ChemInform Abstract: Geometric Parameters in Nucleic Acids: Nitrogenous Bases

ARTICLE *in* CHEMINFORM · JANUARY 1996

Impact Factor: 0.74 · DOI: 10.1021/ja952883d

Geometric Parameters in Nucleic Acids: Nitrogenous Bases

Lester Clowney, Shri C. Jain, A. R. Srinivasan, John Westbrook, Wilma K. Olson, and Helen M. Berman*

Contribution from the Department of Chemistry, Rutgers University, Piscataway, New Jersey 08855-0939

Received August 21, 1995[⊗]

Abstract: We present estimates of the bond-length and bond-angle parameters for the nitrogenous base side groups of nucleic acids. These values are the result of a statistical survey of small molecules in the Cambridge Structural Database for which high-resolution X-ray and neutron crystal structures are available. The statistics include arithmetic means and standard deviations for the different samples, as well as comparisons of the population distributions for sugar- and non-sugar-derivatized bases. These accumulated data provide appropriate target values for refinements of oligonucleotide structures, as well as sets of standard atomic coordinates for the five common bases.

Introduction

X-ray crystallographic determinations of the structures of nucleic acids and nucleic acid−protein complexes have increased dramatically over the last several years. A survey of the Nucleic Acid Database (NDB)[1] shows that there are over 300 solved oligonucleotide structures and 50 nucleic acid complexes currently available; the number of structure determinations continues to increase. The refinement of such oligonucleotides, most of which are determined with resolution poorer than 1 Å, necessitates the use of geometric restraints. Thus, it is critical to have values for the target bond lengths and valence angles that are as accurate as possible. The best source of these target values are high-resolution crystal structures of nucleic acid analogs, and more than 13 years have passed since Taylor and Kennard[2] first analyzed the bonding geometries of nucleic acid base moieties in the Cambridge Structural Database (CSD).[3] Since then, the number of high-resolution structures containing the nucleobases available has nearly doubled, and there are now sufficient data to determine independent values for uracil and thymine. The larger sample size further allows for the use of more stringent criteria in selecting structures to include in the analyses. For instance, the maximum *R* factor of structures included in this survey was 6% (compared to a value of 8% in Taylor and Kennard[2]) and the maximum average error in C−C bonds (average estimated standard deviation, esd) was 0.01 Å (versus a value of 0.015 Å in Taylor and Kennard[2]). An updated analysis of the base structures in the CSD is presented here.

Methods

Selection of Structures. Sets of high-resolution structures containing the five nitrogenous bases—cytosine, thymine, uracil, adenine, and guanine (Figure 1)—were initially collected from the CSD using the program QUEST.[3] Protonated cytosines and adenines were treated independently from neutral species, while protonated guanines were excluded due to the small sample size. The sampling criteria were established on the basis of both chemical and crystallographic considerations.

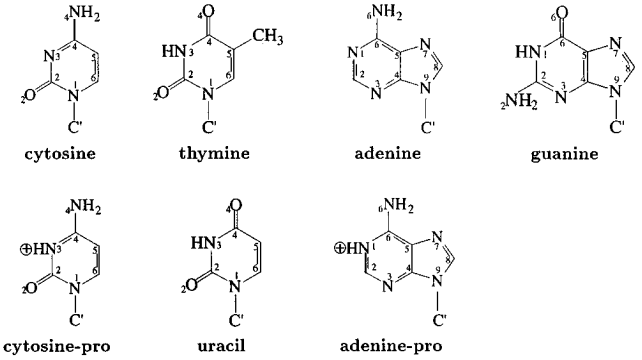

Figure 1. Structures of the nitrogenous bases which are considered in this survey. The N1 nitrogen atoms of pyrimidines and the N9 nitrogens of purines are shown in a linkage to the C1′ carbon of the sugar ring.

Only structures with *R* values better than 6% were used. This value was chosen after considering at what value of the *R* factor there is a statistically significant reduction in the standard deviations of bond lengths and valence angles. Subsets of bond lengths or bond angles were examined where increasingly smaller *R* factors were used as cutoffs for the structures to include, i.e., the initial set included all structures with an *R* factor less than 8%, the second set included those with a maximum *R* factor of 7.5%, and so on, using cutoffs down to $R = 4.5\%$ at 0.5% increments. Means and standard deviations were determined for each set, and the *F* test (see below) was used to compare the variances of the initial set, where the value of *R* was 8%, with those of each succeeding set. A significant reduction in the sample variance was found at $R = 6\%$.

The selected structures had to meet two additional crystallographic criteria. The statistical sample was limited to structures with (1) resolution better than 1 Å, and (2) esd's for C−C bond lengths less than 0.01 Å. Using these criteria, most hydrogen atoms were located directly or with difference Fourier maps.

Several chemical criteria were also used. Only pyrimidines substituted at N1 and purines substituted at N9 were selected. Of these structures, those with a sugar substitution were also treated separately to see if sugar derivatization had a significant effect on base geometry. Neutral bases and protonated bases were considered separately, while hemi-protonated bases, crystal structures with transition metals, atoms as heavy as bromine (Br), and oligonucleotides were excluded from consideration.

The CSD codes for the structures selected are listed in Table 1.

Software. The CSD programs QUEST and GSTAT[3] were initially used to select structures and extract information from the CSD. The program QUEST was used to generate files containing a range of

* To whom correspondence should be addressed.

⊗ Abstract published in *Advance ACS Abstracts,* January 1, 1996.

(1) Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S. H.; Srinivasan, A. R.; Schneider, B. *Biophys. J.* **1992**, *63*, 751−759.

(2) Taylor, R.; Kennard, O. *J. Mol. Struct.* **1982**, *78*, 1−28.

(3) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. *Acta. Crystallogr.* **1979**, *B35*, 2331−2339.

0002-7863/96/1518-0509$12.00/0

Table 1. CSD Codes of Base Structures Used[a]

cytosine	docytc	kuthow	jucpay	araden10	thopad10
acytid	jibdih	kuzxei	katyot	bedlif	vomfos
bivvil	kogbox	methym01	kofguh	betwus	yabgaj
bofwoi	kogbox01	palwuz	lyfura	bifyoe	
botsim10	mecyto10	soddeu	metura01	boyduo	**adenine-pro**
boxgie	tazzup	sosbil	metura02	buvpox	admopm
budway10	vuymad	tmthym	metura03	cezbis	admpot10
cimjen	xfurcc10	tpataa	metura04	cezmez	adoshc
cytcyp20		trfutm	mxeurd	cidgur	adposd
cytidi10	**thymine**	tymcxa	pabkos	coczid	arfuad
decyuf	bicrua	vevbut	suridp	cubrum	arfuad01
fikhow	cedboc	vevbut01	suromm	dhoads01	cugnex
fovxet	cezfoc	vikfea	tanzen	dhoads10	dhpmad
gahhom	clqunb10		thfurc10	dohgem	edpbxu
gigmoy	dayvoo	**uracil**	thpyur	dorjoj	kitrek
gouger10	doxziz	aurcpb	uraraf01	duvheh10	madcmp
jikhoa	duksor	beurid10	uraraf10	eadpba	slcada10
juhlaz	duxxez	biyrik	uridmp10	fabfuj01	tahjiv
kosnov	firpol	bofwic	vanjez	fibgec	
marafc	fixgau01	bufyik	vevhuz	fikhai	**guanine**
metcyt01	fixgau02	cawcae	vomfim	fikhai01	budway10
segmas	fixgau03	cdurid	vomgaf	foylua	cehtak10
sivzus	gebtem	cedbiw	vukgox	gahhig	dahmii
	gexxiq	cekluz	zzzapa10	gahhig01	gebrio
cytosine-pro	gexxiq01	ciryux		kemyeg	gipbiq
aracyp	kasvef	combov	**adenine**	keplia	guansh10
arfcyt10	katyin	daurid01	acados	keplog	guopna10
bzcytn	kexkih	fecfau	adenos01	kubguj	guopna11
cytiac	kezruc	fipkeu	adenos10	meaden01	jafhih
cyticl	kinheu	funten	adprop	naamph10	scgmpt10
cytidn	kinhiy	gathoy	amdoad	opadna	sdgunp
docypo	kitsov	gidzic10	amoada	opadna01	tamxek
docypo03	kogbir	jikbua	amoadb	pakwon	vuwrek
	kunnem	jucnok	amoadc		

[a] The references for these structures are available on the WWW (http://ndbserver.rutgers.edu) in the Archive Section.

Table 2. Cytosine Statistics: Parameter Estimates for Neutral Cytosine ($N = 28$) Compared with Those from Ref 2 ($N = 14$)[m]

parameter	x_{med}[a]	$\bar{x}$[b]	$\bar{x}_{T\&K}$[c]	P_t[d]	σ[e]	$\sigma_{T\&K}$[f]	min(x)[g]	min(x)$_{T\&K}$[h]	max(x)[i]	max(x)$_{T\&K}$[j]	V[k]	P_V[l]
N1−C2	1.397	1.397(2)	1.399(4)	0.624	0.010	0.014	1.379	1.379	1.416	1.416	0.143	−
C2−N3	1.355	1.353(1)	1.356(3)	0.347	0.008	0.012	1.334	1.334	1.363	1.384	0.295	0.025
N3−C4	1.335	1.335(1)	1.334(2)	0.543	0.007	0.006	1.325	1.326	1.359	1.346	0.266	0.100
C4−C5	1.424	1.425(1)	1.426(4)	0.743	0.008	0.015	1.412	1.391	1.447	1.447	0.192	−
C5−C6	1.340	1.339(1)	1.337(2)	0.291	0.008	0.006	1.321	1.327	1.351	1.351	0.156	−
C6−N1	1.365	1.367(1)	1.364(2)	0.242	0.006	0.007	1.357	1.356	1.380	1.376	0.259	0.100
C2−O2	1.242	1.240(2)	1.237(2)	0.223	0.009	0.006	1.225	1.226	1.254	1.247	0.209	−
C4−N4	1.334	1.335(2)	1.337(4)	0.605	0.009	0.015	1.318	1.312	1.358	1.369	0.242	0.150
N1−C1′	1.470	1.470(2)	na	na	0.012	na	1.450	na	1.497	na	0.159	−
C6−N1−C2	120.3	120.3(1)	120.6(1)	0.010	0.4	0.3	119.5	120.0	121.0	121.0	0.173	−
N1−C2−N3	119.1	119.2(1)	118.9(2)	0.148	0.7	0.6	117.8	117.8	120.5	119.9	0.145	−
C2−N3−C4	120.0	119.9(1)	120.0(2)	0.217	0.5	0.7	118.9	118.0	120.8	120.7	0.179	−
N3−C4−C5	121.9	121.9(1)	121.8(2)	0.756	0.4	0.6	121.0	121.0	122.6	123.2	0.143	−
C4−C5−C6	117.4	117.4(1)	117.6(2)	0.383	0.5	0.6	116.2	116.2	118.3	118.4	0.247	0.150
C5−C6−N1	121.0	121.0(1)	121.0(2)	0.921	0.5	0.7	120.0	119.9	122.0	122.0	0.152	−
N1−C2−O2	118.9	118.9(1)	119.2(2)	0.243	0.6	0.8	117.6	117.8	119.9	121.3	0.168	−
N3−C2−O2	121.9	121.9(1)	121.9(2)	0.948	0.7	0.9	120.1	119.5	123.0	123.0	0.170	−
N3−C4−N4	118.1	118.0(1)	117.9(3)	0.648	0.7	1.1	116.4	115.0	119.4	119.2	0.215	−
C5−C4−N4	120.2	120.2(1)	120.3(2)	0.660	0.7	0.8	118.9	119.1	121.8	121.8	0.150	−
C6−N1−C1′	121.0	120.8(2)	na	na	1.2	na	118.7	na	122.7	na	0.305	0.025
C2−N1−C1′	118.5	118.8(2)	na	na	1.1	na	116.6	na	120.7	na	0.201	−

[a] Median value in the set of parameter values. Values for bond lengths are in angstroms and for angles in degrees. [b] Arithmetic mean value with standard error of the mean in parentheses. [c] Arithmetic mean value from ref 2 with the standard error of the mean in parentheses. [d] Significance level for the equivalence of the means. [e] Standard deviation of the sample. [f] Standard deviation of the sample from ref 2. [g] Minimum value in the sample. [h] Minimum value from ref 2. [i] Maximum value in the sample. [j] Maximum value from ref 2. [k] Value of the Kuiper statistic for comparing the current sample against a normal distribution. [l] Significance level for the Kuiper statistic when the mean and standard deviation are estimated from the sample. The probability is according to ref 6. For example, a value of 0.1000 implies that the likelihood of a sample having this large a statistic being normal is only 10 in 100. Where "−" is used, this probability is greater than 0.150, while 0.01 is the lowest possible confidence level. [m] Tables 2−8 share a common format. The first column contains the names of all bonds and angles in a given base, while the data in the other columns are described by the footnotes. The columns with data from ref 2 are labeled T&K.

information about the selected structures (*dat*-format and *bib*-format files), while GSTAT was used to generate files of bond lengths and valence angles (*table*-format files). These files were then used as input with a new program, NDB-dict, which converts *table*-format files to lists of geometries (values of bond lengths and valence angles), *dat*-format files to lists of experimental information (such as *R* factors and space groups), and *bib*-format files to bibliographic lists. The object-oriented design of NDB-dict makes it relatively simple to generate sets

Table 3. Cytosine-pro Statistics: Parameter Estimates for Protonated Cytosine ($N = 17$) Compared with Those from Ref 2 ($N = 17$)

parameter	x_{med}[a]	$\bar{x}$[b]	$\bar{x}_{T\&K}$[c]	P_t[d]	σ[e]	$\sigma_{T\&K}$[f]	min(x)[g]	min(x)$_{T\&K}$[h]	max(x)[i]	max(x)$_{T\&K}$[j]	V[k]	P_V[l]
N1−C2	1.380	1.381(2)	1.381(2)	0.965	0.007	0.008	1.371	1.365	1.401	1.401	0.262	−
C2−N3	1.383	1.384(2)	1.387(2)	0.159	0.007	0.007	1.370	1.376	1.396	1.403	0.200	−
N3−C4	1.352	1.353(2)	1.352(1)	0.763	0.006	0.006	1.339	1.339	1.364	1.363	0.245	−
C4−C5	1.414	1.413(1)	1.413(3)	0.984	0.005	0.011	1.403	1.396	1.422	1.445	0.221	−
C5−C6	1.347	1.346(2)	1.341(2)	0.072	0.006	0.010	1.330	1.314	1.357	1.357	0.249	−
C6−N1	1.365	1.365(2)	1.362(2)	0.264	0.007	0.010	1.350	1.339	1.377	1.380	0.203	−
C2−O2	1.213	1.212(1)	1.211(2)	0.594	0.006	0.007	1.201	1.201	1.221	1.227	0.205	−
C4−N4	1.313	1.315(2)	1.313(3)	0.454	0.007	0.011	1.308	1.279	1.336	1.329	0.376	0.025
N1−C1′	1.475	1.483(4)	na	na	0.015	na	1.469	na	1.510	na	0.465	0.010
C6−N1−C2	121.9	121.7(1)	121.5(1)	0.290	0.5	0.5	120.7	120.7	122.3	122.3	0.337	0.100
N1−C2−N3	114.8	114.7(2)	114.9(2)	0.438	0.7	0.7	113.4	113.4	116.2	116.0	0.318	0.100
C2−N3−C4	125.5	125.3(2)	125.1(2)	0.360	0.7	0.7	123.7	123.7	126.1	126.1	0.382	0.025
N3−C4−C5	117.6	117.6(1)	117.5(2)	0.777	0.5	0.7	116.7	116.4	118.5	118.6	0.269	−
C4−C5−C6	118.4	118.4(1)	118.5(1)	0.504	0.5	0.6	117.2	117.2	119.3	119.7	0.285	−
C5−C6−N1	122.1	122.2(1)	122.5(1)	0.168	0.5	0.5	121.6	121.6	123.8	123.8	0.342	0.050
N1−C2−O2	123.5	123.4(2)	123.5(2)	0.721	0.7	0.6	122.1	122.4	124.5	124.7	0.244	−
N3−C2−O2	121.8	121.9(1)	121.6(1)	0.170	0.5	0.6	121.2	120.2	122.9	122.5	0.218	−
N3−C4−N4	119.7	119.5(2)	119.5(2)	0.886	0.7	0.7	118.3	118.3	120.3	120.6	0.336	0.100
C5−C4−N4	122.9	123.0(2)	123.0(3)	0.929	0.8	1.0	121.8	120.9	124.2	124.8	0.196	−
C6−N1−C1′	121.3	121.2(2)	na	na	0.9	na	119.9	na	122.6	na	0.216	−
C2−N1−C1′	116.7	116.9(2)	na	na	1.0	na	115.4	na	118.7	na	0.239	−

Table 4. Thymine Statistics: Parameter Estimates for Thymine ($N = 50$) Compared with Those for Uracil from Ref 2 ($N = 32$)

parameter	x_{med}[a]	$\bar{x}$[b]	$\bar{x}_{T\&K}$[c]	P_t[d]	σ[e]	$\sigma_{T\&K}$[f]	min(x)[g]	min(x)$_{T\&K}$[h]	max(x)[i]	max(x)$_{T\&K}$[j]	V[k]	P_V[l]
N1−C2	1.376	1.376(1)	1.379(2)	0.131	0.008	0.010	1.358	1.357	1.393	1.397	0.165	−
C2−N3	1.373	1.373(1)	1.373(2)	0.926	0.008	0.009	1.356	1.358	1.394	1.401	0.146	−
N3−C4	1.381	1.382(1)	1.383(2)	0.645	0.008	0.010	1.366	1.363	1.401	1.410	0.154	−
C4−C5	1.446	1.445(1)	1.440(2)	0.044	0.009	0.011	1.419	1.418	1.464	1.458	0.269	0.010
C5−C6	1.339	1.339(1)	1.338(2)	0.756	0.007	0.009	1.324	1.320	1.355	1.356	0.205	0.100
C6−N1	1.379	1.378(1)	1.380(2)	0.436	0.007	0.011	1.361	1.354	1.395	1.403	0.129	−
C2−O2	1.218	1.220(1)	1.218(2)	0.464	0.008	0.010	1.202	1.190	1.243	1.239	0.164	−
C4−O4	1.228	1.228(1)	1.227(2)	0.534	0.009	0.009	1.207	1.200	1.246	1.243	0.119	−
C5−M5	1.497	1.496(1)	na	na	0.006	na	1.484	na	1.510	na	0.228	0.025
N1−C1′	1.470	1.473(2)	na	na	0.014	na	1.441	na	1.506	na	0.220	0.025
C6−N1−C2	121.2	121.3(1)	121.3(1)	0.701	0.5	0.6	119.9	120.0	122.4	122.8	0.219	0.025
N1−C2−N3	114.4	114.6(1)	114.8(1)	0.152	0.6	0.7	113.4	113.6	116.2	116.0	0.244	0.010
C2−N3−C4	127.1	127.2(1)	127.0(1)	0.224	0.6	0.6	126.3	125.6	129.0	128.4	0.205	0.100
N3−C4−C5	115.3	115.2(1)	114.7(2)	0.012	0.6	0.9	113.8	113.3	116.4	116.7	0.170	−
C4−C5−C6	118.0	118.0(1)	119.2(2)	0.000	0.6	1.3	116.3	117.2	119.2	122.3	0.182	−
C5−C6−N1	123.6	123.7(1)	122.8(2)	0.000	0.6	0.9	122.3	120.1	125.7	124.2	0.209	0.050
N1−C2−O2	123.1	123.1(1)	123.2(1)	0.460	0.8	0.8	121.5	120.6	124.9	124.6	0.140	−
N3−C2−O2	122.3	122.3(1)	122.0(1)	0.025	0.6	0.7	120.8	120.7	123.5	123.6	0.178	−
N3−C4−O4	119.9	119.9(1)	119.8(1)	0.593	0.6	0.7	118.4	118.4	120.9	121.4	0.172	−
C5−C4−O4	124.9	124.9(1)	125.4(2)	0.028	0.7	1.0	123.2	123.2	126.4	127.2	0.126	−
C4−C5−M5	119.0	119.0(1)	na	na	0.6	na	117.7	na	120.6	na	0.173	−
C6−C5−M5	122.9	122.9(1)	na	na	0.6	na	121.3	na	124.8	na	0.177	−
C6−N1−C1′	120.2	120.4(2)	na	na	1.5	na	115.4	na	124.1	na	0.160	−
C2−N1−C1′	118.3	118.2(2)	na	na	1.6	na	114.4	na	123.6	na	0.198	0.100

and subsets of structures which can be examined or compared using a number of different statistical tests. NDB-dict was also used to generate commands for a second new program, plot2d, which was used to generate the histograms shown in this paper, and for LaTeX,[4] the text formatting language which was used for report generation.

Statistics. For statistical purposes, samples consist of the bond lengths and valence angles of all structures from the CSD that meet the criteria previously defined. For example, structures containing neutral cytosine are associated with a number of samples of bond lengths and valence angles, such as N1−C2 bond lengths, C2−N3 bond lengths, and N1−C2−N3 valence angles. A complementary set of samples was also formed for the subset of these structures that are derivatized with furanose sugars having obligatory oxygens at the 3′ and 5′ positions. Two further complementary sets of samples are formed for structures containing protonated cytosine. Thus, for each base considered, there is a distinct sample of values for each bond and angle. These samples were characterized and compared using several statistical parameters and tests. The entries in a typical sample, such as the set of N1−C2 bond lengths of neutral cytosine, were sorted to determine the minimum (min(x)), maximum (max(x)), and median (med(x)) values. The mean ($\bar{x}$) and standard deviation (σ) were calculated with standard equations; the standard error of the mean (sem) is $\sigma/\sqrt{N}$, where N is the number of values in a given sample (i.e., number of structures available). Note that σ is used here for both the sample and population standard deviations. For example, the population of N1−C2 bond lengths is the set of all possible measurements of the N1−C2 bond length, while a sample is a subset of this. Thus, the mean of the sample approaches that of the population in the limit of infinite sample size and the parameters presented here are only estimates of the true population values.

A series of statistical tests was used to compare the means, variances, and distributions of two different samples. The hypothesis is made that some property is shared by two samples (the null hypothesis), and the probability that the null hypothesis holds is then determined (the significance level for the test). For example, are the N1−C2 bond lengths in neutral and protonated cytosine significantly different?

The normality of sample distributions was examined with the Kuiper test,[5,6] in which the distribution of a sample was compared with the expected normal distribution. The null hypothesis is that the two

(4) Lamport, L. *Latex−A Document Preparation System−Users Guide and Reference Manual*; Addison-Wesley: Reading, MA, 1985.

(5) Stephens, M. A. *Biometrika* **1965**, *52*, 309−321.

distributions are equivalent, and a low significance level implies the presence of systematic errors, such as those associated with significant environmental effects on the geometry of the crystal structure. The Kuiper test is similar to the more common chi-squared (χ^2) test but has the advantages of not requiring the samples to be binned and of often being more sensitive to differences in distributions. The Kuiper statistic V, given below in eq 1, is a member of a class of statistical equations known as empirical distribution (EDF) statistics. The general form of an EDF is

$$f(z_i) = i/N$$

where i is the ordinal number of the sorted z and N is the total number of values in the sample, so that i/N is simply the fraction of elements with a value less than that of z_i, and the function takes on values from 0 to 1. The normal distribution function, that correspondingly varies from 0 to 1, is

$$P(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-x^2/2}\,dx$$

The Kuiper test is a measure of the difference between these two functions for given data.

To find V, a collection of data (bond lengths or valence angles) is normalized to the corresponding z values using eq 2, where σ is the standard deviation of the sample and $\bar{x}$ the mean. The z values are then sorted in ascending order and used in eq 3 to evaluate the step functions EDF^+ and EDF^-. In eq 1, max denotes the maximal differences between the EDF functions and $P(z)$ for all data considered, with a total sample size of N. Note that i in eqs 1 and 3 is the ordinal value of a given z, i.e., i is ord(z_i), following sorting. $P(z)$ is the corresponding Gaussian function for the distribution of the normal population; the ideal Gaussian distribution is expressed as a cumulative function so that $P(z_i)$ is the integral of the normal curve from $-\infty$ to z_i, a value which can be either calculated or obtained from tables found in most statistics books. The quantity V is thus a measure of the sum of maximal differences between the empirical and expected distribution functions when $P(z_i)$ is (a) greater than and (b) less than the value of the EDF.

$$V = \max_{0<i\le N}[\mathrm{EDF}_i^{+} - P(z_i)] + \max_{0<i\le N}[P(z_i) - \mathrm{EDF}_i^{-}] \qquad (1)$$

$$z = \frac{x - \bar{x}}{\sigma} \qquad (2)$$

$$\mathrm{EDF}_i^{+} = \frac{i}{N} \qquad (3a)$$

$$\mathrm{EDF}_i^{-} = \frac{i-1}{N} \qquad (3b)$$

The significance level for the null hypothesis (P_V) is the probability that V would be as large as the value observed if the sample distribution were indeed normal; it may be found analytically or by using Monte Carlo simulations. The former method requires that the population mean and standard deviation must be known independently and not be estimated from the data. For the other case, where the values are estimated from the data, Monte Carlo simulations can be performed to find the probabilities associated with different values of V and different sample sizes, as was done in Table 4.9 of ref 6. These significance levels are reported here for the observed values of V.

The t test proposed by Behrens[7,8] (as opposed to the more commonly employed Student's t test) was used to compare the means of two samples. The null hypothesis for this test is that the two means are the same, and the significance level for the test (P_t) is the probability that the hypothesis is true. For example, a P_t value of 0.05 corresponds to a 5% probability that the two means are equivalent for normally distributed data. The two-tailed F test[8] was used to determine whether the variances of two samples were significantly different. This test was applied to subsets of structures selected with different R factors used as cutoffs (results not shown).

Regression Analysis. The coordinates for "average" bases have been obtained by finding the set of coordinates which minimizes the function f in eq 4 for one of the bases. The summations are over all m bond lengths and n angles of a particular base, where $\bar{d}_i$ is the average distance observed for bond i in the appropriate sample, while $\bar{\theta}_i$ is the average value of valence angle i. For example, consider the sample of N1−C2 bond lengths found in neutral cytosine residues and the average for this sample, $\bar{d}_1$. The corresponding value of $\bar{d}_1$ is the N1−C2 bond length calculated for a given set of coordinates, while σ_1 is the sample standard deviation which is appropriate for $\bar{d}_1$ or $\bar{\theta}_1$. To find the set of coordinates minimizing the difference between these average and calculated values, the downhill simplex algorithm of Nelder and Mead[8] has been incorporated into NDB-dict for function minimization:

$$f = \left(\sum_{i=1}^{n}\frac{(\theta_i - \bar{\theta}_i)^2}{\sigma_i^2} + \sum_{i=1}^{m}\frac{(d_i - \bar{d}_i)^2}{\sigma_i^2}\right) \qquad (4)$$

Results

The statistical analyses of bond lengths and valence angles for seven sets of neutral and protonated structures are reported in Tables 2−8. The frequency distributions for non-normal samples, as well as the distributions for the sugar-derivatized subsets, are displayed in Figures 2 and 3; distributions of pyrimidines are in Figure 2 and purines in Figure 3.

Cytosine. The geometrical parameters obtained for cytosine are listed in Table 2, and the non-normal frequency distributions are shown in Figure 2. At the 5% level of significance, the Kuiper test shows all samples of bond lengths and valence angles to be normally distributed except for the C2−N3 bond distances and C6−N1−C1′ angles. These same bonds and angles are distributed non-normally in the sugar-derivatized samples (data not shown). Comparison of the current results with those of Taylor and Kennard[2] shows that only the mean C6−N1−C2 angle differs at the 5% significance level according to the t test. The frequency distributions for the complete set of cytosine structures and the sugar-derivatized subset appear similar, and at the 5% level of significance, the means of these samples are indistinguishable according to the t test.

Protonated Cytosine. The geometrical parameters obtained for protonated cytosine are reported in Table 3, and the frequency distributions for the non-normal sample geometries in Figure 2. At the 5% level of significance, the Kuiper test shows the samples to be normal except for two bonds, C4−N4 and N1−C1′, and two angles, C2−N3−C4 and C5−C6−N1. In the sugar-derivatized samples the same parameters are distributed non-normally with the exception of the C5−C6−N1 angle (data not shown). There are no differences, at the 5% significance level, with the corresponding mean values reported in Taylor and Kennard.[2]

Thymine. The results for thymine represent a special case since this base was not included separately in the previous survey, but instead was treated as a substituted uracil. Thus, the geometrical parameters obtained for thymine in Table 4 have been compared with those obtained previously[2] for uracil. The significant differences between the samples are to be expected. The t test shows the sugar-derivatized subset (Table 9) also to differ substantially from the full set of thymines for a number of average parameters involving atom N1. At the 5% level of significance, the Kuiper test shows that a number of the bond distances and angles involving N1 or C5 are distributed non-

(6) Stephens, M. A. In *Goodness of Fit Techniques*; *Statistics: Textbooks and Monographs*; D'Agostino, R., Stephens, M., Eds.; Marcel Dekker, Inc.: New York, 1986; Vol. 68, pp 97−193.

(7) Hamilton, W. *Statistics in Physical Science*; Ronald Press: New York, 1964; pp 92 and 93.

(8) Press, W. H.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: New York, 1992; pp 408−412 and 616−619.

Table 5. Uracil Statistics: Parameter Estimates for Uracil ($N = 46$) Compared with Those from Ref 2 ($N = 32$)

parameter	x_{med}[a]	$\bar{x}$[b]	$\bar{x}_{T\&K}$[c]	P_t[d]	σ[e]	$\sigma_{T\&K}$[f]	min(x)[g]	min(x)$_{T\&K}$[h]	max(x)[i]	max(x)$_{T\&K}$[j]	V[k]	P_V[l]
N1−C2	1.381	1.381(1)	1.379(2)	0.435	0.009	0.010	1.363	1.357	1.399	1.397	0.162	−
C2−N3	1.374	1.373(1)	1.373(2)	0.875	0.007	0.009	1.356	1.358	1.388	1.401	0.151	−
N3−C4	1.382	1.380(1)	1.383(2)	0.235	0.009	0.010	1.362	1.363	1.402	1.410	0.180	−
C4−C5	1.430	1.431(1)	1.440(2)	0.000	0.009	0.011	1.407	1.418	1.452	1.458	0.221	0.050
C5−C6	1.338	1.337(1)	1.338(2)	0.764	0.009	0.009	1.316	1.320	1.357	1.356	0.134	−
C6−N1	1.375	1.375(1)	1.380(2)	0.057	0.009	0.011	1.358	1.354	1.391	1.403	0.180	−
C2−O2	1.219	1.219(1)	1.218(2)	0.813	0.009	0.010	1.190	1.190	1.241	1.239	0.168	−
C4−O4	1.232	1.232(1)	1.227(2)	0.024	0.008	0.009	1.217	1.200	1.249	1.243	0.129	−
N1−C1′	1.465	1.469(2)	na	na	0.014	na	1.448	na	1.503	na	0.254	0.010
C6−N1−C2	121.2	121.0(1)	121.3(1)	0.044	0.6	0.6	119.8	120.0	122.1	122.8	0.192	0.150
N1−C2−N3	114.9	114.9(1)	114.8(1)	0.498	0.6	0.7	113.5	113.6	116.2	116.0	0.174	−
C2−N3−C4	126.9	127.0(1)	127.0(1)	0.750	0.6	0.6	125.8	125.6	128.4	128.4	0.140	−
N3−C4−C5	114.6	114.6(1)	114.7(2)	0.671	0.6	0.9	113.4	113.3	115.9	116.7	0.169	−
C4−C5−C6	119.6	119.7(1)	119.2(2)	0.072	0.6	1.3	118.5	117.2	121.0	122.3	0.182	−
C5−C6−N1	122.6	122.7(1)	122.8(2)	0.487	0.5	0.9	121.4	120.1	123.7	124.2	0.214	0.100
N1−C2−O2	122.8	122.8(1)	123.2(1)	0.046	0.7	0.8	120.6	120.6	124.5	124.6	0.152	−
N3−C2−O2	122.3	122.2(1)	122.0(1)	0.127	0.7	0.7	120.5	120.7	123.7	123.6	0.278	0.010
N3−C4−O4	119.6	119.4(1)	119.8(1)	0.029	0.7	0.7	117.8	118.4	120.4	121.4	0.258	0.010
C5−C4−O4	125.9	125.9(1)	125.4(2)	0.010	0.6	1.0	124.7	123.2	127.2	127.2	0.181	−
C6−N1−C1′	121.2	121.2(2)	na	na	1.4	na	118.1	na	123.7	na	0.148	−
C2−N1−C1′	117.7	117.7(2)	na	na	1.200	na	114.9	na	119.5	na	0.179	−

Table 6. Adenine Statistics: Parameter Estimates for Neutral Adenine ($N = 48$) Compared with Those from Ref 2 ($N = 21$)

parameter	x_{med}[a]	$\bar{x}$[b]	$\bar{x}_{T\&K}$[c]	P_t[d]	σ[e]	$\sigma_{T\&K}$[f]	min(x)[g]	min(x)$_{T\&K}$[h]	max(x)[i]	max(x)$_{T\&K}$[j]	V[k]	P_V[l]
N1−C2	1.340	1.339(1)	1.338(3)	0.701	0.009	0.012	1.323	1.316	1.366	1.367	0.179	−
C2−N3	1.332	1.331(1)	1.332(3)	0.700	0.009	0.014	1.297	1.308	1.345	1.369	0.217	0.050
N3−C4	1.344	1.344(1)	1.342(2)	0.323	0.006	0.009	1.329	1.321	1.361	1.357	0.214	0.050
C4−C5	1.385	1.383(1)	1.382(2)	0.582	0.007	0.010	1.367	1.363	1.396	1.408	0.132	−
C5−C6	1.406	1.406(1)	1.409(1)	0.044	0.009	0.005	1.378	1.398	1.427	1.417	0.184	−
C6−N1	1.351	1.351(1)	1.349(2)	0.398	0.007	0.011	1.339	1.333	1.367	1.371	0.131	−
C5−N7	1.387	1.388(1)	1.385(2)	0.153	0.006	0.010	1.376	1.367	1.406	1.406	0.203	0.100
N7−C8	1.311	1.311(1)	1.312(2)	0.576	0.007	0.007	1.296	1.292	1.329	1.328	0.186	0.150
C8−N9	1.372	1.373(1)	1.367(4)	0.128	0.008	0.016	1.356	1.328	1.390	1.406	0.189	0.150
N9−C4	1.374	1.374(1)	1.376(2)	0.254	0.006	0.009	1.362	1.361	1.386	1.401	0.131	−
C6−N6	1.336	1.335(1)	1.337(3)	0.614	0.008	0.015	1.314	1.321	1.352	1.392	0.180	−
N9−C1′	1.464	1.462(1)	na	na	0.010	na	1.437	na	1.483	na	0.188	0.150
C6−N1−C2	118.5	118.6(1)	118.8(2)	0.230	0.6	0.8	116.6	117.7	120.1	120.8	0.192	0.150
N1−C2−N3	129.2	129.3(1)	129.0(1)	0.057	0.5	0.6	128.3	127.7	130.3	130.5	0.155	−
C2−N3−C4	110.6	110.6(1)	110.8(1)	0.322	0.5	0.6	109.3	109.9	111.6	112.1	0.179	−
N3−C4−C5	126.7	126.8(1)	126.9(2)	0.505	0.7	0.8	125.0	125.2	128.3	128.1	0.117	−
C4−C5−C6	117.0	117.0(1)	116.9(1)	0.373	0.5	0.6	116.1	115.4	118.3	118.4	0.137	−
C5−C6−N1	117.6	117.7(1)	117.6(1)	0.685	0.5	0.6	116.4	116.4	118.7	118.5	0.172	−
C4−C5−N7	110.6	110.7(1)	110.7(1)	0.765	0.5	0.5	109.6	109.8	111.9	111.5	0.267	0.010
C5−N7−C8	103.9	103.9(1)	103.9(2)	0.808	0.5	0.7	101.9	102.4	104.9	105.8	0.185	−
N7−C8−N9	113.8	113.8(1)	113.8(2)	0.990	0.5	0.7	112.9	112.4	115.1	115.7	0.163	−
C8−N9−C4	105.8	105.8(1)	105.9(1)	0.535	0.4	0.5	104.4	104.7	106.8	106.9	0.274	0.010
N9−C4−C5	105.8	105.8(1)	105.7(1)	0.311	0.4	0.6	105.0	104.5	106.6	106.5	0.184	−
N3−C4−N9	127.3	127.4(1)	127.4(1)	0.919	0.8	0.6	125.9	126.2	129.4	128.4	0.211	0.050
C6−C5−N7	132.5	132.3(1)	132.3(2)	0.899	0.7	0.7	130.1	130.2	133.3	133.9	0.287	0.010
N1−C6−N6	118.5	118.6(1)	119.0(2)	0.060	0.6	0.8	117.5	117.7	120.1	120.4	0.209	0.100
C5−C6−N6	123.6	123.7(1)	123.4(2)	0.213	0.8	1.0	122.1	121.7	125.8	125.2	0.147	−
C8−N9−C1′	127.6	127.7(3)	na	na	1.8	na	122.5	na	130.7	na	0.162	−
C4−N9−C1′	125.9	126.3(3)	na	na	1.8	na	123.7	na	130.6	na	0.210	0.050

normally. Restricting the samples to include only sugar-derivatized thymine fragments significantly increases normality in corresponding samples, as is apparent in Figure 2 and Table 9.

Uracil. The geometric parameters obtained for uracil are reported in Table 5, and the frequency distributions for the non-normal sample geometries in Figure 2. The samples are distributed normally except for bonds C4−C5 and N1−C1′ and angles N3−C2−O2 and N3−C4−O4. In the sugar-derivatized samples only the N3−C4−C5 angles are distributed non-normally (data not shown). The mean value for the C6−N1 bond of the sugar-derivatized subset also differs at the 5% significance level from that of the parent set. The means of several bonds and angles differ at the 5% significance level from the corresponding values found by Taylor and Kennard.[2]

The effect of methylation on the geometry of uracil was examined by comparing corresponding parameters for uracil and its C5-methyl derivative, thymine. The means were compared using the *t* test, with the results shown in Table 10. There are highly significant differences in the parameters of the bond lengths and valence angles containing the C5 atom, except for the C5−C6 bond.

Adenine. The geometrical parameters obtained for adenine are reported in Table 6, and the non-normal frequency distributions for the sample geometries in Figure 3. At the 5% level of significance, the Kuiper test shows the distributions of many bond distances and angles—C2−N3, N3−C4, C4−C5−N7, C8−N9−C4, N3−C4−N9, C6−C5−N7, and C4−N9−C1′—to be non-normal at the 5% level. In contrast, the only samples distributed non-normally for the sugar-derivatized set are those for bonds N3−C4, C5−C6, and C5−N7 and angles C4−C5−N7 and N9−C4−C5 (data not shown). Only the mean value

Table 7. Adenine-pro Statistics: Parameter Estimates for Protonated Adenine ($N = 15$) Compared with Those from Ref 2 ($N = 13$)

parameter	x_{med}[a]	$\bar{x}$[b]	$\bar{x}_{T\&K}$[c]	P_t[d]	σ[e]	$\sigma_{T\&K}$[f]	min(x)[g]	min(x)$_{T\&K}$[h]	max(x)[i]	max(x)$_{T\&K}$[j]	V[k]	P_V[l]
N1−C2	1.356	1.357(2)	1.362(4)	0.269	0.009	0.013	1.348	1.349	1.387	1.387	0.465	0.010
C2−N3	1.304	1.305(2)	1.306(2)	0.690	0.008	0.008	1.292	1.292	1.323	1.323	0.330	0.150
N3−C4	1.354	1.356(1)	1.354(3)	0.414	0.006	0.009	1.349	1.336	1.366	1.366	0.338	0.100
C4−C5	1.380	1.378(2)	1.385(6)	0.264	0.008	0.020	1.354	1.354	1.386	1.442	0.357	0.100
C5−C6	1.402	1.403(2)	1.405(4)	0.696	0.007	0.015	1.391	1.389	1.412	1.448	0.293	−
C6−N1	1.359	1.359(2)	1.360(2)	0.607	0.007	0.008	1.350	1.349	1.371	1.371	0.230	−
C5−N7	1.379	1.379(1)	1.378(2)	0.586	0.005	0.007	1.370	1.364	1.387	1.389	0.238	−
N7−C8	1.312	1.312(2)	1.316(2)	0.212	0.008	0.008	1.301	1.305	1.330	1.330	0.232	−
C8−N9	1.375	1.373(2)	1.378(4)	0.196	0.009	0.012	1.357	1.357	1.384	1.398	0.272	−
N9−C4	1.365	1.365(2)	1.366(3)	0.816	0.007	0.009	1.353	1.350	1.376	1.378	0.187	−
C6−N6	1.321	1.320(2)	1.322(3)	0.640	0.008	0.012	1.304	1.312	1.332	1.355	0.199	−
N9−C1′	1.466	1.466(2)	na	na	0.009	na	1.45	na	1.479	na	0.220	−
C6−N1−C2	123.3	123.3(2)	123.2(2)	0.821	0.6	0.6	122.1	121.9	124.3	124.1	0.159	−
N1−C2−N3	125.8	125.7(1)	125.5(2)	0.376	0.6	0.6	124.9	124.5	126.9	126.9	0.330	0.150
C2−N3−C4	111.6	111.6(1)	112.0(2)	0.179	0.4	0.9	111.0	111.0	112.4	114.6	0.377	0.050
N3−C4−C5	127.4	127.4(1)	127.4(2)	0.894	0.6	0.7	126.2	125.7	128.7	128.7	0.338	0.100
C4−C5−C6	118.1	117.9(1)	117.7(2)	0.440	0.5	0.9	116.6	115.6	118.4	118.4	0.379	0.050
C5−C6−N1	113.9	114.0(1)	114.3(3)	0.322	0.4	1.0	113.4	113.4	115.1	117.2	0.307	−
C4−C5−N7	111.0	111.0(1)	111.0(2)	0.936	0.3	0.5	110.5	110.4	111.6	112.3	0.248	−
C5−N7−C8	103.7	103.7(1)	104.1(2)	0.074	0.4	0.6	103.3	103.3	104.8	105.2	0.312	−
N7−C8−N9	113.6	113.5(1)	113.0(2)	0.098	0.6	0.9	111.9	111.8	114.1	114.1	0.316	−
C8−N9−C4	106.0	105.9(1)	106.3(2)	0.085	0.4	0.7	104.9	105.6	106.6	108.3	0.305	−
N9−C4−C5	105.8	105.8(1)	105.6(2)	0.364	0.5	0.8	105.1	104.1	107.0	107.0	0.282	−
N3−C4−N9	126.7	126.7(2)	127.0(4)	0.487	0.8	1.2	124.3	124.3	127.7	130.1	0.423	0.010
C6−C5−N7	130.9	131.0(1)	131.3(2)	0.308	0.5	0.8	130.5	130.5	132.0	133.5	0.283	−
N1−C6−N6	120.1	120.2(2)	120.2(2)	0.968	0.7	0.8	118.7	118.7	121.6	121.6	0.385	0.025
C5−C6−N6	126.1	125.8(2)	125.5(3)	0.430	0.8	1.2	124.1	123.5	127.3	127.3	0.339	0.100
C8−N9−C1′	127.5	127.2(4)	na	na	1.6	na	124.5	na	130.7	na	0.282	−
C4−N9−C1′	126.5	126.8(5)	na	na	1.8	na	123.2	na	129.9	na	0.258	−

Table 8. Guanine Statistics: Parameter Estimates for Guanine ($N = 21$) Compared with Those from Ref 2 ($N = 7$)

parameter	x_{med}[a]	$\bar{x}$[b]	$\bar{x}_{T\&K}$[c]	P_t[d]	σ[e]	$\sigma_{T\&K}$[f]	min(x)[g]	min(x)$_{T\&K}$[h]	max(x)[i]	max(x)$_{T\&K}$[j]	V[k]	P_V[l]
N1−C2	1.371	1.373(2)	1.375(3)	0.670	0.008	0.008	1.362	1.365	1.389	1.387	0.310	0.050
C2−N3	1.324	1.323(2)	1.327(2)	0.199	0.008	0.006	1.301	1.320	1.335	1.335	0.195	−
N3−C4	1.351	1.350(2)	1.355(2)	0.061	0.007	0.005	1.338	1.345	1.368	1.362	0.231	−
C4−C5	1.378	1.379(2)	1.377(2)	0.474	0.007	0.006	1.367	1.369	1.399	1.388	0.268	−
C5−C6	1.418	1.419(2)	1.415(5)	0.444	0.010	0.012	1.402	1.402	1.439	1.439	0.211	−
C6−N1	1.391	1.391(2)	1.393(2)	0.476	0.007	0.005	1.375	1.385	1.405	1.400	0.194	−
C5−N7	1.388	1.388(1)	1.389(3)	0.673	0.006	0.007	1.373	1.380	1.402	1.401	0.229	−
N7−C8	1.304	1.305(1)	1.304(3)	0.672	0.006	0.008	1.292	1.392	1.317	1.316	0.241	−
C8−N9	1.373	1.374(1)	1.374(4)	0.911	0.007	0.009	1.362	1.363	1.388	1.388	0.175	−
N9−C4	1.374	1.375(2)	1.377(2)	0.452	0.008	0.006	1.361	1.371	1.397	1.389	0.341	0.025
C2−N2	1.337	1.341(2)	1.341(3)	0.931	0.010	0.008	1.328	1.328	1.368	1.352	0.286	0.150
C6−O6	1.238	1.237(2)	1.239(5)	0.750	0.009	0.014	1.223	1.225	1.258	1.270	0.184	−
N9−C1′	1.461	1.459(2)	na	na	0.009	na	1.438	na	1.469	na	0.279	0.150
C6−N1−C2	125.1	125.1(1)	124.9(2)	0.474	0.6	0.5	123.0	123.9	125.8	125.7	0.336	0.025
N1−C2−N3	123.7	123.9(1)	124.0(2)	0.735	0.6	0.4	123.2	123.3	125.4	124.5	0.286	0.150
C2−N3−C4	112.0	111.9(1)	111.8(1)	0.334	0.5	0.2	110.7	111.5	112.9	112.1	0.283	0.150
N3−C4−C5	128.7	128.6(1)	128.4(2)	0.282	0.5	0.4	127.6	127.8	129.5	129.2	0.176	−
C4−C5−C6	118.8	118.8(1)	119.1(1)	0.087	0.6	0.2	117.9	118.7	120.2	119.3	0.191	−
C5−C6−N1	111.4	111.5(1)	111.7(2)	0.548	0.5	0.6	110.7	111.0	113.0	112.8	0.260	−
C4−C5−N7	110.8	110.8(1)	110.8(2)	0.866	0.4	0.4	109.5	110.2	111.6	111.4	0.390	0.010
C5−N7−C8	104.4	104.3(1)	104.2(3)	0.681	0.5	0.8	103.3	102.6	105.1	105.0	0.218	−
N7−C8−N9	113.1	113.1(1)	113.5(4)	0.336	0.5	0.9	112.0	112.7	114.5	115.4	0.357	0.010
C8−N9−C4	106.4	106.4(1)	106.0(2)	0.168	0.4	0.6	105.7	105.0	107.2	106.6	0.201	−
N9−C4−C5	105.4	105.4(1)	105.6(1)	0.069	0.4	0.2	104.7	105.3	106.4	105.8	0.165	−
N3−C4−N9	125.9	126.0(1)	126.0(2)	0.992	0.6	0.5	124.9	125.4	127.1	126.9	0.244	−
C6−C5−N7	130.5	130.4(1)	130.1(2)	0.254	0.6	0.5	128.8	129.3	131.5	130.7	0.258	−
N1−C2−N2	116.4	116.2(2)	116.3(2)	0.720	0.9	0.5	113.2	115.8	117.5	117.0	0.317	0.050
N3−C2−N2	119.8	119.9(1)	119.7(2)	0.522	0.7	0.5	118.9	119.3	121.5	120.8	0.222	−
N1−C6−O6	120.0	119.9(1)	120.0(2)	0.654	0.6	0.6	117.7	119.2	121.0	121.0	0.341	0.025
C5−C6−O6	128.7	128.6(1)	128.3(2)	0.183	0.6	0.4	127.7	127.7	129.6	128.7	0.266	−
C8−N9−C1′	127.4	127.0(3)	na	na	1.3	na	124.6	na	129.2	na	0.236	−
C4−N9−C1′	126.3	126.5(3)	na	na	1.3	na	124.3	na	129.0	na	0.159	−

of the C5−C6 bond differs at the 5% significance level from the corresponding value in Taylor and Kennard.[2]

Protonated Adenine. The geometrical parameters obtained for protonated adenine are listed in Table 7, and the frequency distributions for the sample geometries are shown in Figure 3. At the 5% level of significance, the Kuiper test shows the sample of N1−C2 bonds to be non-normal in their distributions at the 5% level, as well as the C2−N3−C4, C4−C5−C6, N3−C4−N9, and N1−C6−N6 angles. For the sugar-derivatized structures the C4−C5−C6 and C5−C6−N1 samples (data not shown) are non-normal. None of the means differ from the corresponding values in Taylor and Kennard[2] at the 5% significance level.

Guanine. The geometrical parameters obtained for guanine

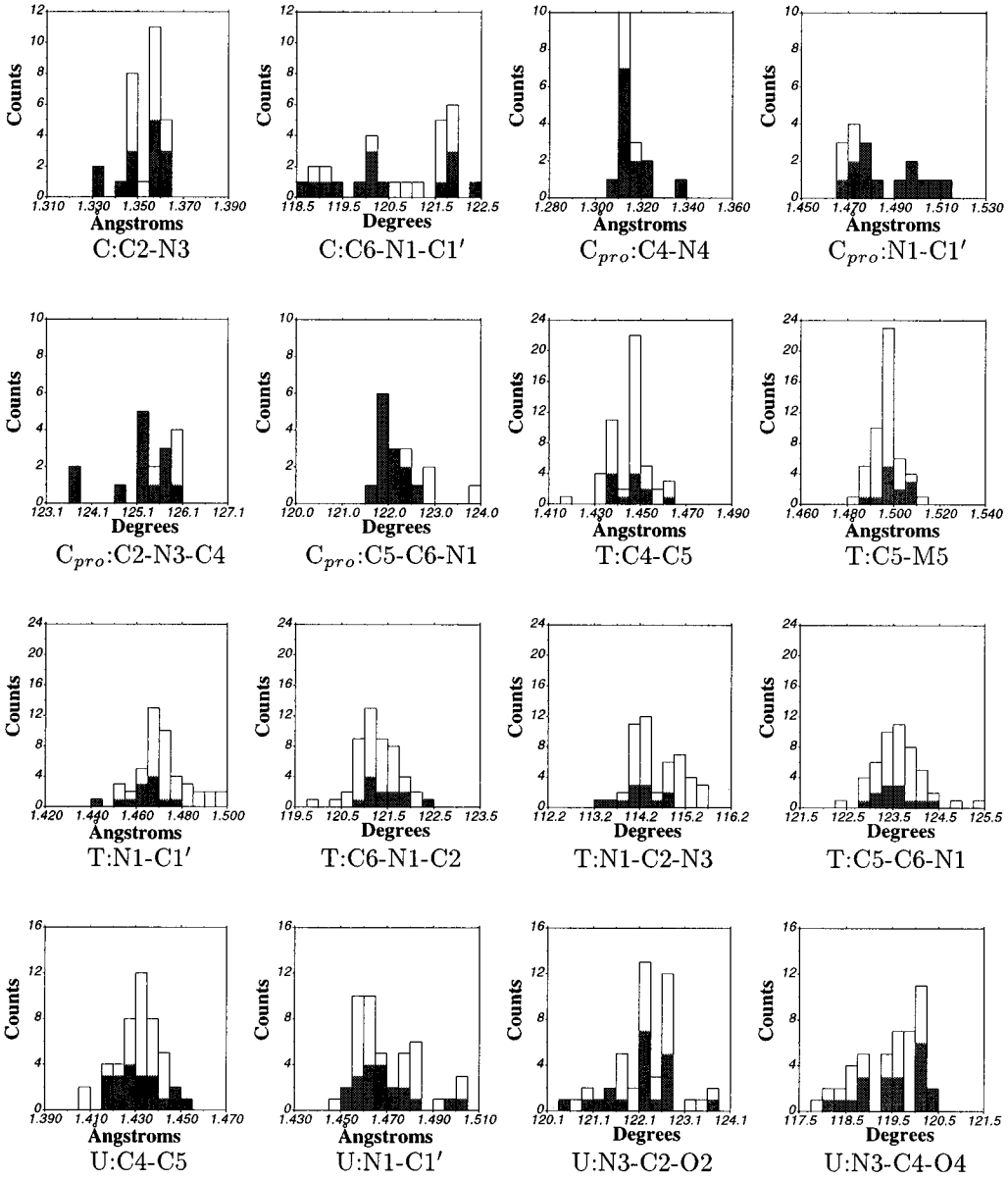

Figure 2. Frequency distributions of bond lengths and valence angles for pyrimidines which are non-normal at the 5% level or less. The label under each histogram corresponds to base and geometric parameter, e.g., C:N1−C2 is the N1−C2 bond of cytosine. The widths of the bond frequency plots are 0.08 Å with bin widths of 0.005 Å, while those of the angle frequencies have widths of 3° with bin widths of 0.6°. The dark areas correspond to the frequencies (shown as counts) for the sugar-derivatized bases, and the empty boxes, to the bases in the full set.

are reported in Table 8, and the frequency distributions for the non-normal sample geometries in Figure 3. At the 5% level of significance, the Kuiper test shows non-normal distributions for N1−C2 and N9−C4 bond lengths and the C6−N1−C2, C4−C5−N7, N7−C8−N9, N1−C2−N2, and N1−C6−O6 angles. None of the parameters have means distinguishable at the 5% significance level from those reported by Taylor and Kennard.[2] Among the geometries for sugar-derivatized structures, only bond N9−C4 and angles C5−C6−N1, N1−C2−N2, and N1−C6−O6 are distributed non-normally at the 5% significance level (data not shown).

Base Planarity and Consistency of Parameters. The planarity of the bases in the sample was judged by examining the average values for a number of torsion angles. The torsions generally lie within 1 sem of either 0° or 180° (the angles in a perfect plane), and the differences are never more than two times this. The average values of the bonds and angles were used to find sets of coordinates for "average" base residues that minimize eq 4. Planarity was enforced by fixing the value of the *z* coordinates at 0. These coordinates were then used to calculate bond distances and angles; the largest difference between the bond distances and angles for the idealized base coordinates and the corresponding averages in small molecules (as seen in Tables 2−8) was 0.001 Å for distances and 0.1° for angles. These were also compared with the corresponding structures generated by Parkinson et al.[9] (using the program X-PLOR with a dictionary based on these average geometries). The rms deviations between the two sets of standard coordinates are less than 0.001 Å.

Discussion

This report presents an updated survey of the bond distances and angles of the nucleic acid bases found in small molecule crystal structures from the Cambridge Structural Database. The values obtained are the best estimates for these parameters and can form the basis for dictionaries used for refinement and model building of nucleic acids. The values are generally similar to

(9) Parkinson, G.; Vojtechovsky, J.; Clowney, L.; Brünger, A. T.; Berman, H. M. *Acta Crystallogr. D* **1996**, in press.

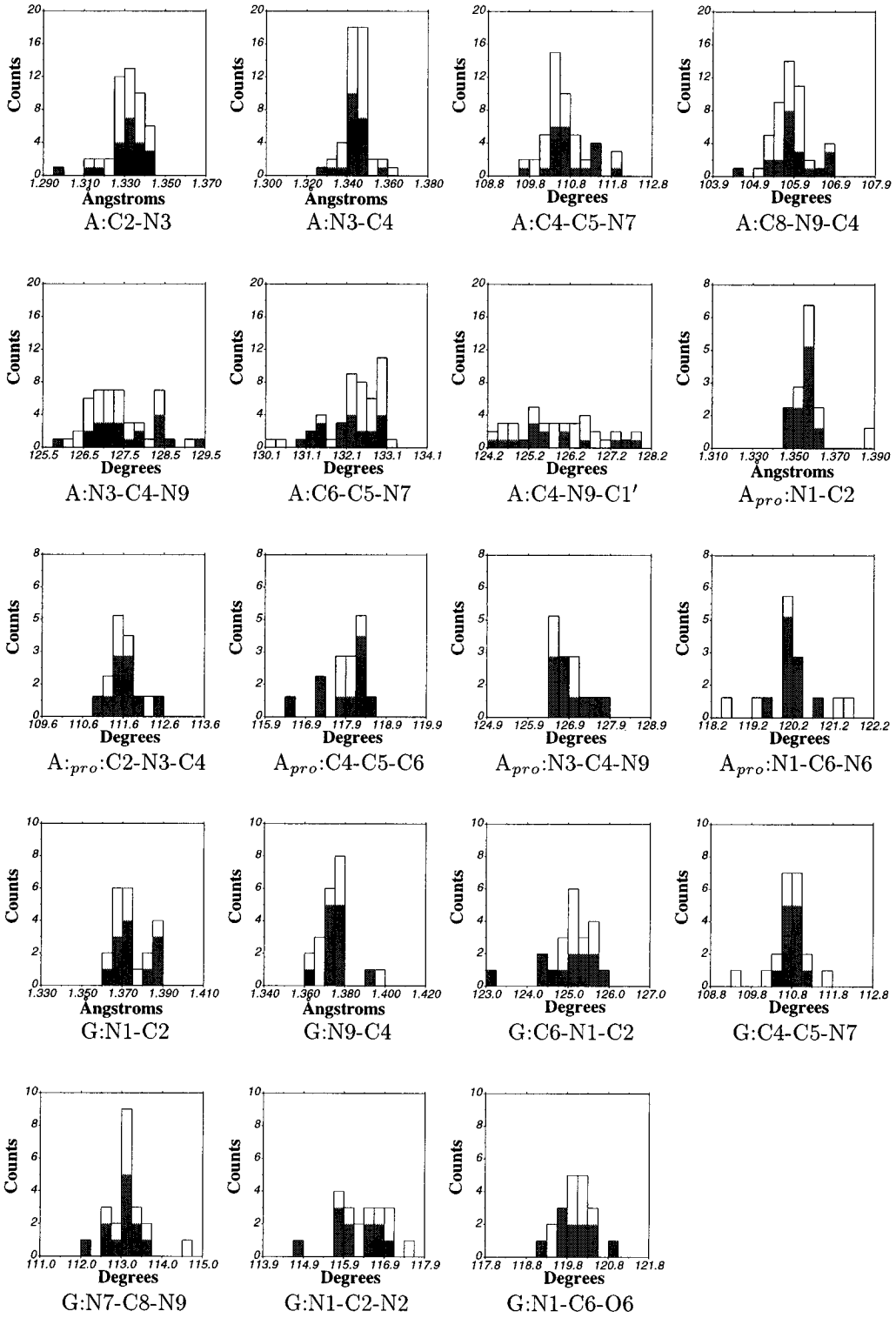

Figure 3. Same as Figure 2 except for purines.

those obtained by Taylor and Kennard,[2] but because the selection criteria for structures are more strict and the sample sizes larger, the standard errors of the mean are often cut in half. In addition, because of the larger sample sizes, it is possible to obtain independent sets of standard values for thymine and uracil.

The significance of inhomogeneities in the samples due to environmental effects, such as differences in chemical substitutions, forces of crystal packing, or varying degrees of hydrogen bonding, has been examined in two ways. First, comparison of corresponding parameters in sugar-derivatized bases versus all bases of a given type shows that, with few exceptions, the mean values do not differ at the 5% significance level. The histograms generally share the same overall distribution as well. Secondly, the effect of general environmental factors was estimated by quantifying the normality of samples and by examining the corresponding histograms of frequency distributions. The results can be understood by noting that crystal packing forces should have a much greater effect when bonds or angles include exocyclic atoms or in the cases when hydrogen bonding occurs.

Excluding atoms N1 or N9 where variable substitutions occur, the non-normal distributions are associated with the presence of an exocyclic carbonyl or amino group or with inclusion of potential hydrogen bond donor or acceptor atoms in the bond

Table 9. Effect of Sugar Substitution on Thymine Geometry[a]

	thymine, $N = 50$, $N_{sugar} = 12$						
parameter	$\bar{x}$	$\bar{x}_{sugar}$	P_t	σ	σ_{sugar}	P_V	$P_{V\ sugar}$
N1−C2	1.376	1.380	0.049	0.008	0.005	−	0.150
C2−N3	1.373	1.376	0.267	0.008	0.009	−	0.025
N3−C4	1.382	1.383	0.662	0.008	0.008	−	−
C4−C5	1.445	1.446	0.612	0.009	0.007	0.010	−
C5−C6	1.339	1.340	0.418	0.007	0.005	0.100	−
C6−N1	1.378	1.383	0.020	0.007	0.006	−	−
C2−O2	1.220	1.216	0.075	0.008	0.005	−	0.150
C4−O4	1.228	1.229	0.748	0.009	0.008	−	−
C5−M5	1.496	1.499	0.249	0.006	0.007	0.025	−
N1−C1′	1.473	1.463	0.006	0.014	0.009	0.025	−
C6−N1−C2	121.3	121.5	0.178	0.5	0.4	0.025	−
N1−C2−N3	114.6	114.2	0.030	0.6	0.5	0.010	−
C2−N3−C4	127.2	127.3	0.427	0.6	0.7	0.100	0.100
N3−C4−C5	115.2	115.3	0.636	0.6	0.6	−	−
C4−C5−C6	118.0	118.0	0.833	0.6	0.4	−	−
C5−C6−N1	123.7	123.5	0.399	0.6	0.4	0.050	−
N1−C2−O2	123.1	123.7	0.023	0.8	0.7	−	0.150
N3−C2−O2	122.3	122.1	0.289	0.6	0.7	−	−
N3−C4−O4	119.9	120.1	0.413	0.6	0.7	−	−
C5−C4−O4	124.9	124.7	0.305	0.7	0.8	−	−
C4−C5−M5	119.0	118.8	0.326	0.6	0.6	−	−
C6−C5−M5	122.9	123.1	0.125	0.6	0.4	−	−
C6−N1−C1′	120.4	119.5	0.096	1.5	1.6	−	0.100
C2−N1−C1′	118.2	118.8	0.232	1.6	1.7	0.100	0.025

[a] The parent set, where any substitution is allowed, is compared with the sugar-derivatized one. P_t is the significance level for the *t* test, and P_V, the significance level for the Kuiper test, as described in the text. The sugar-derivatized subset is denoted by the subscript sugar.

Table 10. Differences between Uracil and Thymine[a]

	$\bar{x}$		
parameter	uracil	thymine	P_t
N1−C2	1.381(1)	1.376(1)	0.005
C2−N3	1.373(1)	1.373(1)	0.754
N3−C4	1.380(1)	1.382(1)	0.346
C4−C5	1.431(1)	1.445(1)	0.000
C5−C6	1.337(1)	1.339(1)	0.469
C6−N1	1.375(1)	1.378(1)	0.079
C2−O2	1.219(1)	1.220(1)	0.553
C4−O4	1.232(1)	1.228(1)	0.054
C6−N1−C2	121.0(1)	121.3(1)	0.032
N1−C2−N3	114.9(1)	114.6(1)	0.012
C2−N3−C4	127.0(1)	127.2(1)	0.078
N3−C4−C5	114.6(1)	115.2(1)	0.000
C4−C5−C6	119.7(1)	118.0(1)	0.000
C5−C6−N1	122.7(1)	123.7(1)	0.000
N1−C2−O2	122.8(1)	123.1(1)	0.135
N3−C2−O2	122.2(1)	122.3(1)	0.471
N3−C4−O4	119.4(1)	119.9(1)	0.001
C5−C4−O4	125.9(1)	124.9(1)	0.000

[a] The means of parameters common to both bases are compared using the *t* test.

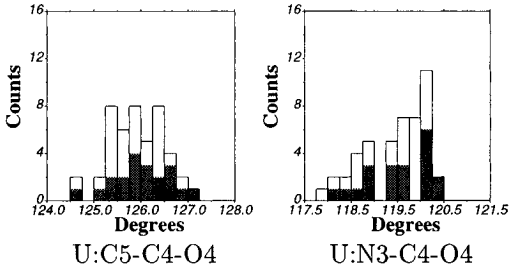

Figure 4. Frequency distribution of uracil valence angles C5−C4−O4 and N3−C4−O4. See legend to Figure 2. The valence angle C5−C4−O4 shows a normal distribution, whereas N3−C4−O4 does not.

or angle under consideration. The most likely hydrogen bond acceptors in the rings are those atoms with the highest electron density. The N3 of cytosine, for example, is a potential acceptor in the neutral base but is a potential H-bond donor when protonated. In adenine, in descending order of preference, acceptors are N7, N1, and N3, with N1 being a potential H-bond donor in protonated adenine, whereas in guanine, the order of acceptors is N7 and N3, with N1 again being a potential proton donor. Except for N3 of guanine, all of the above atoms are involved in non-normal bond-length and/or valence-angle distributions (see Figures 2 and 3).

An alternative test to check for environmental effects on nucleobase structures was used in Taylor and Kennard,[10] where a weighted χ^2 test was used to determine whether the observed variances of bond or angle distributions could be accounted for by uncertainty in the data. Since the test requires knowledge of the experimental standard deviations for individual bonds and angles, information not available in CSD, the χ^2 analyses of only a few distances and angles were considered in Taylor and Kennard.[10] These few χ^2 analyses compared with the current ones that have been analyzed using the Kuiper test.

The current results agree with those reported in Taylor and Kennard[10] for bond distance N7−C8 of adenine and angle N1−C2−N3 of cytosine. In contrast to the previous work, the current study found no bias evident in the sample of uracil C5−C4−O4 angles, according to either the Kuiper test (Table 5) or the corresponding histogram (Figure 4). Also, the current study found the adjacent N3−C4−O4 angle to be significantly non-normal (Table 5), and the frequency distribution clearly skewed (Figure 4). While in principle both these estimators of environmental effects are reasonable, the lack of experimental errors for individual atoms in the CSD makes using the χ^2 test impossible without reference to the original papers describing the structures of interest. The Kuiper test does not have this requirement and is in general consistent with the histograms.

It should be noted that while sugar substitution generally has a negligible effect on the mean bond lengths and valence angles, this is not to say that the nature of the N-linkage is unimportant. Ideally the model structures should be as similar to the target nucleic acids as possible so that when the number of nucleoside structures becomes great enough, or when high-resolution

(10) Taylor, R.; Kennard, O. *Acta. Crystallogr.* **1983**, *B39*, 517−525.

oligonucleotide structures become available, the sugar-derivatized bases would be the preferred structures to use for statistical analysis.

The values presented here are the best current estimates of nucleobase geometry in high-resolution X-ray structures. They confirm and extend an earlier survey of base geometry[2] and should be of immediate use as target values in the constrained refinement of nucleic acid structures[11] and for parameterizing force fields such as those used in molecular dynamics programs.[12,13] The set of mean values and standard deviations, as well as the coordinates for the idealized base geometries, are available electronically over the World Wide Web (http://ndbserver.rutgers.edu) and will be updated as more high-resolution structures are collected.

Acknowledgment. This work was supported by a grant from the NSF (BIR9305135) for the Nucleic Acid Database Project. We would like to thank Nitya Srinivasan for collecting preliminary results and Francisco Figueirido and Zhilrang Ying for helpful discussions.

JA952883D

(11) Brünger, A. T.; Kuriyan, J.; Karplus, M. *Science* **1987**, *235*, 458−460.

(12) Mackerell, A. D., Jr.; Wiórkiewicz-Kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946−11975.

(13) Cornell, W.; Cieplak, P; Bayly, C.; Gould, I.; Merz, K., Jr.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. *J. Am. Chem. Soc*. **1995**, *117*, 5179−5197.