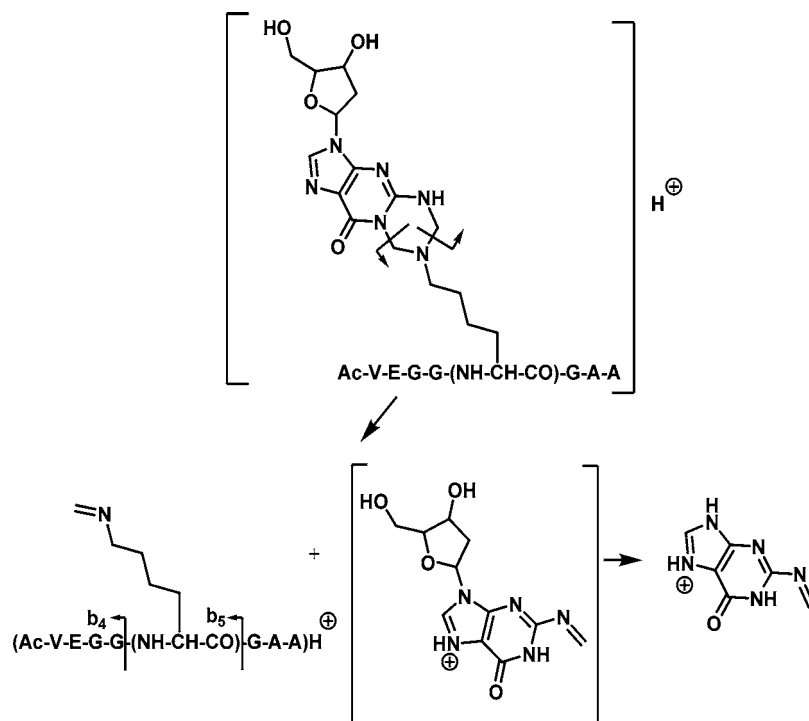
Figure 3. ESI-QTOF-MS/MS spectrum of the protonated molecule acetyl-VEGG(K−CH₂−dG)GAA: (green) ions containing Schiff base derivative of Lys; (red) ions containing unmodified Lys; (black) ions that do not contain residue 5.

Scheme 3. Fragmentation of AcVEGG(TPHA-1)GAA

deoxyribose to give an ion at *m/z* 164 in low abundance. The absence of an ion corresponding to [MH − dG]⁺ (nominal mass *m/z* 730) accompanied by product ions from its backbone fragmentation, such as observed above for a single cross-linking methylene, supports the triazino structure. The difference in mass between b_5 and b_4 ions corresponds to the Schiff base adduct at the terminal -NH₂ of the Lys residue as required by Scheme 3.

The MS/MS spectrum of the product the containing the tricyclic (TPHA-2) cross-linking structure with an $N5$ hydroxymethyl-substituted triazinopurine (Figure S15) is dominated by two series of ions derived from peptide backbone fragmentations of the intermediate product ions shown in Scheme 4. The y_4, b_4, and b_5 ions can be identified for both series, and the compositions establish that the terminal -NH₂ of Lys has been incorporated into the linking structure. The sequential fragmentations of the protonated molecule that lead to the intermediate products in Scheme 4 are suggested by the presence of product ions both at *m/z* 164 for the protonated

N^2-Schiff base adduct of G and at *m/z* 178 compatible with the protonated tricyclic base 1,N^2-ethano-G.

In the MS/MS spectrum of the His-containing 8-mer cross-linked to dA, a single series of backbone fragmentations arises from an intermediate product ion via loss of the Schiff base adduct of dA. This fragmentation pattern is not informative regarding the site of methylene attachment to the 8-mer, and the structure of the cross-link is inferred from the structures determined by NMR studies for cross-linked monomeric units dA−CH₂−His.

$N^α$-Boc-Protected Amino Acids Cross-Linked to Deoxynucleosides. Formaldehyde-induced coupling reactions between the $N^α$-Boc-protected amino acids and nucleoside monomers were investigated on a scale that allowed detailed structural determination of the linkages by NMR spectrometry. Table 3 gives the exact masses and corresponding elemental compositions of the cross-linked products, and MS/MS spectra are given in Figures S17−S24 in Supporting Information. LC−MS analysis revealed no cross-links formed between deoxynucleo-

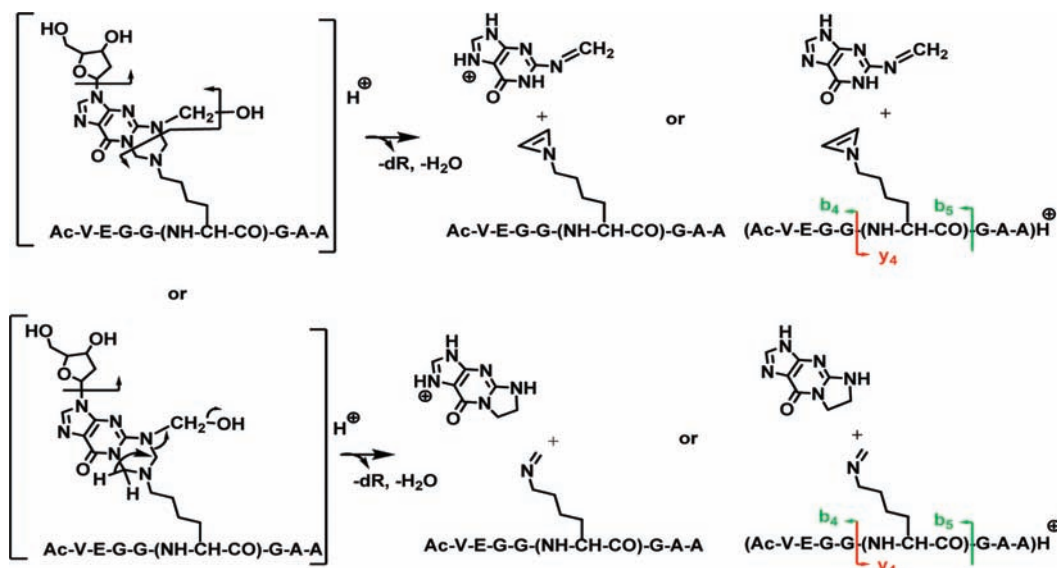**Scheme 4.** Fragmentation of AcVEGG(TPHA-2)GAA

Table 3. Exact Masses of Formaldehyde-Induced Deoxynucleoside-Amino Acid Cross-Links by FTICR-MS

cross-links	exptl mass ([MH]$^+$)	calcd mass ([MH]$^+$)	composition	Δ (ppm)
Lys—CH$_2$—dG	526.2622	526.2619	C$_{22}$H$_{35}$N$_7$O$_8$	0.6
TPHA-1	538.2623	538.2619	C$_{23}$H$_{35}$N$_7$O$_8$	0.7
TPHA-2	568.2726	568.2725	C$_{24}$H$_{37}$N$_7$O$_9$	0.2
Trp—CH$_2$—dG	584.2465	584.2463	C$_{27}$H$_{33}$N$_7$O$_8$	0.3
Cys—CH$_2$—dG	501.1763	501.1762	C$_{19}$H$_{28}$N$_6$O$_8$S	0.2
His—CH$_2$—dA	519.2312	519.2310	C$_{22}$H$_{30}$N$_8$O$_7$	0.4
Cys—CH$_2$—dA	485.1813	485.1813	C$_{19}$H$_{28}$N$_6$O$_7$S	0
Cys—CH$_2$—dC	461.1701	461.1700	C$_{19}$H$_{28}$N$_6$O$_8$S	0.2

sides or between amino acids. As observed for the trinucleotide and 8-mer, coupling products of Lys with dG incorporated one or two methylene groups or two methylenes with a hydroxymethyl adduct. All of the Lys coupling products were labile and were isolated and stored at subambient temperature. The lability of the Lys coupling products is consistent with reports that formaldehyde-induced DPCs involving the Lys- and Arg-rich major histones are hydrolytically unstable.[15,20] The structures of formaldehyde-induced cross-links formed between the amino acids and nucleosides are given in Chart 1.

Lys Cross-Links with dG. In the presence of formaldehyde, Lys and dG formed three coupling products with the structures given in Chart 1. All three structures were detected in coupling reactions run at the lowest formaldehyde concentration. The proportion of tricyclic and *N*-hydroxymethyl-substituted tricyclic nucleosides increased with increasing formaldehyde concentration and reaction time (confirmed with ^{13}C-formaldehyde), consistent with progressive incorporation of formaldehyde molecules into the initial coupling product (Figure S25). As described above, the products were all labile at ambient temperature and required care in isolation and characterization. The tricyclic structures TPHA-1 and TPHA-2 were definitively established by mass spectrometry and NMR. Lys—CH$_2$—dG was too labile for characterization by NMR, and its assignment as a product of cross-linking is based on indirect evidence from the analyses below.

Exact mass measurements by ESI-MS in the positive ion mode of major ions at *m/z* 526, 538, and 568 in the coupling reaction mixture were compatible with the protonated molecules Lys—CH$_2$—dG, TPHA-1 and TPHA-2, respectively (Table 3).

When ^{13}C-formaldehyde was used in the coupling reaction, the protonated molecules are observed at *m/z* 527, 540, and 571, confirming formaldehyde as the origin of the methylene linkers and the hydroxymethylene group.

Well-resolved peaks were collected by semipreparative HPLC (Figure 4) and characterized by NMR (Figures S26—S31 in Supporting Information). ^{1}H NMR spectra of the fractions collected at 17.2 and 26.5 min indicated they were dG-containing mixtures, suggesting postcolumn decomposition.

The peak eluting at 17.2 min contained dG/Lys/TPHA-1 in 1:1:1 ratio, whereas the peak at 26.5 min contained a 1:1.4:7 mixture of dG/TPHA-1/TPHA-2 based on signals in the H1′ region of the NMR spectra. In the early eluting fraction, Lys and dG were readily identified by comparison of well resolved signals in critical regions of the spectrum with authentic standards, and in the late-eluting fraction, signals of TPHA-1 and TPHA-2 were readily distinguished because of the difference in signal intensities and peak integrals. The composition of the fractions may be explained by co-chromatography of TPHA-1 and Lys—CH$_2$—dG at 17.2 min and sequential uncoupling of formaldehyde units in both fractions. The absence of signals in either fraction attributable to singly bridged Lys—CH$_2$—dG suggests that Lys—CH$_2$—dG degrades significantly faster than the tricyclic adducts. The appearance of dG and TPHA-1 in the 26.5 min fraction suggests that both TPHA-1 and TPHA-2 are unstable. The mixtures observed in the ^{1}H NMR spectra are consistent with the degradation at 37 °C of the 17.2 and 26.5 min fractions monitored by mass spectrometry (Figure S32). Uncoupling is supported additionally by a signal
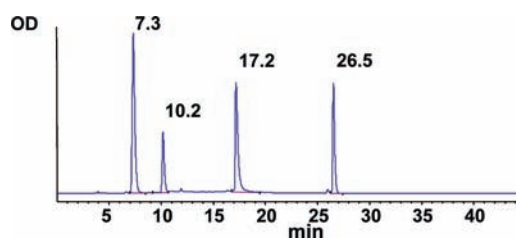

Figure 4. Trace (254 nm) of semipreparative HPLC of the reaction mixture from the formaldehyde-induced coupling of Lys and dG. The peaks at 17.2 and 26.5 min were characterized by NMR. Peaks at 7.3 and 10.2 min are dG and *N*^{2}-hydroxymethyl-dG, respectively.
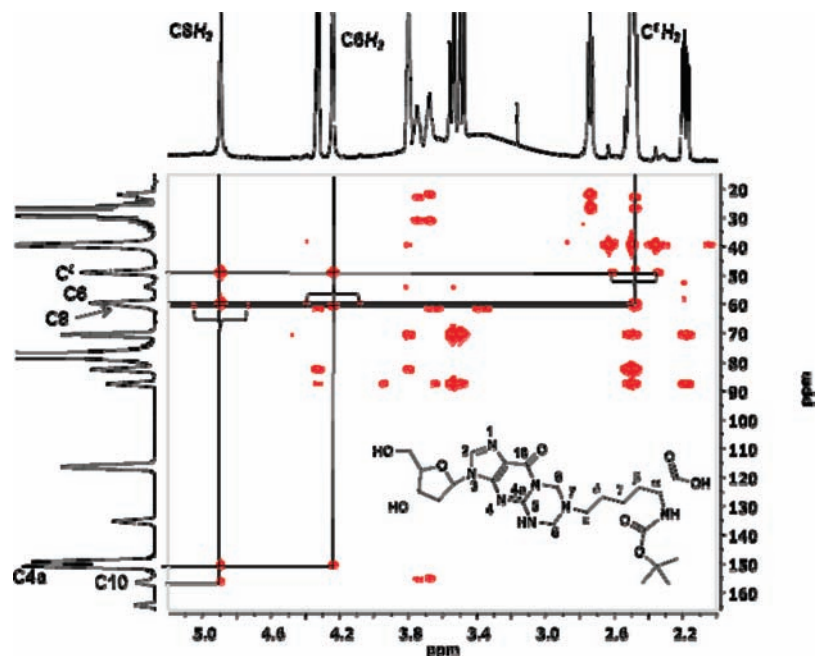
Figure 5. HPLC peak at 17.2 min containing TPHA-1: expansion of the HMBC spectrum between 2.0 and 5.2 ppm on the [1]H-axis and 15 and 166 ppm on the ^{13}C-axis. Key signals are identified on the marginal ^{1}H and ^{13}C traces. Unsuppressed 1-bond couplings are indicated by brackets.
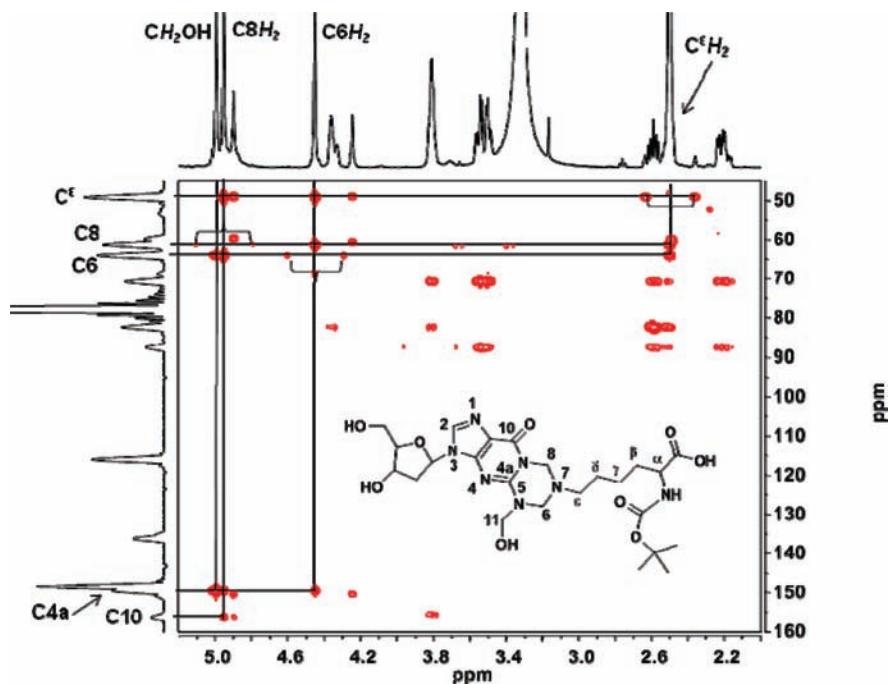

Figure 6. HPLC peak at 26.5 min containing TPHA-2: expansion of the HMBC spectrum between 2.0 and 5.2 ppm on the [1]H-axis and 40 and 160 ppm on the ^{13}C-axis. Key signals are identified on the marginal ^{1}H and ^{13}C traces. Unsuppressed 1-bond couplings are indicated by brackets.

in the ^{1}H NMR spectra of both early and late-eluting fractions attributable to formaldehyde or formaldehyde hydrate (Figures S26, S29) in accord with reported reversibility of formaldehyde-induced DPCs.[20] The formaldehyde/formaldehyde hydrate assignment is based on the appearance of a proton singlet below 8 ppm having unsuppressed 1-bond coupling with a carbon signal at 164.4 ppm that has no connectivity with any component in the HMBC spectra of the mixtures (Figures S26, S29).

In each fraction, the NMR signals of the components of the mixtures could be resolved, allowing the structures of TPHA-1 and TPHA-2 to be unambiguously assigned. Critical in estab-

lishing the triazine ring is the presence of two formaldehyde-derived methylenes and identification of the connectivities between the methylene groups, guanine, and Lys moieties. Figures 5 and 6 show expansions of the HMBC spectra that are key to establishing the fused triazine linkages. In the HMBC spectrum of TPHA-1 (Figure 5), a broad two-proton singlet at 4.89 with three-bond coupling to a ^{13}C signal at 59.5 ppm and a second two-proton singlet at 4.24 ppm with $^3J_{C-H}$ coupling to a ^{13}C signal at 60.5 ppm are assigned to methylene groups at positions 8 and 6, respectively, of the triazino[1,2-a]purine framework. Unsuppressed one-bond couplings, confirmed by

the corresponding C/H cross peaks in the HSQC spectrum (Figure S27) are observed for both methylene signals and allow assignment of the methylene carbon shifts. The protons attached to the methylene carbon assigned to C8 show the expected connectivities within the tricyclic framework between C4a at 150.2 ppm and C10 at 155.8 ppm, while the methylene protons at position 6 couples only with the carbon at C4a. Attachment of the hexanoic acid moiety at N7 is confirmed by cross peaks between the methylene protons at positions 6 and 8 and a carbon signal at 49.2 ppm, which can be assigned to hexanoic acid C^ε by virtue of unsuppressed one-bond coupling in the HMBC spectrum and the corresponding C/H cross peak in the HSQC spectrum. The complementary three-bond coupling between the methylene protons attached to C^ε (overlapping with the DMSO signal and sugar H2″ signals at ~2.5 ppm) and C6 and C8 is also observed.

By a similar analysis of the HMBC spectrum in Figure 6, connectivities can be identified to confirm the triazino-[1,2*a*]purine framework of TPHA-2, the major component of the late-eluting peak. The critical cross peaks are between C8H_2 (4.95 ppm) and C6 (63.9 ppm), C4a (149.4 ppm), C10 (156.0 ppm), and C^ε (49.0 ppm); between C6H_2 at 4.45 ppm and C8 (61.2 ppm), C4a (149.4 ppm), and C^ε (49.0 ppm); and between $C^\varepsilon H_2$ (2.5 ppm, overlapping with DMSO) and C6 and C8. A methylene proton signal at 4.99 ppm can be assigned to the hydroxymethylene group with the position of attachment fixed at N5 by virtue of cross peaks with C6 and C4a. In the ^{1}H NMR spectrum (Figure S29), a singlet, assignable to formaldehyde by the rationale described above, strongly supports the conclusion that the presence of THPA-1 in the late-eluting fraction results from the hydrolytic elimination of the N5-hydroxymethylene substituent of TPHA-2. The dynamic nature of formaldehyde-induced Lys-dG cross-links (Scheme S1) is supported by the lability of the products and the presence of formaldehyde in the NMR spectra of TPHA-1 and TPHA-2, which indicate that the condensations are reversible.

Cys Cross-Links with dG, dA, and dC. A single product was isolated from the cross-linking of Cys and dG with formaldehyde. In contrast to the cross-linked products of Lys, the product with Cys was stable and readily isolated and purified. The exact mass corresponded in elemental composition to addition of one methylene group. ESI-MS/MS of the protonated molecule (Figure S20) yielded product ions from loss of deoxyribose and *t*-Boc. The base peak of the MS/MS spectrum appears at *m/z* 164 (Figure S20), in accord with fragmentation to a Schiff base derivative of guanine, which indicates that the methylene cross-link bridges N^2 of guanine and the Cys sulfhydryl group.

NMR spectrometry (Figures S33, 34) confirmed guanine N^2 and Cys SH as the points of attachment of the methylene cross-link. The carbon and proton signals of formaldehyde-derived methylene cross-link are assigned from the HMBC spectrum on the basis of unsuppressed one-bond coupling between the carbon signal at 43.2 ppm and a 2-proton methylene multiplet centered at 4.52 ppm. Attachment at the Cys sulfhydryl is then fixed by coupling between the diastereotopic $C^\beta H_2$ protons of Cys at 2.84 and 3.03 ppm and the formaldehyde-derived methylene bridging carbon, while attachment at guanine N^2 is fixed by coupling between the bridging methylene protons and C2 of guanine at 151.9 pm. The presence of the guanine imino N1*H* at 10.82 ppm in the ^{1}H NMR spectrum, unambiguously assigned from coupling with guanine C5 at 120 ppm in the HMBC spectrum (Figure S33), rules out attachment of the cross-link at guanine N1. Attachment to Cys at N^α is ruled out by

identification of Cys-$N^\alpha H$ as one of two incompletely resolved proton signals at ~7.10 ppm based on cross peaks with Cys C^α (53.9 ppm) and C^β (31.8 ppm).

Cys formed stable cross-links with dA and dC in minor amounts. By high resolution mass spectrometry, both products had empirical compositions consistent with a methylene bridge linking the nucleoside and Cys. Both products were characterized by NMR (Figures S35–S38 in Supporting Information). Critical C/H connectivities in the HMBC spectra (Figures S35, S37) were observed between the proton signals of the bridging methylene and a carbon signal of the base, and between the bridging methylene and Cys $^\beta CH_2$. In the case of Cys−CH_2−dA, attachment to dA at the exocyclic amino group was established by coupling between the diastereotopic methylene proton multiplets of the bridging groups centered at 4.56 and 4.73 ppm and C6 of dA at 153.6 ppm (Figure S35). Attachment to the Cys sulfhydryl group was established by coupling between the diastereotopic methylene bridge protons and the C^β signal of Cys at 34.7 ppm. Identification of a one-proton N^6H signal confirmed the substitution at the exocyclic amino group of dA, while presence of the Boc-$N^\alpha H$ proton signal ruled out attachment at the α-amino group of Cys. The cross-link between dC and Cys was similarly determined to be between the exocyclic amino group of dC and the Cys sulfhydryl because of the observed coupling between the bridging methylene protons at 4.48 and 4.26 ppm and C4 of dC at 162.5 ppm and between the bridging methylene protons and C^β of Cys at 35.1 ppm (Figure S37). As in the case of Cys−CH_2−dA, the position of the cross-link determined from the HMBC spectrum is supported by one-proton signals attributable to N^4H of dC and Boc-$N^\alpha H$.

His Cross-Link with dA. A low yield of stable cross-linked product was isolated from the reaction of His and dA with formaldehyde, incorporating one methylene linkage identified by accurate mass measurement. Despite the absence of C,H cross-peaks between the linker methylene and His, attachment at $N^{\varepsilon 2}$ could readily be established by the presence of imidazole ring C−H proton signals $H^{\delta 2}$ at 6.96 and $H^{\varepsilon 1}$ at 7.61 ppm, which were definitively assigned by unsuppressed one-bond coupling to carbon signals at 115.5 and 136.1 ppm, respectively, in the HMBC spectrum (Figure S39). Site of attachment to dA rests on the presence of a signal having an integral value of one proton assigned to N^6H. When the cross-link is derived from ^{13}C-formaldehyde, a broad two-proton signal at 5.59 ppm in the ^{1}H NMR spectrum is replaced by a two-proton doublet centered at 5.59 ppm, $^1J_{C-H}$ = 153.2 Hz (Figure S40), and can be assigned to the formaldehyde-derived methylene bridge. This signal has the expected NOESY interactions with the imidazole protons in the ROESY spectrum (Figure S41).

Trp Cross-Link with dG. The Trp coupling product incorporated one methylene cross-link determined by accurate mass measurement. The site of attachment to Trp can be definitively fixed at C2 of the indole ring and is analogous to the reported formation of tetrahydro-β-carboline from Trp via intramolecular incorporation of a formaldehyde-derived methylene when the α-amino group is unprotected.[27,28] The proton signal of the formaldehyde-derived methylene forming the cross-link to dG is identified by a broad two-proton resonance at 5.30 ppm in the ^{1}H NMR spectrum attached to a carbon having a signal at 79.8 ppm in the HSQC spectrum (Figure S42), which is also

(27) Tammler, U.; Quillan, J. M.; Lehmann, J.; Sadee, W.; Kassack, M. U. *Eur. J. Med. Chem.* **2003**, *38*, 481–493.
(28) Zhao, M.; Bi, L.; Bi, W.; Wang, C.; Yang, Z.; Ju, J.; Peng, S. *Bioorg. Med. Chem.* **2006**, *14*, 4761–4774.

detected by unsuppressed one-bond coupling ($^1J_{C-H}$ = 164.4 Hz; Figure S43) in the HMBC spectrum. Attachment of the methylene at indole C2 is supported by the detection of indole benzo-ring carbons as the only indole carbons having attached protons in the HSQC spectrum and by NOESY interactions between the bridge methylene and Trp $C^\beta H_2$ proton signals in the ROESY spectrum (Figure S44), which would not be observed if attachment of the methylene were at the indole nitrogen. While the proton-bearing carbons of the Trp indole ring were readily identified in the HSQC spectrum, overlapping of the quaternary carbon signals of both indole and dG moieties precludes establishing connectivity between the methylene bridge and dG from the HMBC spectrum. Attachment of the methylene bridge at the exocyclic amino group is inferred from the absence of a 2-proton signal assignable to an unsubstituted exocyclic amino group of guanine and the fragmentation of the protonated molecule to yield the Schiff base of dG as the base peak in the MS/MS spectrum (Figure S24).

Conclusion

We have investigated the formation of formaldehyde-induced cross-links between nucleosides and amino acids and provide the first rigorous structural characterizations of DNA−protein cross-links induced by coupling with formaldehyde. Eight cross-linked structures have been characterized. In five of these structures, the cross-linking arose via a methylene bridge connecting the exocyclic amino group of a nucleoside to a nucleophilic nitrogen or sulfur of an amino acid side chain. In the case of Trp, the points of attachment were the N^2 exocyclic amino group of dG and C2 of the indole ring. Lys, in addition to forming an expected dG-N^2−CH$_2$−N$^\varepsilon$-Lys cross-link, formed two tricyclic adducts with dG, a 10-oxo-triazino[1,2-*a*]purinyl derivative incorporating two formaldehyde molecules and a 5-hydroxymethyl-10-oxo-triazino[1,2-*a*]purinyl adduct incorporating three formaldehyde molecules. While the formation of the tricyclic adducts was favored at formaldehyde levels of 50 and 100 mM, all three adducts were identified at the lowest

levels tested (5 mM). The three Lys cross-linked products were labile in solution, and the cross-link formed by a single methylene bridge was too unstable for characterization by NMR spectrometry. The lability of the Lys-dG cross-links supports the reported reversibility of formaldehyde-induced cross-links formed in vivo between histones and DNA.[20] The cross-links Cys−CH$_2$−dG, Cys−CH$_2$−dA, and Cys−CH$_2$−dC were stable and readily isolated and characterized. The cross-links characterized in this work will contribute to a better understanding of DPC formation induced by formaldehyde and of the mode of action of this known human carcinogen. The stable structures identified in this study have potential as biomarkers for the occurrence of DPCs following formaldehyde exposure. Furthermore, by using [^{13}CD$_2$]-formaldehyde, specific exposure-related exogenous cross-links can be differentiated from endogenous cross-links. Such sophisticated methods will be necessary to carefully examine formaldehyde adducts formed at the site of contact versus adduction at distant sites to determine the plausibility of inhaled formaldehyde as a causative agent for leukemia.[29]

Supporting Information Available: Complete HMBC and ROESY NMR spectra of N^α-Boc-amino acid-deoxynucleoside cross-links. Full HSQC spectra of TPHA-1, TPHA-2, and Trp−CH$_2$−dG. Complete QTOF-MS/MS spectra of N^α-Boc-amino acid-trinucleotide cross-links and complete QTOF-MS/MS spectra of *N*-terminal acetylated octapeptide-deoxynucleoside cross-links. This material is available free of charge via the Internet at http://pubs.acs.org.

JA908282F

(29) Zhang, L.; Steinmaus, C.; Eastmond, D. A.; Xin, X. K.; Smith, M. T. *Mutat. Res.* **2009**, *681*, 150–168.