Global mapping of nucleic acid conformational space: dinucleoside monophosphate conformations and transition pathways among conformational classes

Gregory E. Sims and Sung-Hou Kim*

Department of Chemistry, University of California at Berkeley and Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Received June 11, 2003; Revised and Accepted August 4, 2003

ABSTRACT

A global conformational space of 6253 dinucleoside monophosphate (DMP) units consisting of RNA and DNA (free and protein/drug-bound) was ‘mapped’ using high resolution crystal structures cataloged in the Nucleic Acid Database (NDB). The torsion angles of each DMP were clustered in a reduced three-dimensional space using a classical multi-dimensional scaling method. The mapping of the conformational space reveals nine primary clusters which distinguish among the common A-, B- and Z-forms and their various substates, plus five secondary clusters for kinked or bent structures. Conformational relationships and possible transitional pathways among the substates are also examined using the conformational states of DNA and RNA bound with proteins or drugs as potential pathway intermediates.

INTRODUCTION

As the 50th anniversary of the double helix passes, there has been an increasing amount of interest in the variation observed in common nucleic acid conformations and their functional implications. Nucleic acid structure is far from the perfectly regular helices first observed in fiber diffraction studies. It is now known that they are conformationally dynamic and there are several stable substates of more common forms. Initially viewed as perturbations caused by crystal packing (1), these substates are important not just for their structural variation but also for their implications in protein–DNA/RNA complexes. As classified by the dinucleoside monophosphate conformation (Fig. 1A), A-form nucleic acids are known to exist in two main conformations, the canonical (A_I) and the minority ‘crankshaft’ substate (A_{II}). One substate differs from the other about the P-O5′-C5′ linkage, generating a crankshaft effect (Fig. 1B), while keeping the relative positions and orientations of bases approximately the same. In addition, B-form DNA is observed in the Watson–Crick B_I substate and the B_{II} substate, which differ about the P-O3′-C3′ linkage (2).

Also, Z-DNA (3) is found in two conformational families designated Z_I and Z_{II} (4).

The biological importance of these conformational substates is becoming more evident. For example, B- to A-form transitions have been implicated in the mechanism of DNA ligases and the recognition of single-strand nicks (5). Some have suggested that the B_I – B_{II} substate interconversion is also important in protein–DNA recognition (5,6). Recent studies by Zhao *et al.* have indicated that Z-DNA might be vital in controlling gene activity. They showed that part of the regulatory sequence of an immune system gene must flip into a Z-DNA conformation before gene activation (7). In combination with the sequence specificity of protein–nucleic acid base contacts, the addition of substates of backbone conformations allows for a highly variable lexicon of interactions.

Nucleic acids are dynamic molecules converting among different substates in response to their changing environment. Studies of protein-bound nucleic acids present a unique opportunity for studying DNA/RNA structures which are intermediate between canonical forms. Jones *et al.* revealed that helical parameters for protein- or drug-bound DNA were intermediate to those observed for free and unbound B- and A-form DNA (8). Some free oligomer crystal structures have also revealed helical properties which lie on a continuum between A- and B-form DNA (9). Also, 13 brominated and methylated structures which possess helical properties that could be arranged as a series of ‘intermediates’ along a transition from B to A were solved under relatively mild conditions by Vargason *et al.* (10,11). This leads one to conjecture that a possible transition pathway between various substates can be constructed by sampling and arranging empirically observed structures in a systematic fashion. To accomplish this analysis, DNA structures must first be mapped into a unified space which describes the conformational relationships among substates and more exotic structures.

Since Ramachandran and Sasisekharan (12) constructed a plot describing the conformations of polypeptides, many groups have tried to apply the same mapping technique to the conformational space of nucleic acids. This proved very difficult because although the simple dipeptide conformational space can be completely described by a two-dimensional ϕ – ψ angle plot, a complete description of the analogous unit in

*To whom correspondence should be addressed. Tel: +1 510 486 4333; Fax: +1 510 486 5272; Email: shkim@cchem.berkeley.edu

Figure 1. (A) The 10 torsion angle representation of the DMP unit. The first residue (subscript 1) is truncated at the γ angle and the second residue (subscript 2) does not include torsion angles ϵ and ζ . In both residues the torsion angle δ is ignored in preference for the pseudorotation and P , which yields more information about sugar pucker. Superscripts are used when referencing torsion angles to implicitly indicate angle location. The atoms outlined in bold were used in both RMSD and atom-atom distance comparisons. (B) Superpositions of DMP substates (weighting the C3', C4' and P atoms), emphasizing the 'crankshaft' nature of phosphate-sugar linkages. The bases have been truncated and replaced with a cyclohexyl group indicating base planarity. (Top) A-form type I (red) and type II (blue) differ at the P-O5'-C5'-C4' linker, in a crankshaft fashion involving the angles α , β and γ . (Bottom) B-form type I (red) and type II (blue) differ by the P-O3'-C3' linkage, involving backbone angles ϵ and ζ .

nucleic acids, the dinucleoside monophosphate (DMP), requires 10 torsion angles (Fig. 1A). One approach has been to plot torsion angles in a pair-wise manner, examining their interdependence and interpreting various correlations (13–15); another has been to reduce structure to several variables like pseudo-angles (16) or helical parameters (17). Attempting to correlate multiple two-angle plots into a unified conformational space becomes a formidable task and selective reduction of variables is subject to some interpretive bias. Nonetheless, it has proven difficult to reduce the high dimensionality space of nucleic acids to a simple planar representation.

Multivariate analysis provides a more straightforward approach for visualizing DMP space, by removing the problems associated with interpreting multiple correlations between angles. For example, Beckers and Buydens analyzed a small data set of A-DNA and B-DNA DMP steps employing

the singular value decomposition method (18). In this study, a matrix of torsion angles was decomposed and projected into a three-dimensional space, revealing the clustering of A-form for both the common state (A_I) and the crankshaft class (A_{II}) and B-form for both substates B_I and B_{II} . Techniques of this sort have several advantages over two-dimensional two-angle representations, namely all torsion angles from the model are used to form the multi-dimensional map and global clustering may be easily observed.

Since the work of Becker and Buydens in 1998, a tremendous body of high quality nucleic acid structures has become available. In our method we have constructed a three-dimensional representation of the conformational space of 6253 DMP structures using the classical multi-dimensional scaling (MDS) method, a type of principal components analysis. MDS clusters data by grouping DMPs into natural clusters by comparing all the pair-wise distances

(conformational dissimilarities) between two structures including DMPs in free nucleic acid structures as well as those bound to proteins or drugs. We also make inferences about transitions between conformations and substates by analyzing the DMP structures from nucleic acids bound to proteins and drugs as potential transition intermediates between substates. Our map provides a unique space where all DMP structures are pre-ordered in a systematic way, i.e. the structural dissimilarity is proportional to the distance observed in the map space. We employ a single source shortest path algorithm to link these intermediates between a given two substates to predict hypothetical transition pathways between the substates.

MATERIALS AND METHODS

Data

In this work, we started by culling the nucleic acid database (NDB) (19) to a subset of structures with better than 2.0 Å resolution. This resolution cut-off was chosen because the orientation of the phosphate group can be fully distinguished from experimentally observed electron density maps, which is necessary for assigning substates which differ by, for example, a 'crankshaft' movement. By limiting the analysis to structures better than this resolution, only structures without ambiguity in the electron density fitting are included. Next, each PDB structure was divided into overlapping dinucleotide units by stepping through each nucleic acid chain excluding 5' and 3' terminal nucleotides as well as modified synthetic nucleotides. The ensemble of structures contains the common A-, B- and Z-form nucleic acids, including both RNA and DNA in free as well as protein- or drug-bound states. Homologous nucleic acid sequences were included, so the data set contains some redundancy in an effort to represent the true conformational variability within sequences. The list of structures and residues employed has been included as Supplementary Material at NAR Online. A total of 6253 DMPs form the complete data set. A smaller 204 DMP unit data set was also constructed from high resolution structures at better than 1.0 Å resolution for the purpose of evaluating the effects of various distance metrics (described below). The torsion angles for each unit were extracted from the coordinates using AMIGOS (16). As seen in Figure 1A, the DMP unit is described by 10 backbone torsion angles γ , ϵ , ζ , P and χ for the 5' nucleotide (denoted with subscript 1) and α , β , γ , P and χ for the 3' nucleotide (subscript 2). The pseudorotation angle, P , which describes the puckering of the ribose sugar, was determined using the method of Altona and Sundaralingam (20). The conformation of each DMP unit can then be described as a 10 member vector.

Multidimensional scaling

Our task is to embed the higher dimensional DMP space into a lower dimensional Cartesian space by employing classical (metric) MDS (21). We used MDS as implemented in the cmdscale function from the multivariate analysis library package of R (22). MDS will transform a matrix of pair-wise distances, $\mathbf{D}$, into a three-dimensional Cartesian space where the Euclidean distances in the reduced space are proportional to the original higher dimensional distances. A

more detailed discussion of the method has been included as Supplementary Material at NAR Online. If a more detailed account of multi-dimensional scaling is desired see Duda and Hart (23).

Choice of distance matrices

A matrix of pair-wise dissimilarities, $\mathbf{D}$, must be constructed using a defined distance metric. Three distance metrics are possible for the study of DMPs: (i) angular Euclidean distance; (ii) root mean square deviation (RMSD); (iii) pair-wise atomic distance matrices. Reijmers *et al.* have tested the consistency of results from these three different distance metrics in clustering trinucleoside structures (24).

(i) The customary metric in cluster analysis is the Euclidean distance (25). However, when measuring the dissimilarity between two torsion angles one must take account of the circular properties of angles. Reijmers *et al.* point out that all angle differences must be $<180^\circ$ or structural similarities will be distorted near angles around 360° (26). A distance matrix, D , can be constructed using a modified form of Euclidean distance which accounts for circularity by always calculating the angular distance which is $\leq 180^\circ$. The form of the distance matrix is calculated as

$$D_{ij} = d(x_i, x_j) = \sqrt{\sum_{k=1}^T \begin{cases} \text{if } 180 < |x_{ik} - x_{jk}| < (360 - |x_{ik} - x_{jk}|) \\ \text{if } 180 \geq |x_{ik} - x_{jk}| < (x_{ik} - x_{jk})^2 \end{cases}} \quad \mathbf{1}$$

where T , the number of torsion angles in each DMP vector, is equal to 10.

(ii) Simple RMSDs of each Cartesian representation, using a quaternion superposition method (27), were also used to find distances from the atomic coordinates of the atoms associated with highlighted bonds in Figure 1A. The phosphate atoms were weighted twice as heavily as all other atoms during superpositioning.

(iii) Pair-wise atomic distance based dissimilarity scores were constructed from an all atom-atom distance matrix for each DMP. The atoms employed in the distance comparisons are the same as those highlighted in Figure 1A. The norm of the pair-wise difference between each atom-atom distance matrix was used as a representation of the dissimilarity.

Since we were not able to distinguish which dissimilarity metric was best a priori, all methods were initially tested using the high 1.0 Å resolution data set. The eigenvalues from the MDS procedure, using all of the dissimilarity metrics, were extracted and plotted against the index value in a scree plot (Fig. 2A). For our purpose, the reduced dimensional space is most valuable for us if the first three largest eigenvalues are much larger than all other eigenvalues, which indicates that most of the variance in the data can be explained employing a basis set composed of the first three eigenvalues. This form of visual inspection is commonly known as Cattell's scree test (28). The shoulder point (or point of inflection) for the torsion angle metric occurs at the third variable and approaches a minimum value faster than other metrics, indicating that the 10 angle representation is slightly better and is adequate to represent the data in a three-dimensional space. In addition, Euclidean angle distances were employed in the final results with the large data set because it distinguished between 'crankshaft' linkages best, where RMSD comparisons fail. Disparate arrangements of torsion angles can have very

similar global conformations in DMP space, as can be seen from superpositions of type I A-form and B-form with their type II substates (Fig. 1B). Like the higher resolution data set,

Figure 2. Eigenvalue index versus normalized eigenvalue magnitude. (A) The results of MDS on a small 204 DMP unit data set (1.0 Å resolution or better) using three distance metrics: RMSD superposition with a quaternion algorithm (square), angular Euclidean distance (circle) and an all atom–atom distance matrix comparison (triangle). The torsion angles approach an inflection point more rapidly and it appears that three eigenvalues are adequate to represent the data set. (B) Eigenvalues of MDS on 6253 DMP unit data set (2.0 Å or better) using angular Euclidean distance.

most of the variance is explained in the 2.0 Å data set by the first three or four eigenvalues using a torsion angle distance metric (Fig. 2B).

Global mapping of nucleic acid conformations

After producing a 6253 point global map of DMP space (Fig. 3A), all bound structures (either from protein–DNA/RNA or drug–DNA/RNA complex structures) were temporarily removed from the map space, leaving only the free nucleic acid DMPs (Fig. 3B). The DMP space is well clustered and a visual inspection easily reveals the presence of nine clusters. A sphere of a radius sufficient to fully encompass the densest region of each cluster was circumscribed around the center of each cluster. These spheres served as approximate boundaries defining the volume and locus of each cluster (Fig. 3C). Next, all DMP units within each of these spheres (from both the bound and free data sets) were extracted and the torsion angle distributions were analyzed. The results for each cluster are listed in Table 1. The reduced space coordinates for all 6253 DMPs and the coordinates of each boundary sphere (representing the ideal DMP structure) were appended together as one list of coordinates. This set is designated as the graph data set. Five secondary clusters (outlined in Fig. 3C), consisting of kinked or bent structures, were also isolated from the DMP space and categorized in Table 2.

Transitional pathways

Graph theory is applied to the graph data sets and used to investigate ‘transitions pathways’ between substates. The goal is to extract any empirically observed evidence for hypothetical transition paths between canonical forms through ‘perturbed’ (protein- or drug-bound) DMP structures. It can be argued that the intermediates in one transition pathway (among several possible pathways) lie along the shortest path between empirically observed structures. A simple graph is depicted in Figure 4. Each DMP can be represented as a vertex, where edges connect various vertices. The weight w_{ij} of each edge is set to be equal to the distance between the two vertices i and j . If the weight between two vertices i and j exceeds a cut-off value w_c , then the weight is reset to infinity. A direct path from vertex i and j with infinite weight is now impossible. Our goal is to navigate from some source vertex in the graph to a target vertex, following an optimal path through the graph. Additionally, no vertices may be connected if they

Table 1. Average torsion angles for principal nucleic acid conformers

	A _I	σ ^a	A _{II}	σ	B _I	σ	B _{II} A	σ	B _{II} B	σ	Z _I A (YR)	σ	Z _I B (RY)	σ	Z _{II} A (YR)	σ	Z _{II} B (RY)	σ
γ ₁	174	16	180	10	171	22	171	15	174	18	184	9	208	33	187	5	215	30
P ₁	40	8	38	4	301	96	322	6	322	62	26	43	325	4	24	6	327	4
ε ₁	82	6	82	5	127	18	143	8	128	16	96	8	143	7	94	6	147	5
ζ ₁	208	10	200	10	185	15	239	24	191	12	239	9	267	6	187	9	262	6
χ ₁	287	10	290	7	262	14	172	15	269	10	300	14	75	6	66	21	74	4
α ₂	201	11	188	7	246	17	251	22	269	10	205	7	64	5	213	7	58	3
β ₂	294	11	145	10	298	32	291	45	294	11	204	11	69	10	168	11	69	6
γ ₂	172	10	190	10	176	16	145	13	173	10	228	15	184	6	163	13	186	6
P ₂	41	16	38	5	305	92	327	33	324	6	326	5	35	69	319	52	24	6
χ ₂	287	13	285	20	262	17	265	12	175	15	74	9	299	13	73	5	62	6
N	2108		180		1771		395		420		123		63		32		22	
[Σ _{i=1} ^N d(a _i , <X>)]/N ^b	25		26		61		38		43		29		30		24		41	

^aStandard deviation values for angles in cluster.

^bAverage distance (equation 1) of all a_i DMPs from mean DMP, <X>.

Table 2. Average torsion angles for secondary nucleic acid conformers

	S ₁	σ	S ₂	σ	S ₃	σ	S ₄	σ	S ₅	σ
γ_1	194	2	165	7	186	11	171	9	178	2
P_1	34	0.5	329	3	37	2	325	8	331	2
ϵ_1	84	1	144	5	92	3	144	5	145	1
ζ_1	208	1	244	2	216	6	255	13	274	4
χ_1	295	1	88	4	297	7	185	5	145	2
α_2	197	2	266	1	252	8	201	5	197	2
β_2	138	3	279	3	312	8	62	7	111	3
γ_2	188	1	157	2	176	2	226	8	243	3
P_2	35	0.4	32	1	330	3	30	12	35	3
χ_2	354	3	274	2	141	4	303	30	222	3
N	8		6		8		15		4	

S₁, monovalent cation-binding site (I); S₂, left-handed single-stranded RNA helix (GA); S₃, left-handed single-stranded RNA helix (AU); S₄, kinked triplex formation and poly(A) tract; S₅, uridine bridge.

represent DMPs with different ribose sugars, which prevent transitions between RNA and DNA structures.

To create the DMP graph, a pattern of edges connecting all neighboring points with w_c set to 0.05 (in the reduced space) was constructed from the graph data set. Using proportionality, the corresponding cumulative angle distances can be retrieved from the MDS calculated distances. A reduced space distance of 0.05 is equivalent to a cumulative difference of 37° in all 10 torsion angles. This value is relatively small and on a per angle basis the changes are <4°, less than thermal fluctuation. The closest neighbor vertex of a DMP should be a reasonably similar structure and it should be thermally accessible. The 37° value for w_c seems reasonable and was arrived at from two observations. First, from examinations of the individual clusters the average distance of individual DMPs from the mean structure ranges from 25 to 40° (Table 1). The variation within clusters resembles what is observed in thermal fluctuations. Second, all reasonable paths between clusters become navigable at a w_c of ~35°, whereas at lower values some transitions are impossible because of the large changes in DMP structure that must be made. The parameter w_c is somewhat analogous to the choice of time step chosen in a molecular dynamics simulation. If a larger value of w_c is chosen then larger changes in conformation occur.

Dijkstra's shortest path algorithm is employed to determine the shortest path between each of the boundary centers which represent each substate (29). In this way we can approximate the transitions between different canonical forms and substates by following a path (among many others) through experimentally observed structures. Figure 5 illustrates some of the most biologically important transitions in RNA and DNA, A- to B-form and B- to Z-form. The k th shortest paths were also found for $k = 400$ using Eppstein's implementation (30) to illustrate that many sub-optimal paths exist in the graph set which are similar to the optimal path (Fig. 5, inset).

RESULTS

Nine major (principal) conformations of nucleic acids

In Figure 3A we describe a representation of the DMP space generated from a data set of 6253 DMP structures. Most noticeably, the bound structures (blue) widely populate the

intermediate space bound by three conformational classes: B- (yellow), A- (red) and Z-form (green) structures, although the space is not as well populated in the region between the A- and Z-form clusters. After removing the bound structures, the remaining free nucleic acids present nine well-defined A-, B- and Z-form clusters (Fig. 3B). The resulting clusters have torsion angle distributions which agree with those analyzed from smaller data sets by Schneider *et al.* (31) (Table 1).

In general, RNA and DNA are dispersed throughout the map. However, higher densities of DNA and RNA are localized to particular clusters (Fig. 3D). Helical RNA primarily occupies clusters A_I and A_{II}. RNA (not necessarily helical) dominates the regions of bound nucleic acid conformations which fill the inter-cluster space of the B- and Z-forms. DNA is the overwhelming majority of the content of clusters B_I, B_{II} and Z; however, the A_I cluster space is shared by both RNA and DNA. The hybrid RNA–DNA structures are spread out in the inter-cluster space with the bound RNA. For the following discussion of dihedral angle values we use the following convention: $g+$, $60 \pm 60^\circ$; tr , $180 \pm 60^\circ$; $g-$, $300 \pm 60^\circ$.

A and B forms

Two substates of the A-form are observed, both the canonical form, A_I, and the crankshaft form, A_{II}. Values of α_2 and β_2 discriminate between these two substates, where type A_I has the configuration $\alpha_2(tr)/\beta_2(g-)$ and A_{II} the configuration $\alpha_2(tr-)/\beta_2(tr)$. In the B-form three important clusters exist, the most populated of which is the canonical Watson–Crick form, B_I. The other two forms are two variants of type B_{II}, one of which, B_{IIa}, is most readily distinguished from type B_I by the angles ϵ_1 , ζ_1 , χ_1 , where B_I adopts $\epsilon_1(g-)/\zeta_1(tr)/\chi_1(g-)$ and B_{II} $\epsilon_1(tr)/\zeta_1(g-)/\chi_1(tr)$. Substate B_{IIb} is a cluster of DMP steps immediately preceding a B_{IIa} step in the nucleic acid chain. This second cluster is partially a consequence of the particular representation used for the DMP unit (Fig. 1A). The end of each B_{IIb} step is truncated at ϵ_2 , ζ_2 ; however, the decreased χ_2 angle (from $g-$ to tr) is sufficient to discriminate this substate from B_I. As can be seen from the standard deviations, B_I angles adopt a wider range of conformations (especially the pseudorotation angle) than either A_I, A_{II} or B_{IIa}. Repeating DMPs with type B_{II} configurations are rare, significantly rarer than the B_I–B_{II} (B_{IIa}) and B_{II}–B_I (B_{IIb}) DMP steps, and are not represented in the DMP cluster space.

Z-form

Typically the Z-form is observed as alternating dinucleotide pyrimidine (Y) and purine (R) steps, existing in either of two separate substates, Z_I and Z_{II}. Therefore, we see clusters based on these repeating patterns as either a YR (pyrimidine–purine) DMP step or the reverse RY step in either the common Z_I or the less common Z_{II} variant. This creates four clusters of left-handed nucleic acids in the DMP space. Both Z_{Ia} and Z_{IIa} are RY steps, while Z_{Ib} and Z_{IIb} are YR steps. A distinguishing feature of Z_{II}-type DMPs is the adoption of the $\chi(g-)/\chi_2(g+)$ and Z_{Ib} adopts $\chi_1(g+)/\chi_2(g-)$. Type I and type II alternate sugar puckering, with the pyrimidine nucleosides preferring a C2'-endo *P* and the purines a C3'-endo *P*.

Minor (secondary) conformations of nucleic acids

Several minor clusters which possess interesting features are dispersed within the bound region of the DMP space, as seen in Figure 3C colored in blue. The torsion angle values are characterized in Table 2 for each cluster.

Monovalent cation-binding site (S_1) (e.g. AD0011 residues 4 and 5) (32). Metal binding of an alkali metal cation creates a 12–13° kink at this DMP step.

Short left-handed single-stranded RNA helix (S_2) (PR0009 residues 109 and 110 chain W) (33) (S_3) (PR0009 residues 102 and 103 chain C). The first and last nucleotides of a GAGAU repeating RNA sequence form a short piece of left-handed helix. This structure is stabilized by interactions with the TRAP protein, and might not be stable if unbound.

Kinked (S_4) (BD0051 residues 16 and 17 chain B) (34). TA dinucleotide steps within poly(A) tract DNA exhibit a kinking towards the major groove of ~8°. A similar type of deformation is also observed in some triplex formations (BD0006).

Uridine bulge (S_5) (AR0034 residues 3 and 4 chain A) (35). The uridine bulge assumes a conformation with the uracil flipped out and protruding into the minor groove. This bulge

produces a large twist between flanking nucleotides, creating a kink in the helical axis.

Currently, the majority of high resolution structures are regular helical structures and we expect the minor conformation space to become more defined as the databases become more populated by non-canonical structures.

Correlation of eigenvalues with torsional angles

The correlation between eigenvalues and torsional angles is not simple. The dominant torsion angles can be determined by analyzing R^2 correlation coefficients between angle and component values. The first eigenvalue, the x -axis, separates DMPs by angles $\chi_2 > \chi_1 > \zeta_1 > \gamma_2$ (in order of greatest R^2) dividing B_I from B_{IIa} and B_{IIb} , Z_{IIa} from Z_{IIb} and Z_{Ia} from Z_{Ib} . The second eigenvalue, on the y -axis, separates DMPs by torsions $P_1 > P_2 > \alpha_1 > \epsilon_1$. At the two extremes along this component there are the C3'-*endo* DMPs (A-form) and C2'-*endo* (B-form) and intervening between the two is the Z-form (alternating C2'-*endo* and C3'-*endo*). The third eigenvalue, represented by the z -axis, divides structures by a combination of angles $\chi_1 > \epsilon_1 > P_1 > \chi_2 > P_2 > \beta_2$. The 90° rotation of the DMP space (Fig. 3B) shows the extent of the z -axis separation most clearly. The contribution of one angle to several dimensions makes it difficult to assign the variation observed along an axis to any one torsion component.

Figure 3. Three-dimensional projections of MDS results. DMPs from bound crystal structures (bound either to a protein, drug or part of a hybrid duplex) are colored in blue, while free structures are colored by form: A, red; B, yellow; Z, green. The first three principle components are represented by the x -, y - and z -axes. The dominant DMP torsion angles of each component are indicated, ordered by R^2 correlation coefficient: x ($\chi_2 > \chi_1 > \zeta_1 > \gamma_2$), y ($P_1 > P_2 > \alpha_1 > \epsilon_1$) and z ($\chi_1 > \epsilon_1 > P_1 > \chi_2 > P_2 > \beta_2$). (A) A projection of both the bound and free DMPs viewed from two orientations. This lower view (a 90° rotation about the y -axis) better illustrates the dispersion of DMPs along the z -axis. (B) Free DMPs structures viewed from two orientations. Nine major clusters (principal) are distinguishable and divide DMP space by form and substrate. (C) Depiction of spherical boundaries used to isolate each cluster and to determine which exotic DMPs are not members of any cluster. Examples of small minor (secondary) clusters are depicted in blue: [S₁], monovalent cation-binding site; [S₂], left-handed single-stranded RNA helix (start-GA); [S₃], left-handed single-stranded RNA helix (end-AU); [S₄], kinked structures; [S₅], uridine bridge. (D) 2.0 Å data set with coloring arranged to highlight the locations of DNA (cyan), RNA (blue) and DNA-RNA hybrid structure (purple). (E) A graphical representation of the shortest paths through the 2.0 Å DMP graph space. Only the free DMPs are depicted to reduce the complexity of the image. The shortest paths between several primary substrates of interest lie along connecting lines: DNA transitions, blue; RNA transitions, cyan. The cut-off value, w_c , for the graph is 0.05 U (37 cumulative degrees). All maps were rendered using PYMOL v.78[®] (Delano Scientific, <http://www.pymol.org>).

Figure 4. Schematic diagram of some graph space G in a two-dimensional Euclidean space. Vertices 1 and 3 cannot be connected because they are more distant than w_c , the distance cut-off. Therefore the edge weight, w_{13} , must be set to infinity. We wish to find the shortest path (shown highlighted) from some source vertex S to some target vertex T given this set of vertices and edge weights.

‘Transitional pathways’

The basic assumption we are making here is that most of the transition intermediates are represented among a large number of diverse DMP conformations found in the protein- or drug-bound nucleic acid structures. The lack of bound structures in the intermediate space adjoining Z-form and A-form clusters suggests that this region is ‘forbidden’. Also, the most direct path for an A- to Z-form transition would lie through this unpopulated region. The population distribution indicates in a very general manner that an A- to Z-form conversion most likely lies through a B-form intermediate. This is especially curious because it is known that an equilibrium between both the Z- and A-forms and B- and Z-forms has been observed at various conditions of hydrostatic pressure (36–39). Additionally, the space between the type Z_a clusters (both

Figure 5. A 90° rotated view of the shortest paths through the 2.0 Å DMP graph space. The inset depicts the shortest 400 paths for DNA transitions between A_I and B_I , which are very similar in trajectory.

type I and II) and Z_b is unpopulated, consistent with the fact that Z_a is for pyrimidine–purine steps and Z_b is for purine–pyrimidine steps.

Figure 5 graphically represents a series of hypothetical transition paths that link intermediate states between primary substates at $w_c = 37^\circ$. The vertices representing each DMP structure which lie along the shortest path connecting each substate cluster are connected by thick lines. Transitions from an RNA to a DNA vertex are disallowed, so separate transitions are recorded for RNA (blue) and DNA (cyan). Figure 6 shows the angle ‘trajectories’ of each series of transition pathways. To show the fluctuation in each substate cluster, the first 10 DMPs in the trajectory are sampled from the center of each starting cluster and the last 10 from the target cluster. The standard deviations for each angle from Table 1 for each cluster are outlined by a rectangular box, which depicts the typical fluctuations in each substate at equilibrium. The intermediate transitional structures lie between each of these subsets and lie on a shortest path traversal through the graph space. Most of the pathways do not traverse many vertices and have about five to six intermediate structures; however, the two longest pathways, B_I – Z_{IIa} and B_I – Z_{IIb} , navigate through about 10 vertices. RNA transitions are only allowed between A_I and A_{II} and A_I and B_I .

A to B transition

Surprisingly, the A_I to B_I path shows an almost instantaneous change in pseudorotation angle (Fig. 6A) in the intermediates. However, the P_2 angle is subjected to some wobbling in sugar pucker between $C2'$ -endo and $C3'$ -endo values through the transition. The ζ_1 angle initially rotates to a g -configuration, then rotates in the opposite direction to a tr configuration to accommodate the puckering change of P_1 from $C3'$ -endo to $C2'$ -endo. The base orientation angles, χ_1 and χ_2 , are relatively undisturbed and decrease gradually to lower values. The bases must decrease their tilt in the B-form from 20° , becoming roughly perpendicular to the helical axis in the B-form. Base de-stacking would not be observed in this type of transition, but in order to accommodate the sugar pucker change, ϵ_1 and ζ_1 are perturbed and shifted from higher and lower values, respectively, in a concerted fashion. The

above type of pathway suggests that changes in the ϵ_1 and ζ_1 torsions (i.e. the phosphate group) lead changes in the sugar pucker.

B to Z transition

The changes that occur upon moving from B_I to Z_{IIb} can be summarized as two crankshaft movements about both $C3'$ -O-P and $C4'$ - $C5'$ -O-P linkers in conjunction with a sugar pucker flip and a change in base orientation. In the B_I to Z_{IIb} pathway P_2 oscillates between a $C2'$ -endo and $C3'$ -endo conformation (Fig. 6A). Angles ζ_1 and χ_1 steadily change from tr to g - and tr to g +, respectively. Similar trends are observed in the B_I to Z_{IIa} pathway, however, in the opposite fashion (i.e. χ_2 changes from tr to g).

The key feature of each of these trajectories is that very rarely are the changes in conformation smooth, like the idealized reaction coordinates used as teaching aids. The important torsion angles involved often flutter between source-like and target-like characteristics before settling into final equilibrium values. For example, the pseudorotation angle oscillates between $C2'$ -endo and $C3'$ -endo values in several pathways, with no values intermediate to these two extremes. Some of the more abrupt changes in angles were investigated further by decreasing w_c (as with the treatment of the pseudorotation angle); however, no paths could be revealed with smoother transitions. It is worth noting that there are several near optimal paths which can be investigated, suggesting that there may be multiple pathways for any transition. For example, the first 400 shortest paths for the A_I – B_I transition (Fig. 5, inset) all appear very similar in trajectory to the shortest path. We also recognize that the set of structures we observe in the NDB database is an empirically observed sampling of nucleic acid conformational and energetic space, and finer sampling is expected as more structures are determined.

DISCUSSION

We have created a global view of the conformational space of nucleic acids employing a large data set of 6253 DMP structures. Within this space, nine major conformations of nucleic acid structures, A, B and Z and their subtypes, are

Figure 6. Hypothetical transitions between substates through the shortest path between DMPs. Each trajectory is plotted as a function of the 10 angle vector representing each unit. The first 10 steps were sampled from the starting cluster and the last 10 steps from the ending cluster. Standard deviation values for each angle in a particular substate (from Table 1) outline each trajectory. (A) DNA transitions only: (left) A_I – B_I ; (right) B_I – $Z_I B$. (B) RNA transition: A_I – B_I .

clearly clustered objectively, and protein- or drug-bound DMPs are seen to adopt more diverse conformations spanning outside the nine clusters. We have also identified five minor (secondary) clusters outside the major conformational space, which have interesting structural features that are often coupled to interactions with other molecules. The dynamic interconversions between substates have become increasingly important in the study of protein–nucleic acid interactions and gene regulation. This nucleic acid conformational map provides a platform to study the structural relationships of various substates and possible transition pathways among conformations.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank our colleagues Chao Zhang, In-Geol Choi, Jingtong Hou, Se-Ran Jun, Steve Holbrook and Jennifer Stidman for their help, advice and discussions throughout the course of this work. We are grateful for the support of the National Science Foundation (DBI-0114707), the National Institute of Health for a training grant for G.E.S. (PHS-GM0829515) and the National Energy Research Scientific Computing Center at Lawrence Berkeley National Laboratory, which is supported by the Department of Energy.

REFERENCES

- Dickerson, R.E., Goodsell, D.S., Kopka, M.L. and Pjura, P.E. (1987) The effect of crystal packing on oligonucleotide double helix structure. *J. Biomol. Struct. Dyn.*, **5**, 557–579.
- Prive, G.G., Heinemann, U., Chandrasegaran, S., Kan, L.S., Kopka, M.L. and Dickerson, R.E. (1987) Helix geometry, hydration and G-A mismatch in a B-DNA decamer. *Science*, **238**, 498–504.
- Wang, A.H.-J., Quigley, G.J., Kolpak, F.J., Crawford, J.L., Van Boom, J.H., Van Der Marel, G.A. and Rich, A. (1979) Molecular-structure of a left-handed DNA fragment at atomic resolution. *Nature*, **282**, 680–686.
- Gessner, R.V., Federick, C.A., Quigley, G.J., Rich, A. and Wang, A.H.J. (1989) The molecular-structure of the left-handed Z-DNA double helix at 1.0-Å atomic resolution—geometry, conformation and ionic interactions of d(CGCGCG). *J. Biol. Chem.*, **264**, 7921–7935.
- Peticolas, W.L. (1995) Raman spectroscopy of DNA and proteins. *Methods Enzymol.*, **246**, 389–416.
- Pichler, A., Rudisser, S., Mitterbock, M., Huber, C.G., Winger, R.H., Liedl, A.H. and Mayer, E. (1999) Unexpected BII conformer substate population in unoriented hydrated films of the d(CGCGAATTCGCG)₂ dodecamer and of native B-DNA from salmon testes. *Biophys. J.*, **77**, 398–409.
- Liu, R., Liu, H., Chen, X., Kirby, M., Brown, P.O. and Zhao, K. (2001) Regulation of CSF1 promoter by the SWI/SNF-like BAF complex. *Cell*, **106**, 309–318.
- Jones, S., van Heyningen, P., Berman, H.M. and Thornton, J.M. (1999) Protein–DNA interactions: a structural analysis. *J. Mol. Biol.*, **287**, 877–896.
- Ng, H.L. and Dickerson, R.E. (2002) Mediation of the A/B-DNA helix transition by G-tracts in the crystal structure of duplex CATGGGCCCATG. *Nucleic Acids Res.*, **30**, 4061–4067.
- Vargason, J.M., Henderson, K. and Ho, P.S. (2001) A crystallographic map of the transition from B-DNA to A-DNA. *Proc. Natl Acad. Sci. USA*, **98**, 7265–7270.
- Dickerson, R.E. and Ng, H.L. (2001) DNA structure from A to B. *Proc. Natl Acad. Sci. USA*, **98**, 6986–6988.
- Ramachandran, G.N. and Sasisekharan, V. (1963) Conformation of polypeptides and proteins. *J. Mol. Biol.*, **7**, 95–99.
- Murthy, V.L., Srinivasan, R., Draper, D.E. and Rose, G.D. (1999) A complete conformational map for RNA. *J. Mol. Biol.*, **291**, 313–327.
- Fratini, A.V., Kopka, M.L., Drew, H.R. and Dickerson, R.E. (1982) Reversible bending and helix geometry in a B-DNA dodecamer CGCGAATTBRCGCG. *J. Biol. Chem.*, **257**, 4686–4707.
- Conner, B.N., Yoon, C., Dickerson, J.L. and Dickerson, R.E. (1984) Helix geometry and hydration in an A-DNA tetramer—IC-C-G-G. *J. Mol. Biol.*, **174**, 663–695.
- Duarte, C.M. and Pyle, A.M. (1998) Stepping through an RNA structure: a novel approach to conformational analysis. *J. Mol. Biol.*, **284**, 1465–1478.
- Ng, H.L. and Dickerson, R.E. (2002) Mediation of the A/B-DNA helix transition by G-tracts in the crystal structure of duplex CATGGGCCCATG. *Nucleic Acids Res.*, **30**, 4061–4067.
- Beckers, M.L.M. and Buydens, L.M.C. (1998) Multivariate analysis of a data matrix containing A-DNA and B-DNA dinucleoside monophosphate steps: multidimensional Ramachandran plots for nucleic acids. *J. Comp. Chem.*, **19**, 695–715.
- Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.H., Srinivasan, A.R. and Schneider, B. (1992) The nucleic acid database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
- Altona, C. and Sundaralingam, M. (1972) Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J. Am. Chem. Soc.*, **94**, 8205–8212.
- Torgerson, W.S. (1952) Multidimensional scaling. 1. Theory and method. *Psychometrika*, **17**, 401–419.
- Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Comput. Graphical Statist.*, **5**, 299–314.
- Duda, R.O. and Hart, P.E. (1973) *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, NY.
- Reijmers, T.H., Wehrens, R. and Buydens, L.M.C. (2001) The influence of different structure representations on the clustering of an RNA nucleotides data set. *J. Chem. Inf. Comput. Sci.*, **41**, 1388–1394.
- Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) Data clustering: a review. *ACM Comput. Surv.*, **31**, 264–323.
- Reijmers, T.H., Wehrens, R. and Buydens, L.M.C. (2001) Circular effects in representations of an RNA nucleotides data set in relation with principal components analysis. *Chemometr. Intell. Lab.*, **56**, 61–71.
- Kearsley, S.K. (1989) On the orthogonal transformation used for structural comparisons. *Acta Crystallogr. A*, **45**, 208–210.
- Cattell, R.B. (1966) The scree test for the number of factors. *Multivar. Behav. Res.*, **1**, 245–276.
- Dijkstra, E.W. (1959) A note on two problems in connection with graphs. *Numer. Math.*, **1**, 269–271.
- Eppstein, D. (1998) Finding the *k* shortest paths. *SIAM J. Comput.*, **28**, 652–673.
- Schneider, B., Neidle, S. and Berman, H.M. (1997) Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers*, **42**, 113–124.
- Tereshko, V., Wilds, C.J., Minasov, G., Prakash, T.P., Maier, M.A., Howard, A., Wawrzak, Z., Manoharan, M. and Egli, M. (2001) Detection of alkali metal ions in DNA crystals using state-of-the-art X-ray diffraction experiments. *Nucleic Acids Res.*, **29**, 1208–1215.
- Antson, A.A., Dodson, E.J., Dodson, G.G., Greaves, R.B., Chen, X.-P. and Gollnick, P. (1999) Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature*, **401**, 235–242.
- Mack, D.R., Chiu, T.K. and Dickerson, R.E. (2001) Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: a comparative analysis of T-A and A-T steps within A-tracts. *J. Mol. Biol.*, **312**, 1037–1049.
- Xiong, Y., Deng, J., Sudarsanakumar, C. and Sundaralingam, M. (2001) Crystal structure of an RNA duplex r(gugucgcac)₂ with uridine bulges. *J. Mol. Biol.*, **313**, 573–582.
- Macgregor, R.B., Jr and Chen, M.Y. (1990) Delta-VBARO of the Na⁺-induced B-Z transition of poly[d(GC)] is positive. *Biopolymers*, **29**, 1069–1076.
- Krzyzaniak, A., Salanski, P., Jurczak, J. and Barciszewski, J. (1991) B-Z DNA reversible conformation changes effected by high pressure. *FEBS Lett.*, **279**, 1–4.
- Krzyzaniak, A., Barciszewski, J., Furst, J.P., Bald, R., Erdmann, V.A., Salanski, P. and Jurczak, J. (1994) A-Z-RNA conformational changes effected by high-pressure. *Int. J. Biol. Macromol.*, **16**, 158–162.
- Macgregor, R.B. (1998) Effect of hydrostatic, pressure on nucleic acids. *Biopolymers*, **48**, 253–263.