ELSEVIER

Biophysical Chemistry 50 (1994) 33–45

Biophysical Chemistry

Modelling DNA conformational mechanics

Richard Lavery, Brigitte Hartmann

Laboratoire de Biochimie Théorique, CNRS Institut de Biologie Physico-Chimique 13, Rue Pierre et Marie Curie, Paris 75005, France

(Received 5 January 1994)

Abstract

Using a modelling technique specifically adapted to DNA helices, we have systematically studied the impact of base sequence on the geometry of the double helix. The results obtained show that each repetitive base sequence leads to several stable conformations belonging to the B-DNA family. These conformational sub-states generally have similar stabilities, but often differ considerably in terms of their helical and backbone parameters. Each sub-state can be characterised by the puckering of its sugar rings. Surface energy mapping and combinatorial search techniques are used to further understand the DNA conformational hypersurface and to extend our study from dinucleotide repeats to tetranucleotide sequences. The resulting structural database should be useful for predicting the properties of longer and more irregular base sequences and thus should contribute to understanding how DNA target sites are recognised.

Key words: DNA; Base sequence effects; Molecular modelling; Energy mapping; Conformational transitions; Molecular recognition

1. Introduction

Understanding the process by which specific DNA binding sites are recognised by proteins, drugs, mutagens and other molecules represents a fundamental step towards understanding the biological activity of DNA. It is clear that such understanding must pass by a detailed knowledge of the structure and dynamics of DNA as a function of its base sequence. There is ample experimental evidence today to show that the base sequence indeed has an important impact on DNA structure and thus on the behaviour of DNA within the cell. X-ray and NMR studies of oligonucleotides point to important conformational heterogeneities as a function of sequence and nucleosome positioning analysis, gel migration and cyclisation studies have brought to light specific sequences which lead to static curvature or influence the overall flexibility of DNA [1,2]. Finally, DNA–protein interaction studies have also shown that so-called indirect read out of information, which relies upon local variations in the structure and dynamics of DNA, plays an important role in recognition processes [3,4].

Despite this progress, detailed knowledge of DNA structure is accumulating relatively slowly, on one hand due to difficulties in crystallising DNA oligomers and, on the other hand, due to the restrictions in obtaining full conformational data on helical nucleic acid fragments using NMR spectroscopy. Since understanding, and subse-

SSDI 0301-4622(94)00007-7

quently predicting, the structure of DNA for an arbitrary base sequence will undoubtedly require data on a very large number of test sequences, it seems worthwhile to consider whether molecular modelling can help in creating such a structural database. Over the last few years we have been working towards this goal, firstly, by designing a modelling algorithm specifically adapted to studying helical nucleic acids and, secondly, by developing reliable techniques for locating stable conformations and then analysing the features which characterise these states. The present article summarises progress made to date and also discusses future directions of research.

Successful modelling of DNA fine structure will certainly be useful for the fundamental task of deconvoluting base sequence effects. This knowledge will, in turn, contribute to understanding some of the most selective and subtle of biological recognition processes. Modelling is already helping experimentalists to translate raw data into macromolecular structure and, as detailed structural data accumulates, it should be possible to continue to refine the quality and applicability of modelling techniques.

2. Modelling and analysing nucleic acid conformation

In attempting to model nucleic acid fine structure, we began with the decision to build a simulation algorithm which would explicitly take into account the stereochemistry of the nucleic acids and their common helical structure. We also chose to simplify our representation in order to reduce as far as possible the number of variables necessary for modelling a given fragment of DNA. These choices led to a number of different modelling techniques [5–7] which evolved towards the Jumna (Junction minimisation of nucleic acids) algorithm we use today [8]. This approach combines the direct use of helicoidal coordinates for positioning individual nucleotides with respect to a common helical axis system and internal coordinates (dihedral angles and, within the sugar rings, valence angles) for describing the internal flexibility of each nucleotide. Junctions between successive 3′-monophosphate nucleotides are closed using harmonic distance and valence angle constraints. In this way, we achieve a ten-fold reduction in the number of degrees of freedom representing the system compared to a classical molecular mechanics calculation in Cartesian coordinates. Although we necessarily loose some fine detail with our model, this is offset by improved possibilities for searching conformational space.

Jumna has another important advantage, namely, the possibility of controlling chosen features of nucleic acid conformation during the simulation. The first, and most important aspect, of this control is the possibility of introducing helical symmetry by simply grouping together sets of helically equivalent variables. This can lead to a further order of magnitude reduction in the number of variables representing the system and, as we will see below, opens the way for treating infinite regular polymers very efficiently. The second aspect of control is the possibility of introducing a wide variety of constraints (inter-atomic distances, dihedral angles, helical parameters, sugar puckers, axis curvature, etc.) which can be used to fit structures to available experimental data, to make energy maps and to induce conformational transitions. Lastly, since the stereochemistry of the nucleic acids is explicitly included in the program, it is possible to build standard helical conformations from a pre-existing nucleotide library by simply specifying the desired base sequence. This approach also simplifies the construction and optimisation of triple helices, quadruplexes, irregular structures and complexes between nucleic acids and other molecules.

The force field FLEX used within Jumna [6,7] is based on Lennard-Jones parameters [9] and atomic charges [10] developed specifically for representing nucleic acids. Quantum chemical calculations were used to obtain angle-dependent hydrogen bonding parameters and also to estimate the most of the dihedral barriers used. Careful parameterisation of the sugar rings was made to ensure a good representation of repuckering and this included the introduction of anomeric dihedral terms (also employed for the representation of phosphate group flexibility). One change has

recently been made to the dihedral term for the thymine C5-methyl rotation, which we tested in the course of recent quantum chemical calculations concerning the strength of C–H...O interactions. Originally, this rotation was assigned a six-fold potential, implying that the methyl protons would be equally repelled by the O4 and HC6 atoms on either side of C5. Ab initio quantum calculations with a 3-21G basis however showed that the only significant repulsion comes from the larger O4 atom. This leads to a three-fold rotation potential and thus to stable rotamers where one proton lies in the plane of the base pointing towards H5, rather than perpendicular to the base plane as before. The result of this minor adjustment was to remove artifactual splitting of certain minima into states differing only in the position of the out-of-plane thymine proton, which pointed either upwards or downwards and led to slightly different base–base interactions.

Concerning the representation of the environment surrounding the nucleic acids studied, we currently treat the electrostatic effects of the water and counterions in an approximate way. Introducing explicit water shells around the macromolecule can be done during the course of dynamic simulations, but this choice increases the cost of the simulations so radically that it would no longer be possible to carry out the large number of studies required to understand base sequence effects. Our present treatment of solvent is thus limited to a simple sigmoidal distance dependence of the dielectric constant, based on model calculations by Hingerty and co-workers [11] and reparameterised by us to allow for adjustable slope and asymptote values [8]. In addition, we reduce the net charge on each phosphate group to $-0.5\ e$ to mimic the effects of counterion binding. Within helical DNA, this crude technique leads to good results, as our own work and recent dynamic simulations have shown [12]. However, this level of approximation is not appropriate either for more irregular nucleic acid structures or for nucleic acid complexes where desolvation and counterion reorganisation effects cannot be ignored. In attempt to treat such problems, we are currently working on a rapid numerical solution of the Poisson–Boltzmann model which would allow the inclusion of a much more refined electrostatic treatment.

Finally, it is remarked that once simulations have been performed, it is very important to be able to precisely analyse the conformations obtained. This is particularly important in the case of studying sequence effects, where it is necessary to quantify small deformations within short fragments of the double helix. We have developed the Curves algorithm for this purpose [13,14]. Curves determines the optimal (and possibly curved) helical axis describing any nucleic acid fragment and also calculates a complete set of helical and backbone parameters. The helical parameters generated by Curves, and used in the present article, obey the Cambridge convention [15].

3. Characterising conformational sub-states

In order to understand base sequence effects on DNA it seemed necessary to take as systematic an approach as possible. Most importantly, this involves studying a wide range of base sequences and also ensuring that the energy minimisations performed lead to the global minimum associated with each sequence. We decided from the beginning of our studies to work on infinite, regular polymers so that the complicating end-effects of oligonucleotides could be eliminated. This is easily achieved using Jumna, firstly, by imposing helical symmetry and, secondly, by optimising the energy per unit cell within the polymer in question, rather than optimising the total energy of a finite length fragment. This choice also has the advantage of reducing the time necessary for each energy calculation (as it is only necessary to calculate the internal conformational energy of one unit cell and its interactions with a set of neighbouring cells).

Our initial studies [16] involved a set of six regular polymers having base sequences containing all the possible dinucleotide steps which can be built from the four standard DNA bases. These polymers are shown below, followed by the dinucleotide steps that they contain. Note that al-

though 16 dinucleotide sequences can be formed from the 4 bases, there are only 10 unique double stranded dinucleotide steps, since the dinucleotides indicated by stars are simply the paired strands of the unstarred dinucleotides on the same line.

$(AA)_n$	AA	TT*		
$(GG)_n$	GG	CC*		
$(CG)_n$	CG	GC		
$(TA)_n$	TA	AT		
$(CA)_n$	CA	AC	TG*	GT*
$(GA)_n$	GA	AG	TC*	CT*

Following the base sequences, these polymers were constrained to obey mononucleotide (AA..., GG...) or dinucleotide (CG..., TA..., TG..., GA...) symmetry. In addition, when appropriate (CG..., TA...) interstrand dyad symmetry was imposed (homonomous constraints). Early calculations were made taking into account the interactions of 7 neighbouring nucleotide pairs with the central unit cell, but more recently we have been able to increase this number to 10 pairs.

Looking for the stable minima of these polymers involved a very large number of minimisations, using a wide variety of starting points. The first calculations were made with two sets of fibre coordinates for B-DNA [17,18] [#1]. The results obtained for each base sequence were then used as new starting points for all the other polymers. This new round of minimisations either led to stable conformations that had already been found or to new minima. In the latter case, the new conformations were again used a starting points for all the remaining polymers. (Using minima obtained with one sequence as starting points for other sequences is facilitated in Jumna by the fact that the output conformation file contains all necessary helicoidal and backbone parameters, but does not contain the base sequence.)

[#1] See also coordinates communicated to our laboratory by S. Arnott.

Once this series of calculations had been completed a second technique was used to verify the stability of the minima found. This involved profiting from the constraint possibilities available in Jumna to perturb the minimal energy conformations found and thus to show that each lay within a clearly defined energy well. We used two types of perturbation for this purpose, overwinding–underwinding and stretching–compression, achieved by respectively constraining the total twist or the total rise of the polymer in question. These calculations lead to a clearly defined set of energy minima and also enabled several new minima to be located. It was also noted that the perturbations imposed on the double helix could, in certain cases, change the relative stabilities of the known minima and could also provoke transitions between different minima [19].

More recently it has been possible to automate this approach to finding stable conformational states using techniques which will be described below. These improvements, as well as the minor modifications introduced since our earlier work (in particular, changes to the thymine methyl rotation and an increase in the number of neighbours interacting with the central unit cell), have led to some changes in the number of minima detected and in the detailed conformation of certain minima (Table 1). Notably, we have detected a number of conformations containing purine nucleotides with O1'-endo sugar puckers (indicated by the symbol # in Table 1). These minima are generally the least stable of those belonging to the B-family, although there is one exception to this rule for the GA polymer. We also discovered several new minima for the AA and GG homopolymers, once mononucleotide symmetry was relaxed (indicated by the symbol * in Table 1). In the case of poly(dA).poly(dT), even the most stable adopts a dinucleotide repeat as has been discussed elsewhere [20].

The first important result of this study was that each of the sequences investigated has a number of distinct minima with different conformations, but very similar energies. The range of conformations covered can be judged in Table 2, which lists the extreme values adopted by each of the helical and backbone parameters. These re-

Table 1
Energies per unit cell for the stable sub-state conformations of the six mono- and dinucleotide polymers studied (values in kcal/mol) [a]

Sequence	1	2	3	4	5	6
$(AA)_n$	−82.5 *	−82.4 *	−81.9 *			
$(GG)_n$	−118.1	−115.2	−115.1 *	−114.2 *		
$(CG)_n$	−119.4	−118.7	−118.0	−114.0 #		
$(TA)_n$	−81.6	−80.8	−80.2	−80.0		
$(CA)_n$	−100.9	−100.8	−100.6	−100.4	−100.3	−100.2
	−99.6	−99.6	−99.2	−97.2 #	−97.1 #	−95.2 #
	−95.0 #					
$(GA)_n$	−99.5	−98.4 #	−98.3	−97.6	−97.0	−96.7 #

[a] Conformations containing purine nucleotides with O1′-endo sugar puckers are indicated by the symbol #. Homopolymer conformations which in fact display dinucleotide symmetry are indicated by the symbol *.

sults concur with recent crystallographic and NMR studies in confirming that the term B-DNA actually represents a rather large volume of conformational space. It will be noted that a number of backbone parameters vary by 20°–50° and that the sugar phase angle covers a remarkable 100° range. Equally, several helical parameters are very variable, twist covering a 17° interval and propeller and buckle roughly 30°, while the translational parameters rise and Xdisp vary respectively by 1.2 and 3.3 Å. Despite these variations, the mean values resulting from our simulations and from experimental data are very similar. (It is remarked that the mean helicoidal values coming from high-resolution oligomer crystals are somewhat weighted towards positive Xdisp due to the relatively common occurrence of B_{II} conformations [21] and to negative propellers by an over representation of AT base pairs).

It should however be stressed that important changes in conformation occur, not only between polymers with different base sequences, but also between the conformational sub-states of each given polymer. This can most easily be seen from

Table 2
Backbone and helical parameter ranges for the B-DNA sub-states compared to experimental X-ray conformations

Parameter	Jumna simulations			Experiment	
	minimum	maximum	mean	B fibre [18]	decamers [a]
α	−72	−58	−65	−41	−64
β	159	188	175	135	166
γ	50	67	58	37	50
δ	91	150	133	139	132
ϵ	−177	−161	−172	−134	−179
ζ	−145	−90	−110	−102	−94
χ	−149	−94	−118	−102	−104
phase	74	184	150	154	150
amplitude	28	46	38	37	39
Xdisp	−3.5	−0.2	−1.6	0.0	0.5
inclin	−18.0	12.0	−0.3	1.5	−1.2
propeller	−22.0	4.0	−6.4	−13.3	−11.0
buckle	−8.0	21.0	−1.8	0.0	0.9
rise	2.8	4.0	3.3	3.4	3.4
twist	28.0	45.0	35.5	36.0	36.2

[a] The experimental results for decamers refer to 8 crystal structures resolved to better than 2 Å. Backbone steps involving $\alpha\gamma$ crankshaft or B_{II} backbone conformations were excluded from the analysis [28].

Table 3
Parameter ranges for the stable sub-states of each polymer

	Xdisp		Inclin		Propeller		Buckle		Rise		Twist	
AA	−1.6	−0.7	4	12	−22	−17	13	21	2.8	3.4	33	41
GG	−1.7	−0.2	−7	−1	−19	−12	5	14	2.8	4.0	35	40
CG	−2.9	−1.5	−7	1	−3	4	−7	6	3.1	3.9	29	42
TA	−3.5	−1.7	0	6	−7	−2	−6	6	3.0	3.7	29	43
CA	−2.0	−0.9	−4	9	−18	1	−8	7	2.9	3.8	31	45
GA	−2.4	−1.0	−12	6	−12	1	−4	16	2.9	3.9	28	42
	α		β		γ		δ		ϵ		ζ	
AA	−72	−58	162	179	55	63	93	147	−177	−165	−139	−92
GG	−67	−59	159	181	54	67	92	146	−177	−164	−141	−94
CG	−71	−64	173	181	54	60	92	149	−176	−166	−134	−90
TA	−71	−63	171	184	51	59	96	146	−175	−169	−131	−91
CA	−72	−60	162	188	52	61	91	150	−177	−161	−144	−90
GA	−72	−58	164	187	50	63	92	148	−176	−163	−143	−91
	χ		phase		amplitude							
AA	−140	−94	81	183	32	42						
GG	−144	−104	83	180	31	45						
CG	−147	−111	90	180	34	44						
TA	−144	−110	85	178	35	46						
CA	−146	−100	74	184	28	46						
GA	−149	−100	81	180	32	45						

the parameters ranges for each polymer given in Table 3. This data shows that a given base sequence is only rarely associated with distinct values of either helical or backbone parameters, if all of its possible energy minima are taken into account. This appears to imply that all sub-states cannot be significantly populated at room temperature, since, if this were the case, base sequence would have little residual impact on the double helix.

In looking for a way to characterise the conformational sub-states, we noted that the sub-states of each polymer always differed in terms of their sugar puckers. The sugars were found to lie in three groups: C2′-endo with low amplitude (termed 'S'), C2′-endo with high amplitude (termed 'X') and O1′-endo (termed 'E'). In the case of the purines, the S and X groups were also separated by a decrease in phase, but this was less marked for the pyrimidines. It was subse-

Table 4
Sub-state conformations classified according to their sugar puckering (S: low amplitude C2′-endo, X: high amplitude C2′-endo, E: O1′-endo). The unique sugars in each strand are listed in the 5′ → 3′ sense and the order of the bases in the first strand is as shown in the sequence column [a]

Sequence	1	2	3	4	5	6
$(AA)_n$	SX:SS *	SX:ES *	SX:SE *			
$(GG)_n$	XX:XX	SS:XX	XS:SX *	SX:ES *		
$(CG)_n$	SX:SX	XS:XS	ES:ES	SE:SE #		
$(TA)_n$	SX:SX	XS:XS	ES:ES	EX:EX		
$(CA)_n$	ES:XS	XS:SX	XS:XS	ES:SX	SX:XS	ES:ES
	XS:ES	SX:SX	SX:ES	SE:XS #	SE:XX #	SE:SE #
	SE:ES #					
$(GA)_n$	XS:SX	ES:SX #	SX:ES	XS:SE	SX:SE	ES:SE #

[a] Conformations containing purine nucleotides with O1′-endo sugar puckers are indicated by the symbol #. Homopolymer conformations which in fact display dinucleotide symmetry are indicated by the symbol *.

quently found that each sub-state could be uniquely identified by a label describing its sugar pucker groups (Table 4). In consequence, it appears that the sugars play an important role in defining the local conformation of DNA. Although base–base interactions, and notably base stacking, are necessarily at the origin of local deformations in helical conformation, these interactions seem to be modulated by the fact that the sugars have a limited set of preferred puckers and that these puckers also control local conformation by strong coupling to other helical and backbone parameters.

4. Energy surface mapping

Since sugar puckers have been found to be sufficient for characterising the conformational sub-states we located, it seemed worthwhile to investigate the energy surface defined by these variables [22]. This approach was designed to test our view that the sugars indeed play a fundamental role in determining the local conformation of the double helix. Jumna was consequently modified to make 1D and 2D energy maps in terms of sugar phase angles, all other variables being relaxed by an energy minimisation at each fixed value of phase.

Since both the mononucleotide conformations of the homopolymers poly(dT).poly(dA) and poly(dC).poly(dG) and the dinucleotide conformations of the alternating polymers with dyad symmetry poly(dCG).poly(dCG) and poly(dTA).poly(dTA) each have only two symmetry distinct sugars, it was possible to search all the sugar conformations of these polymers with individual 2D maps. One such map is shown in Fig. 1, covering the B-family conformations of the CG alternating polymer. The results of this study led to several important conclusions. Firstly, the maps obtained were independent of the starting conformation used and maps built up from several independently calculated zones could be fitted together smoothly. Fig. 2 shows an example of this for the CG alternating polymer, where four separate maps have been put together to form a continuous energy surface covering the full range

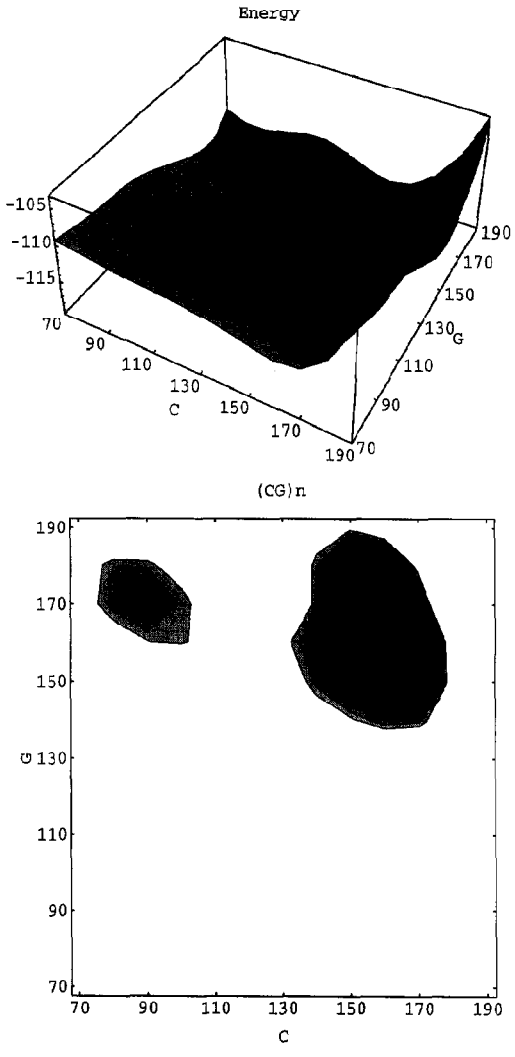

Fig. 1. Sugar phase energy map of the B-family domain of poly(dCG).poly(dCG).

of sugar puckers from C2′-endo to C3′-endo. Secondly, all the energy minima which had been found for the polymers studied were also present within the sugar maps. Both these findings indicate that, for a given base sequence, once sugar puckers are fixed, the conformation of the double helix is also fixed. Sugars therefore do play a central role in determining DNA helical conformation.

With the help of these maps it is also possible to determine the energy barriers separating the various sub-states of each polymer. In general, it

is found that the S and X puckering domain minima are separated by very small barriers amounting to only a few tenths of a kcal/mol. In contrast, sub-states containing E (O1'-endo) sugars are generally isolated by larger barriers of the order of 1–2 kcal/mol. Similar barriers also oppose the passage from the E to the N domain (C3'-endo) as seen in Fig. 2, which also illustrates the optimal pathway for passing between the B (top left) and A (bottom right) forms of the double helix. It should be recalled that since all

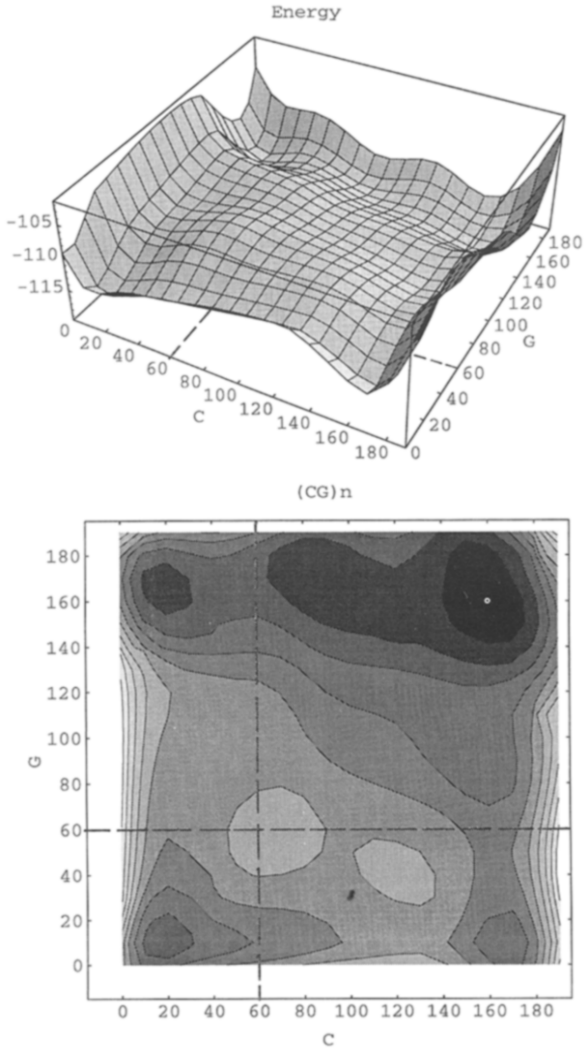

Fig. 2. Sugar phase energy map for the full A/B domain of poly(dCG).poly(dCG). The dotted lines indicate the junctions between separately calculated partial maps.

the energy minima seen on the maps correspond to conformations which differ by their sugar puckers, the energy barriers defined by the maps (where all other variables are optimised at each point) are also the lowest energy pathways for passing between the different sub-state minima.

The map shown in Fig. 2 contains the results of roughly 600 energy minimisations. Consequently, making sugar maps in more than 2D would become prohibitively expensive. It is therefore generally impossible to use this technique for polymers with more than two symmetry distinct sugars. However, an extension of mapping can be made in special cases, since it was noted in our early studies [16] that successive sugars within each strand of the double helix tend to have alternating phase angles. This implies that, rather than looking directly at phase angles themselves, we should also be able to map phase angle differences (Δphase) between successive nucleotides. In the case of the CG alternating polymer, constraining Δphase is equivalent to moving along the leading diagonal of the map in Fig. 1, optimising the conformation along lines parallel to the trailing diagonal at each value of Δphase. The result of this procedure is to move along the energy valleys of the 2D map, thus creating a 1D map. This map nevertheless still contains information on all the energy minima and defines the barriers between them. This being so, it becomes possible to treat alternating polymers with four symmetry distinct sugars, poly(dCA).poly(dTG) and poly(dGA).poly(dTC), as 2D maps, which fix the Δphase value within each strand. An example of this technique is shown in Fig. 3 for the GA alternating polymer. This map is naturally more complex, but it indeed contains all the sub-states listed in Table 1 (the most stable state occurs in the centre of the map with the second state at the top right-hand corner, the third state lies centre left, and the three least stable states fall along the bottom edge of the map).

Sugar maps can help us to take one further step towards looking at more complicated sequences and, notably, certain examples of the tetranucleotide sequences discussed in the following section. If we consider tetranucleotide repeats which contain central dyad symmetry axes, such

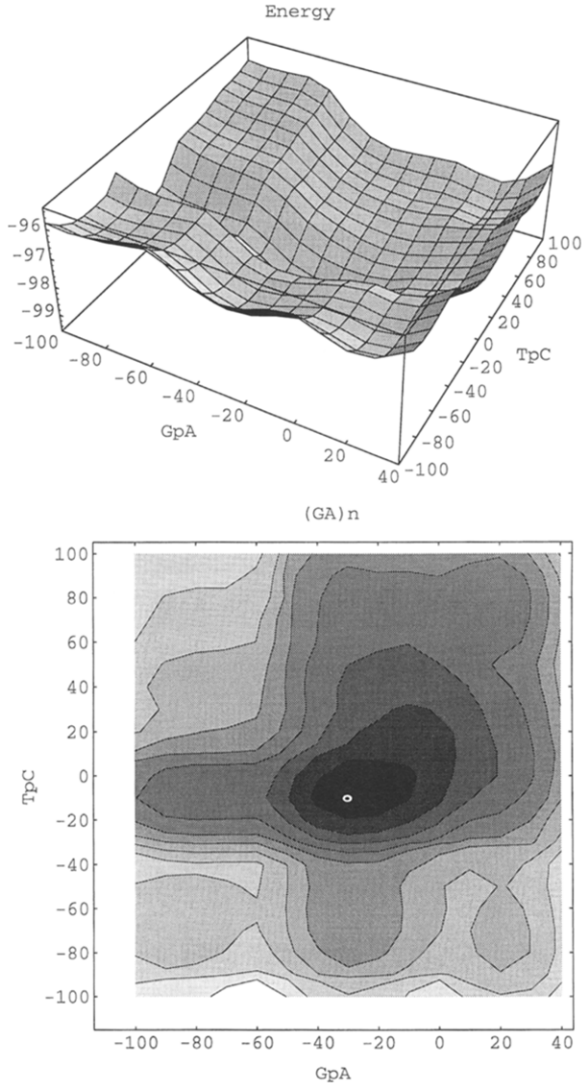

Fig. 3. 2D Δphase energy map for poly(dGA).poly(dTC).

as the sequence ACGT, these sequences will also, in principle, contain only four symmetry distinct sugars, as for the alternating dinucleotide sequences discussed above. However, since the four unique sugars now all occur within one strand, they are associated with three independent Δphase values (for example the dinucleotide steps TpA, ApC and CpG in the ACGT tetranucleotide chosen. Note that the Δphase of the fourth step GpT is uniquely determined by the other three values). It is therefore not possible to cover all possible sugar conformations with a single map, but this can be done by a set of 2D Δphase maps each calculated for a fixed value of the third Δphase value. This rather lengthy procedure has been applied to the ACGT polymer (with tetranucleotide, homonomous symmetry constraints) and leads to a total of seven stable

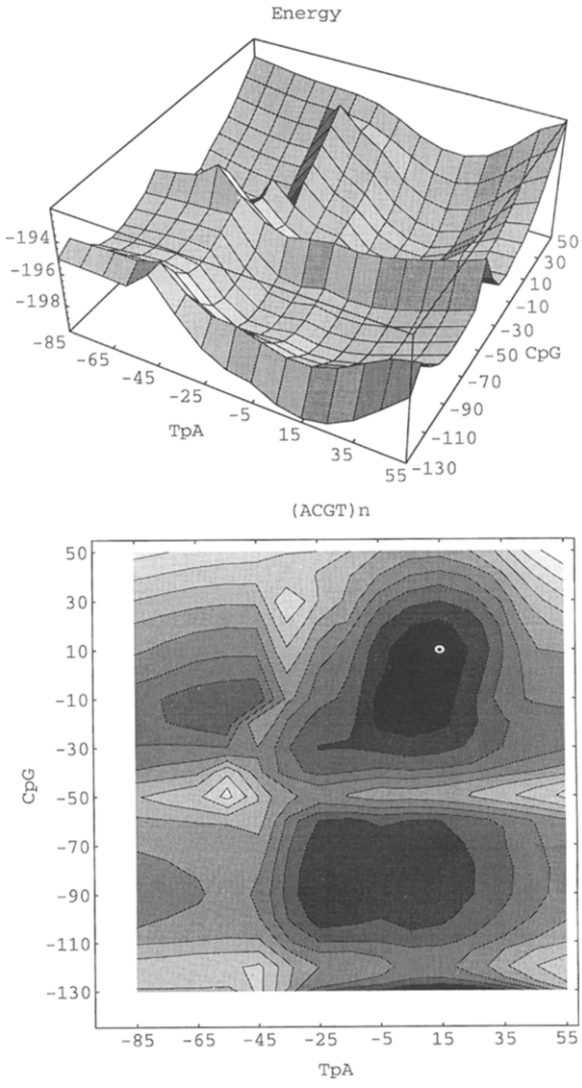

Fig. 4. 2D Δphase map for the TpA and CpG steps of the $(ACGT)_n$ tetranucleotide repeat polymer, the ApC step being left free to vary. Moving clockwise from the top left, the successive quadrants of the map contain respectively 3 (S,X domain), 2 (E sugars for C), 1 (E sugars for T and C) and 1 (E sugar for T) minima.

sub-states. In fact, with this polymer, the same result can surprisingly be obtained using a single map, plotting TpA versus CpG Δphase angles and leaving the ApC Δphase free to vary (Fig. 4).

5. Combinatorial searches

All the sub-state conformations of DNA detected so far have sugars belonging to the S, X, E or N families and can be uniquely characterised by their sugars. It therefore seems reasonable to suppose that if all possible combinations of sugar conformations are used as starting points for energy minimisation, it will be possible to locate all of the stable sub-states for any given sequence. This simple idea was put into practice by modifying Jumna to read a chosen set of possible puckers for each nucleotide within a given polymer and then to cycle automatically through full combinatorial set of puckers. Tests of this procedure for the dinucleotide repeat polymers discussed above confirmed that it is an easy and reliable way of finding sub-states. Its application to the tetranucleotide sequence ACGT also led to the full set of sub-states detected by energy mapping. However, these calculations, which did not impose homonomous constraints and thus allowed for eight symmetry distinct sugars, also revealed a further set of stable sub-states which no longer obey the dyad symmetry of the base sequence (Table 5). 15 such states were located, which, added to the 8 states with dyad symmetry, makes a total of 23 minima for this polymer (it should be remarked that the combinatorial search used did not allow for E sugars within purine nucleotides and thus there are probably several further minima to be located). As with the dinucleotide sequences, many of the sub-states found for $(ACGT)_n$ have virtually identical energies, but their conformations typically differ by backbone RMS values of roughly 10° (for sub-states containing only S and X sugars) and roughly 20° (for conformations containing any E sugars).

Automated combinatorial searches have thus enabled us to begin studying a more sophisticated model of sequence effects. We began, as described earlier, by looking at the 10 unique dinucleotide sequences. Since base stacking is at the origin of sequence effects upon the double helix, the simplest (nearest-neighbour) model reasonably assumes that any given sequence can be considered as formed from a series of mutually independent dinucleotide steps. This model underlies all current attempts to predict sequence-dependent helix curvature ([23–25], and references therein). Although such models have had successes, it has already become clear, both from experimental oligonucleotide structures and from modelling studies, that the bases which flank each dinucleotide step can have a non-negligible effect upon its conformation. This suggests that a next-nearest-neighbour model, taking into account such effects would be more reliable [26]. This is equivalent to saying that a long sequence can be built up from a series of overlapping tetranu-

Table 5
Sub-state conformations of the tetranucleotide repeat polymer $(ACGT)_n$. The unique sugars in each strand are listed in the 5′ → 3′ sense for the tetranucleotide ACGT. Energies per unit cell given in kcal/mol [a]

No.	Sugar puckers	Energy
1	XSXS:XSXS	−199.5 *
2	XXSS:XXSS	−199.5 *
3	XSXS:XXSS	−199.5
4	XESS:XXSS	−199.2
5	XXSS:SEXX	−199.2
6	SEXX:SXSX	−199.0
7	SESX:SESX	−199.0 *
8	XSXS:SXSX	−198.9
9	XESS:XESS	−198.8 *
10	XXSS:XXSE	−198.7
11	SEXX:XESS	−198.7
12	XXSS:SXSX	−198.7
13	SXSX:SXSX	−198.4 *
14	XXSE:XXSE	−198.1 *
15	SESE:XXSS	−197.8
16	SESE:SESX	−197.5
17	SESE:SXSX	−197.4
18	SXSE:SXSX	−197.4
19	XESE:XSXS	−197.2
20	SESE:XESS	−196.9
21	XESS:XESE	−196.9
22	SESE:SESE	−196.5 *
23	SSXX:SESE	−196.3

[a] Conformations with dyad inversion symmetry between the two strands (as implied by the base sequence) are indicated by the symbol *.

cleotides. This model however represents a considerable increase in complexity since, while there are only 10 unique dinucleotide fragments, there are 136 unique tetranucleotides.

Using the same approach as in our earlier work, the 136 tetranucleotides can be studied with the help of 39 repeating sequence polymers. These polymers and the tetranucleotides they contain are shown in Table 6. Each polymer is constrained to tetranucleotide symmetry. No dyad constraints are imposed, even in the cases of sequences possessing dyad symmetry (see discussion of ACGT above). This means that each polymer contains a total of eight symmetry distinct sugars. The combinatorial study of these polymers is now underway and, although it is too early to draw the final conclusions, several interesting points have already come to light. Firstly, each polymer, and therefore the tetranucleotide sequences it contains, again exhibits a small number of energy sub-states (typically around 20) with quite widely differing conformations. Secondly, it is still possible to characterise these minima by their sugar pucker classes (S, X, E), although these classes are less tightly defined than in the case of the dinucleotide sequences. Thirdly, mono- and dinucleotide repeat sequences and also sequences with dyad symmetry (the first three sections of Table 6) commonly exhibit stable energy minima with lower symmetry than that implied by their base sequences. The symmetry reduction effect found with poly(dA).poly(dT) [20] and with the $(ACGT)_n$ polymer is thus a general phenomenon, although it should be added that the most stable sub-states of each tetranucleotide do, in general, reflect the symmetry of its base sequence. This finding nevertheless suggests that conformational asymmetries induced by neighbouring sequences, by thermal excitation, or by interactions with other molecules are probably common events.

It should be added that the number of sub-states found for each polymer is very much less than the total number of combinations which may be formed from the sugar pucker classes. If we consider that purines most commonly adopt either S or X puckers, while pyrimidines adopt S, X or E puckers, then the eight symmetry distinct sugars in a tetranucleotide can lead to a total of $2^4 \times 3^4 = 1296$ different combinations. This number is roughly 60 times bigger than the number of sub-states typically found for each tetranucleotide. Each base sequence therefore has a

Table 6
Polymers containing the 136 unique tetranucleotide sequences. The tetranucleotides contained within each polymer are listed to the right of its sequence

Tetranucleotides containing mononucleotide repeats				
$(GGGG)_n$	GGGG			
$(AAAA)_n$	AAAA			
Tetranucleotides containing dinucleotide repeats				
$(CGCG)_n$	CGCG	GCGC		
$(TATA)_n$	TATA	ATAT		
$(GTGT)_n$	GTGT	TGTG		
$(GAGA)_n$	GAGA	AGAG		
Tetranucleotides with inversion symmetry				
$(CCGG)_n$	CCGG	CGGC	GGCC	
$(TTAA)_n$	TTAA	TAAT	AATT	
$(TCGA)_n$	TCGA	CGAT	GATC	
$(TGCA)_n$	TGCA	ATGC	CATG	
$(ACGT)_n$	ACGT	CGTA	GTAC	
$(AGCT)_n$	AGCT	TAGC	CTAG	
Other tetranucleotides				
$(AGGT)_n$	AGGT	GGTA	GTAG	TAGG
$(AGGA)_n$	AGGA	GGAA	GAAG	AAGG
$(TGGT)_n$	TGGT	GGTT	CAAC	TTGG
$(GCGG)_n$	GCGG	CGGG	GGGC	GGCG
$(ACGA)_n$	ACGA	CGAA	GAAC	CGTT
$(AGCA)_n$	AGCA	TTGC	CAAG	AAGC
$(AGCG)_n$	AGCG	GCGA	CGAG	GAGC
$(ATAA)_n$	ATAA	TAAA	AAAT	AATA
$(ATAC)_n$	ATAC	TGTA	ATGT	CATA
$(GATG)_n$	GATG	ATGG	TGGA	GGAT
$(GATA)_n$	GATA	ATAG	TAGA	AGAT
$(GGAG)_n$	GGAG	GAGG	AGGG	GGGA
$(AGAA)_n$	AGAA	GAAA	AAAG	AAGA
$(CGAC)_n$	CGAC	GGTC	CGGT	CCGA
$(TGAT)_n$	TGAT	AATC	CAAT	TTGA
$(CAGC)_n$	CAGC	AGCC	TGGC	CTGG
$(TAGT)_n$	TAGT	AGTT	TAAC	CTAA
$(CAGA)_n$	CAGA	AGAC	TGTC	CTGT
$(CAGT)_n$	CAGT	AGTC	TGAC	CTGA
$(GGTG)_n$	GGTG	GTGG	TGGG	GGGT
$(AGTA)_n$	AGTA	GTAA	TAAG	AAGT
$(CGTC)_n$	CGTC	GGAC	CGGA	ACGG
$(TGTT)_n$	TGTT	AAAC	CAAA	TTGT
$(CTGC)_n$	CTGC	GGCA	AGGC	CAGG
$(ATGA)_n$	ATGA	TGAA	GAAT	AATG
$(GTGC)_n$	GTGC	CGCA	ACGC	CGTG
$(GTGA)_n$	GTGA	TGAG	GAGT	AGTG

strong discriminatory role in selecting only a small number of pucker combinations, which, in turn, translate this sequence into the specific three-dimensional properties of the corresponding double helix.

6. Conclusions and future prospects

The work carried out so far has shown that most repetitive DNA sequences seem to lead to a small number of sub-states with different conformations, but similar stabilities. It has also been found these sub-states can be best characterised by their sugar puckers, which consequently seem to play a central role in fixing local helical structure. Recent developments in rapidly searching for the energy minima of more complex polymers suggest that it should be feasible to build up a structural database of all tetranucleotide sequences and we may hope that this database will, in turn, open the route for predicting the conformations of longer sequences.

In parallel with these studies we continue attempts to improve the quality of our simulations, notably through the introduction of a more refined solvent and counterion treatment and also by developing methods to include dynamic aspects of DNA conformation. This will be important both for understanding how sub-states will be populated at ambient temperatures and for determining how far base sequence influences the local flexibility of the double helix.

In conclusion, understanding the behaviour of DNA as a function of its sequence will be an important step towards understanding the way DNA targets are recognised within the cell and, thus, a fundamental aspect of genetic control. It should also enhance the possibilities of engineering synthetic molecules which will be able to modify gene activity [27]. Lastly, confronted with the exponentially growing amounts of sequence data from the Human Genome Project, reliable modelling will offer a way of converting one-dimensional sequences into a combination of spatial and temporal data which will hopefully improve our ability to read and profit from this mass of information.

Acknowledgement

We wish to thank the Association for International Cancer Research (St. Andrews University, UK) for their generous support of this research.

References

[1] E.N. Trifonov, CRC Crit. Rev. Biochem. 19 (1989) 89.
[2] E.N. Trifonov, in: Theoretical biochemistry and molecular biophysics, Vol. 1. DNA (Adenine Press, New York, 1991) p. 377.
[3] D. Suck, in: Structural tools for the analysis of protein–nucleic acid complexes, eds. D.M.J. Lilley, H. Heumann and D. Suck (Birkhauser, Basel, 1992) p. 127.
[4] A.A. Travers, Curr. Opin. Struct. Biol. 2 (1992) 71.
[5] H. Sklenar, R. Lavery and B. Pullman, J. Biomol. Struct. Dyn. 3 (1986) 967.
[6] R. Lavery, H. Sklenar, K. Zakrzewska and B. Pullman, J. Biomol. Struct. Dyn. 3 (1986) 989.
[7] R. Lavery, I. Parker and J. Kendrick, J. Biomol. Struct. Dyn. 4 (1986) 443.
[8] R. Lavery, in: Structure and expression, Vol. 3. DNA bending and curvature, eds. W.K. Olson, R.H. Sarma, M.H. Sarma and M. Sundaralingam (Adenine Press, New York, 1988) p. 191.
[9] V.B. Zhurkin, V.I. Poltev and V.L. Florent'ev, Mol. Biol. 14 (1980) 1116.
[10] R. Lavery, K. Zakrzewska and A. Pullman, J. Comput. Chem. 5 (1984) 363.
[11] B. Hingerty, R.H. Richie, T.L. Ferrel and T.E. Turner, Biopolymers 24 (1985) 427.
[12] V. Fritsch and E. Westhof, J. Am. Chem. Soc. 113 (1991) 8271.
[13] R. Lavery and H. Sklenar, J. Biomol. Struct. Dyn. 6 (1989) 655.
[14] R. Lavery and H. Sklenar, in: Structure and methods, Vol. 2. DNA protein complexes and proteins, eds. R.H. Sarma and M.H. Sarma (Adenine Press, New York, 1990) p. 215.
[15] R.E. Dickerson, M. Bansal, C.R. Calladine, S. Diekmann, W.N. Hunter, O. Kennard, R. Lavery, H.C.M. Nelson, W.K. Olson, W. Saenger, Z. Shakked, H. Sklenar, D.M. Soumpasis, C.-S. Tung, E. von Kitzing, A.H.-J. Wang and V.B. Zhurkin, J. Mol. Biol. 205 (1989) 787.
[16] M. Poncin, B. Hartmann and R. Lavery, J. Mol. Biol. 226 (1992) 775.
[17] S. Arnott and D.W.L. Hukins, J. Mol. Biol. 81 (1973) 93.
[18] S. Arnott, R. Chandrasekaran, D.L. Birdsall, A.G.W. Leslie and R.L. Ratliff, Nature 283 (1980) 743.
[19] M. Poncin, D. Piazzola and R. Lavery, Biopolymers 32 (1992) 1077.
[20] K. Zakrzewska, V.I. Poltev, C. Oguey and R. Lavery, J. Mol. Struct. THEOCHEM 286 (1993) 219.
[21] B. Hartmann, D. Piazzola and R. Lavery, Nucl. Ac. Res. 21 (1993) 561.

[22] R. Lavery, in: Advances in computational biology, Vol. 1, ed. H.O. Villar (JAI Press, Connecticut, 1994), in press.
[23] A. Bolshoy, P. McNamara, R.E. Harrington and E.N. Trifonov, Proc. Natl. Acad. Sci. USA 88 (1991) 2312.
[24] S. Cacchione, P. De Santis, D. Foti, A. Palleschi and M. Savino, Biochemistry 28 (1989) 8706.
[25] R.K.-Z. Tan and S.C. Harvey J. Biomol. Struct. Dyn. 5 (1987) 497.
[26] K. Yanagi, G.G. Privé and R.E. Dickerson, J. Mol. Biol. 217 (1991) 201.
[27] C. Hélène and J.J. Toulmé, Biochim. Biophys. Acta 1049 (1990) 99.
[28] A. Lipanov, M.L. Kopka, M. Kaczor-Grzeskowiak, J. Quintana and R.E. Dickerson, Biochemistry 32 (1993) 1373.