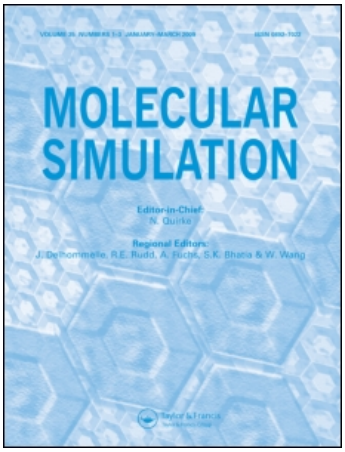

Molecular Simulation

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Full-strand Minimization of Single and Double Stranded B-DNA Using Monte-Carlo Annealing

Konstantinos Sfyrakis[a]; Astero Provata[b]; David C. Povey[a]; Brendan J. Howlin[a]
[a] Department of Chemistry, School of Physics and Chemistry, University of Surrey, Guildford, UK [b] Institute of Physical Chemistry, National Research Center "Demokritos", Athens, Greece

Online publication date: 26 October 2010

To cite this Article Sfyrakis, Konstantinos , Provata, Astero , Povey, David C. and Howlin, Brendan J.(2003) 'Full-strand Minimization of Single and Double Stranded B-DNA Using Monte-Carlo Annealing', Molecular Simulation, 29: 9, 555 — 575

To link to this Article: DOI: 10.1080/0892702021000008858

URL: http://dx.doi.org/10.1080/0892702021000008858

Molecular Simulation, VOL. 29 (9), September 2003, pp. 555–575

Taylor & Francis
Taylor & Francis Group

Full-strand Minimization of Single and Double Stranded B-DNA Using Monte-Carlo Annealing

KONSTANTINOS SFYRAKIS[a], ASTERO PROVATA[b,*], DAVID C. POVEY[a] and BRENDAN J. HOWLIN[a,*]

[a]*Department of Chemistry, School of Physics and Chemistry, University of Surrey, Guildford GU2 7XH, UK;* [b]*Institute of Physical Chemistry, National Research Center "Demokritos", 15310 Athens, Greece*

(Received June 2002; In final form June 2002)

A new algorithm is developed which implements the Monte Carlo annealing method and is used to model and minimize linear, single and double stranded B-DNA sequences. The preliminary B-DNA structures are modeled using initial structures obtained from the Brookhaven database. The model contains structural details at the atomic level and is therefore more elaborate and accurate than the pseudo-atomic and elastomechanical models. The minimization concerns the entire chain length and not only local nucleotide complexes. A variety of DNA sequences (coding or non-coding, random or real, homogeneous or heterogeneous) are investigated in the range of 20–40 base pairs. The potential energy function is written in terms of a set of internal coordinates defined to account for the helical parameters such as twist, tilt and rise, which are important parameters for the description of the global shape of any type of DNA or RNA molecule. The force field used is composed of a limited number of bonded and non-bonded interactions such as bond stretch, angle bend and Lennard–Jones interactions with the Dreiding II force field parameter set used for these interactions. From the minimized structures the angles between Phosphate–Oxygen–Carbon "A_1" and Oxygen–Phosphate–Oxygen "A_2" and the average helical twist were calculated. For single strands it is shown that the bond angles are $A_1 = 107 \pm 1°$ and $A_2 = 122 \pm 1°$, while the helical twist is $37.8 \pm 1°$. For double stranded DNA our model predicts the helical twist of $h = 35.5 \pm 2°$ well, in the A strand, while the prediction is less accurate, $h = 47 \pm 2°$ for the complementary strand B. The average values for the angles A_1 and A_2 are $130 \pm 1°$ and $150 \pm 3°$ for strand A and $102 \pm 4°$ and $123 \pm 5°$ for strand B. The reason for this discrepancy is attributed to the different conformations initially adopted by the sugars in the A and B strands.

Keywords: Coding/non-coding DNA; Single/double strand; B-DNA; Monte-Carlo; Annealing

INTRODUCTION

The folding and unfolding of DNA plays a critical role in its super-coiling behavior and for the binding of proteins to the DNA structure during the biochemical processes in the cell [1]. The structure of DNA is now studied from different points of view by medical/biological science through gene makers, bio-engineering and nanotechnology [2,3]. It is particularly important to investigate the different spatial conformations of this hetero-polymer and the different processes that allow or prevent its functioning. In particular, the different functional role of coding and non-coding DNA may be reflected in the various conformations adopted by the molecule during diverse biological process. It is also important to understand the effects of the chemical structure on the 3D structure and of the local or environmental forces, responsible for the processes of folding and unfolding. The aim of this work is to compare structural and coiling characteristics of single and double stranded DNA segments originating (a) from coding regions of DNA (b) homogeneous DNA segments found frequently in non-coding regions and (c) from artificial random DNA sequences. The increase in computational power in the last years makes possible this investigation of the folding and unfolding of large DNA segments under an accurate generic force field.

To this end, it is necessary to develop a method for modeling the three-dimensional structure of any DNA and RNA segment under an accurate and simple force field, for studying the dynamics of the folding process and for examining associated sequence characteristics such as the potential

*Corresponding authors. E-mail: aprovata@limnos.chem.demokritos.gr and b.howlin@surrey.ac.uk

ISSN 0892-7022 print/ISSN 1029-0435 online © 2003 Taylor & Francis Ltd
DOI: 10.1080/0892702021000008858

curvature of the overall structure. It is also important to investigate the major forces responsible for the folding of DNA or RNA polynucleotide strands, in order to form the stable geometrical structure of the double helix. A minimal complexity model will be adopted in this work.

An optimum energy minimization process of any organic molecule, large or small, requires all the bonds, angles and torsional angles of the molecule to be included during the modeling procedure. Although in the last years computing power has significantly increased, it is often unnecessary to include all degrees of structural freedom, especially for large macromolecules such as DNA. Today, orbital theories have proved that specific group of atoms (sugar rings) are very stable during many chemical or structural processes [4–7]. In order to decrease the degrees of freedom in this study, it is assumed that the sugar rings and the side chains of the DNA segments, remain geometrically stable and only the most energetically flexible parts of the macromolecule are taken into account. The three dimensional structure of a linear or non-linear DNA or RNA macromolecule can thus be determined by six degrees of freedom (torsional angles) of the sugar phosphate backbone and one more describing the orientation of the base about the glycosidic bond (Fig. 1). Using these seven degrees of freedom per nucleotide it is possible to construct highly flexible three-dimensional DNA or RNA structures [8] that could be used for energy minimization models. Later in this study, extra bond and angle degrees of freedom will be discussed, in order to monitor other structural changes in the model.

Recent numerical simulations of nucleotide sequences involve a variety of methods ranging from molecular mechanics (MM) to molecular dynamics (MD) [9] and Monte Carlo (MC) and involve works on Dynamics of B-DNA including water and counterions [10] and Dynamics of DNA denaturation [11]. Some models investigate descriptions of the internal dynamics of DNA using MD simulation methods ranging from atomic scales to mesoscopic scales [12] where many others are based on sequence, characteristics, functions and structure of DNA [13–21]. Throughout these methods it is very common to compare the structure and dynamics of modeled dodecamer sequences [22,23] with the Dickerson dodecamer [24–26]. Alexey K. Mazour used MD to obtain accurate and stable B-DNA duplexes where only the minor groove is filled with water and the bulk solvent is represented implicitly [27].

Additionally, other methods are using MC algorithms in order to investigate and equilibrate large double DNA segments. Such works on DNA concern helix deformations upon stretching using internal coordinates and give results that lead to force curves which exhibit a plateau as the conformational transition occurs [14]. Vlahovicek and Pongor use a different construction method to form along DNA chains which use in the final structure constraint MD simulations to find the lowest energy molecular conformation [16]. Besides simulation methods, systems of different equations have been used to describe DNA supercoiling and thus treat sequences containing thousands of base pairs [28]. Other methods investigate large

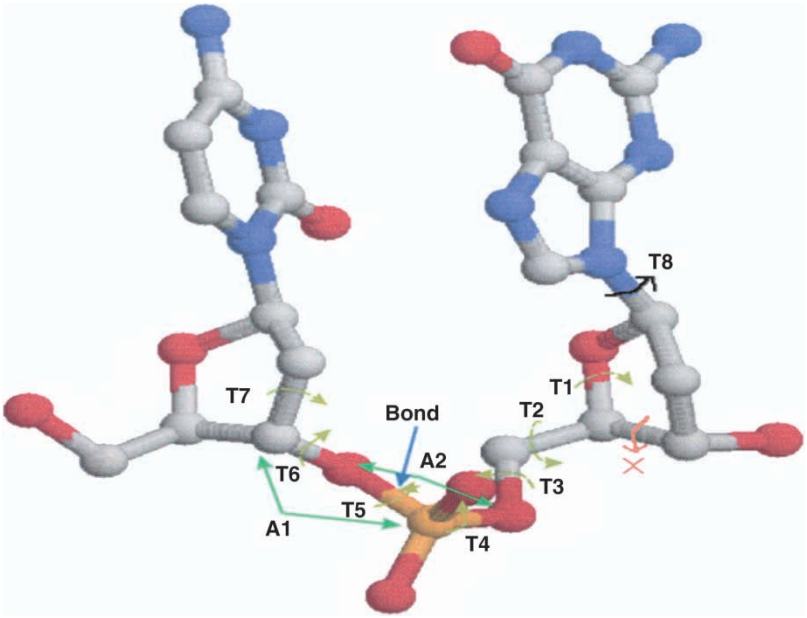

FIGURE 1 Two DNA Bases and Sugars Connected With a P–O Band. The DNA Strand is Rotated Via the O–P–O and P–O–C Angle As Shown. Different Types of Bonds, Angles and Torsions Used in the Simulation Are shown. (Colour version available online.)

supercoiled DNA using Langevin and Euler MD [11,29,30] obtaining trefoil knotting results [31].

MM has been used to study duplex DNA [32], giving good agreement with experimental data and strong sequence effects on both the position of the unpaired base and on the overall curvature induced by the abasic lesion [4]. Such approaches are based on a "multi-copy approach allowing to determine which base sequence favors a given structural change or interaction via a single energy minimization [33]. Wilma K. Olson and Victor B. Zhurkin develop an algorithm that models double-helical DNA at four levels of the three-dimensional structure, the all-atom, the base-pair, the mesoscopic and the scale of several thousand nucleotides level. They show that conformational changes are important for the functioning of the helix and affect its packaging in the close confines of the cell, the mutual fit of DNA and protein in nucleoprotein complexes, and the effective recognition of base pairs in recombination and transcription [15]. Another approach studies DNA oligomers with the MC method using an internal coordinates model associated with pseudorotational representation of sugar repuckering [34]. The MD simulation of the Dickerson dodecamer, by Yong Duan [22] shows, very good agreement with the X-ray structure (1.1 Å) using the particle mesh Ewald sum.

Many of the rules that determine the path through space followed by the central axis of the DNA macromolecule are not yet understood. In the case of linear DNA free in solution, the phenomenon of sequence-dependent bending has received considerable attention [36]. In the past, continuum elastic models have been used to investigate the trajectories of linear molecules [37–40]. These models treat the DNA macromolecules as isotropic, homogeneous, elastic materials without incorporating a good level of detail. Thus, they cannot predict accurately enough local sequence-dependent effects or conformational transitions requiring rearrangements of the secondary structure. More accurate models have been developed, where the structure is described by pseudo-atomic models [41]. These models are less detailed than all-atom models, having less structural complexity and smaller computational demands. However, these models do not help us to investigate the main reasons for DNA backbone bending and DNA folding because they do not contain enough information at the structural atomic level of the DNA pairs.

It is known that the DNA of living organisms is composed of coding segments responsible for the synthesis of proteins and intervening non-coding regions. Although both coding and non-coding regions are constructed as "random" sequences of base pairs, there are important statistical differences in their construction [2,35]. Striking examples are (a) repetitive segments in the non-coding region i.e. many copies of the same nucleotide string are found in juxtaposition in a long non-coding segment and (b) large strands of up to 100 purines (A's or G's) are found in the non-coding regions of higher eucaryotes with relatively high probability. These unexpected features of the non-coding regions may be reflected in their spatial conformations and functionality. Especially, they may be helpful both in the coiling of the helix and super coiling. In this study, we have chosen to study the structural conformations of three different types of single and double DNA stranded sequences, (a) segments of coding sequences obtained from real organisms (lambda virus), (b) segments of DNA consisting of the same base-pairs such as the ones which are found in non-coding regions of real organisms and (c) random, artificial sequences (the coding regions of DNA have the statistical characteristics of random sequences). Differences in the spatial conformations of the previous three categories of sequences may reflect differences in their functionality.

In this study, a model for folding single or double stranded B-DNA segments of any size and sequence is described, using the MC simulated annealing method. The model contains a very accurate representation of all the atoms for every base pair excluding the hydrogens. The hydrogens were not included because the simulation treats each base pair as a unit, assuming that the hydrogen bonds between the base pairs are unchangeable. Each base pair is represented by 14 degrees of freedom in total, containing bond stretching between phosphate and oxygen atoms in the secondary strand, bending angles and torsional angle transformations, with the hydrogen bonding between base pairs remaining unchangeable. Although it is generally assumed that MC simulations are less efficient for exploring the conformational space of macromolecules than MD [34] this research restricts the conformational space of each degree of freedom based on experimental data found in Brookhaven database (see Tables I and II). The force field includes two bonded and one non-bonded interaction whose parameters are taken from the Dreiding force field [42], treating the atoms as hard spheres. Successive base pairs, are connected by phosphate–oxygen bonds whose characteristics (bond lengths) are taken from high-resolution DNA crystal structures [23,24]. Computer algorithms

TABLE I The Range of All Randomly Generated Angles in Degrees for B-DNA structures

Symbol	*Angles*	*Strand*	*Range*
A1	P(a)–O(b)–C(b)	A	$0.0 \leftrightarrow 180.0°$
A2	O(a)–P(a)–O(b)	A	$0.0 \leftrightarrow 180.0°$

The a And b Integer Numbers Corresponding to the Base Counting With $b = a + 1$.

TABLE II The Range of All Randomly Generated Torsions in Degrees Based on Pdb File 166D.pdb

Symbol	*Torsion Angles*	*Strand*	*Range* (*strand A*)	*Range* (*strand B*)
T1	C(a)–C(a)–O(a)–C(a)	A, B	110.0 ↔ 130.0°	110.0 ↔ 130.0°
T2	O(a)–C(a)–C(a)–O(a)	A, B	−50.0 ↔ −90.0° 150.0 ↔ 160.0° −140.0 ↔ −160.0°	−50.0 ↔ −100.0°
T3	P(a)–O(a)–C(a)–C(a)	A, B	140.0 ↔ 180.0° −150.0 ↔ −180.0°	140.0 ↔ 180.0° −150.0 ↔ −180.0°
T4	O(a)–P(b)–O(b)–C(b)	A	0.0 ↔ 360.0°	Not Used
T5	C(a)–O(a)–P(b)–O(b)	A	−60.0 ↔ −180.0°	Not Used
T6	C(a)–C(a)–O(a)–P(b)	A	60.0 ↔ 120.0°	Not Used
T7	C(b)–C(b)–C(b)–O(b)	A, B	60.0 ↔ 110.0°	60.0 ↔ 110.0°

The a And b Are Integer Numbers Corresponding to the Base Counting With $b = a + 1$.

similar to those of MM and MD are used and the conformational space is searched for global minimum structures exploring them in a wide range of temperatures. The same algorithm provides high-resolution structural features at the level of individual atoms, allowing for the determination of local variations in the helix twist and bending and thus, permitting the convenient treatment of sequence-dependent effects. The programs *Rasmol* and *Swiss PDB Viewer* [43,44] are used for displaying and comparing the resulting DNA structures and the program *pdbfit* that calculates the root mean square (RMS) structure fitting [43], is used to compare the crystallographic B-DNA structure with the ones produced by our model.

In the next section, the model of B-DNA is described and the most important characteristics of its structure are explored. The initial structures used for modeling the linear "comb-like" starting DNA segments are described and the different types of DNA sequences used for the minimization are presented. Next, the method of producing random DNA structures and the force field used to calculate the energy of each one of these structures are described followed by explanation of the basic algorithm and its parameters. In the third section, the results of the single strand of B-DNA simulations are presented with emphasis on the characteristics, which will produce, in the next step, the double helix. Additionally, the conformations of single stranded DNA may represent those of RNA when thymine is replaced by uracil. Results are presented for (a) sequences coming from coding DNA of real organisms, (b) for homogeneous segments, which are frequently met in non-coding DNA and (c) for artificial random sequences. In the fourth section, the results on the double stranded B-DNA helix are presented and the cases a, b, and c are studied. In the fifth section, the differences and similarities in the spatial conformations of coding and non-coding DNA sequences are studied by comparing the structural characteristics of the structures produced. It is shown that on the average the bond angles for the single DNA strands are $A_1 = 107 \pm 1°$ and $A_2 = 122 \pm 1°$, while for the double strands $A_1 = 130 \pm 1°$ and $A_2 = 150 \pm 3°$ for strand A and $102 \pm 4°$ and $123 \pm 5°$ for strand B, respectively. The average helical twist h for the single strands are $h = 37.8 \pm 1°$, while for the double stranded simulation, $h = 35.5 \pm 2°$ for the A strand and $h = 47 \pm 2°$ for the B strand. It is obvious that the prediction is less accurate for the B strands owning to the method, which considers initially connected sugars for strand A but initially disconnected sugars for the complementary strand B. Finally, in the concluding section the main conclusions of this work are discussed and future applications, modifications and improvements of this method are proposed.

MODEL AND METHOD

In this section, the most important structural characteristics found in the B-DNA form are described before simulating this DNA type. Also, the initial DNA structure used to predict the single or double B-DNA strands and the parameters needed to develop the force field for this simulation are described as a function of a set of the internal coordinates. The use of MC and simulated annealing [45] is also described in order to determined the lowest energy structure of the diverse sequences.

For the generation of the minimum energy structures, a multiprocessor server containing four CPU running at 150 MHz each is used, running on the IRIX64 operating system. The simulation was written in the C++ programming language using the *gcc* compiler, creating classes for each of the different parts of the algorithm. The most important classes are the generation of the new structures, the calculation of the energy force field and the initialization and storing of the molecule during the minimization. The program is written in such a way as to accept from the command line different types of force field (Van der Waals, Lennard–Jones and bond angle potentials) and different molecular structures (single or double chain). The DNA sequence is also written from the command line. During

the minimization, for each interval the molecule is stored to a different file and all the different parts of the energy force field for each step of the process are stored in the same file.

Structural Characteristics of B-DNA

In the current work the investigation of B-DNA is chosen since it is the most common DNA structure found in nature. Also, it is regarded as the native form because its X-ray pattern resembles that of the DNA in intact sperm heads [46]. However, the general methodology of this research should be similar for minimization of the other forms of DNA, by changing only some of the basic parameters of this minimization procedure.

B-DNA consists of two polynucleotide strands that wind about a common axis with a right-handed twist to form a $\sim$20 Å diameter double helix [1,47]. The two strands are anti-parallel (run in opposite directions) and wrap around each other such that they cannot be separated without unwinding the helix. The natural B-DNA helix has 10 base pairs (bp) per turn (a helical twist of 36° per bp) [1,47], a pitch (rise per turn) of 34 Å and the planes of the bases are nearly perpendicular to the helical axis [1,46]. Later, these basic structural characteristics found from the X-ray crystallographic data are compared with our modeled structures.

General Description of the Algorithm

The algorithm starts with an initial state of very high energy. The choice of this initial state is described in the "Initialization of Structure" section. For the construction of the new structures, the random generator method is used, described in the "Random Generator of New Structures" section. The energy of every new molecular structure is calculated using a general force field described in the "Force Field" section, whereas the methodology and the parameters used to obtain the conformational energy minimum of the structures are presented in the "Monte-Carlo Annealing Simulation" section.

Initialization of Structure

The initial coordinates of the four DNA bases are copied from a high-resolution *pdb* file *166D.pdb* [48], found in the Brookhaven database [49]. The 166D.pdb file contains a B-DNA polynucleotide strand (*Deoxyribonucleic acid*), and a docked *gamma-oxapentamidine*. It is formed by 12 base pairs with the sequence in strand A (*cgcgaattcgcg*) and is resolved at a resolution of 2.2 Å (see Tables AI and III).

This polynucleotide strand is known by the name "Dickerson-Drew Dodecamer" and is widely discussed in the literature [24–26]. Some of the studies in the literature involve comparison with its own X-ray structure and that of the NMR structure of the native counterpart [50], others involve MD including water and counterions [10,22,51] and others investigate the stability and the conformation of the Dodecamer inducing sugar puckers or binding to other molecules [52,53].

In order to produce the initial building blocks of the simulated DNA strands, the *pdb* file was decomposed into its constituent base pairs with attached side-chains. Each pair of bases i.e. *AT*, *TA*, *GC*, *CG* from the *pdb* file were then superimposed, using the side-chains and the sugar part, onto each other using the program *pbdfit*. The average coordinates of the backbone were then generated and used to build the starting model. These pairs remain unchanged at their interconnections (hydrogen bonds), during the energy minimization, in order to make the system computationally simpler.

The initial structure of single stranded DNA is built by connecting single DNA bases in such a way that they form a linear strand and the minimization depends only on the angle bend and Lennard–Jones interactions (Fig. 1).

TABLE III The Values for the Torsion Angles for the B-DNA Strand from the Pdb File 166D.pdb

	Torsions (*degrees*)										
	Strand A							*Strand B*			
Pair Sequence	*T1*	*T2*	*T3*	*T4*	*T5*	*T6*	*T7*	*T1*	*T2*	*T3*	*T7*
cg	116.8	−77.9	173.9	−55.5	−119.6	78.4	89.9	126.0	−70.4	−157.6	71.8
gc	114.8	−96.4	163.8	−65.5	160.3	149.2	80.1	106.4	−62.4	150.8	−68.9
ag	115.2	−65.3	175.8	−73.0	−117.2	88.2	86.4	110.5	−78.5	149.9	81.3
aa	126.9	−77.0	179.6	−59.2	−111.4	83.1	76.0	114.8	−88.4	171.6	71.9
ta	125.7	−53.5	173.9	−69.4	−102.4	91.2	75.0	124.7	−56.2	168.9	84.6
tt	120.3	−58.9	164.8	−69.9	−112.8	83.7	74.7	124.9	−57.7	170.2	78.3
ct	117.7	−68.6	179.7	−60.4	−98.1	62.4	74.4	121.0	−54.1	173.9	71.8
gc	108.7	−65.0	−162.3	−67.6	−106.9	68.8	71.6	113.3	−60.6	168.6	74.5
cg	114.9	−56.8	145.1	−54.6	−152.7	102.1	87.2	108.9	−60.8	177.7	106.5
gc	121.7	−50.0	173.3	−73.7	−95.8	70.9	80.0	107.7	156.9	−161.1	73.9

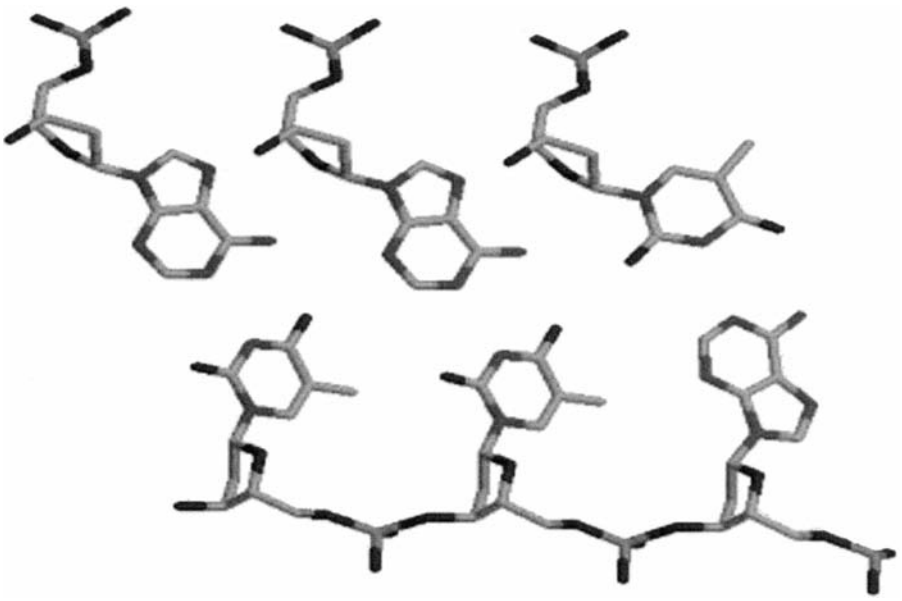

FIGURE 2 Sample of Three DNA Base Pairs, Initialized for the Minimization procedure.

The initial structure of the double stranded DNA is built by connecting DNA base pairs in such a way that they form a linear strand. For double strand minimization, the bases of the second strand are initially not bonded to each other and bonded to the bases of the opposite strand, only with hydrogen bonds (Fig. 2). In that way the second strand is allowed to fit its own conformation without bias, the number of calculations for the minimization is reduced and the bonds are allowed to form normally from the conformation chosen. The current version of the simulation program can handle both, single or double DNA strands assuming that they exist in a vacuum (solvent molecules are excluded). The initial state used here is chosen in order to show how the single chain curls under the influence of the force field. Other authors use different initial configurations as starting states [21].

For the generation of the minimum energy structures both, single and double DNA strands are used. Some of them are homogeneous or artificial (random) and others are extracted from the DNA of test organisms. The real DNA segments are taken from a coding region of DNA, of the *Bacteriophage Lambda*. The *Lambda* virus (as is usually known) is one of the most commonly used test organism and it consists of almost pure coding DNA. In order to understand how DNA strands coil, sequences with variable number of bases (2, 5, 10, ... 40 base pairs) are used during the minimization.

In each run, four different structures are minimized simultaneously in order to decrease the total time length of the minimizations for all the runs. The number of new conformations depends on the number of base pairs minimized and the different energy components forming the force field. A 20-base, double DNA strand needs approximately 17 h for minimization, and a 40-base double DNA strand needs approximately 127 h. The time to complete each simulation increases exponentially, depending on the number of bases. Three different strategies were followed in order to test the algorithm and to compute the structural characteristics of DNA strands: First, single strands were minimized using only the bond angle potential, secondly, both the bond angle potential and Lennard–Jones interactions were used and finally, double strands were minimized including all the force field interactions.

Random Generator of New Structures

An important part of this molecular energy minimization simulator is the random generator of new structures of DNA. During each step, the program randomly finds a base–base pair from the strand and rotates it in three-dimensional space. In order to find a new lower energy conformation, the following steps are taken:

- Choice of a random nucleotide.
- Choice of two randomly generated angles corresponding to the angles A_1 and A_2, (Fig. 1), with a defined step of $\pm 0.1°$, respectively (see Table I).
- Choice of 11 torsion angles, 7 defined for the first strand and 4 for the second, with a defined step for all of them of $\pm 0.1°$. The T8 torsion angle is not used in this method in order to reduce the size of the conformational space, (see Fig. 1 and Table II).

The angles defined in Fig. 1 and Table I do not completely correspond to the standard nucleic acid torsion notation [54] but they are chosen here as more appropriate for the needs of the current algorithm. In the case of the two strand minimization, seven torsion angles are used for the first strand while only four torsion angles are enough for the second strand because the remaining three are defined by the geometrical constraints of the problem. All other angles and lengths shown in Fig. 1 are kept constant during the simulations. More precisely, the sugar ring geometry and the bases (A, T, C and G) remain undistorted. Moves of O–P–O and P–O–C valence angles are allowed as general as possible (0–180°) in order to give the molecule maximum freedom to reach the minimum energy state.

The prediction of the step for both angle and torsion is based on experience to avoid very large or small conformation changes. The range for the angles is defined to give maximum rotation in space, whereas the range for the torsion angles is defined from the file *166D.pdb* and based on the type of DNA that is minimized. In Table II, the three torsions for the second strand are not defined because they will adopt the correct configuration by changing the previously defined angles and torsions. The mathematics behind the transformations and rotations of each segment of DNA are based on simple vector algebra [5,7], using the internal and Cartesian coordinates of the system [55].

Force Field

This molecular energy minimization model is developed with a force field suitable for fast and accurate predictions of the energy of the DNA structure treating the atoms of the same type identically. It is assumed that the potential energy of a molecule with arbitrary geometry is expressed as a superposition of valence (or bonded) interactions (E_{val}) that depend on the specific connections (bonds) of the structure and non-bonded interactions (E_{nb}) that depend only on the distance between the atoms. The valence interactions consist of the bond stretch (E_{BND}, two-body interactions) and the bond-angle bend (E_{ANG}, three-body interactions), whereas the non-bonded interactions contribute to the van der Waals or dispersion potential (E_{VDW}).

$$E_{\text{total}} = E_{\text{val}} + E_{\text{nb}} = E_{\text{BND}} + E_{\text{ANG}} + E_{\text{VDW}}$$

In the current version of the model the atoms are allowed to move by two, three and four body interactions, which are formed by one, two and three bonds, respectively. In order to adjust the total energy of the force field to be independent of the length of the DNA strand the total energy is divided by the total number of bases. To determine the accuracy and the global energy minimum point (defined at 1 arbitrary unit), the energy of the known DNA structures is calculated. Table IV shows these structures and their energies.

TABLE IV The Energy of Real Double DNA Strands Calculated by the Force Field Used in the Current Algorithm. The Equilibrium Energy Values for Each Structure Are recorded

DNA sequences	*Energies* (*Arbitrary Units*)			
	Total	*VDW*	*ANG*	*BND*
cg, Gc	1.039	0.600	0.396	0.042
aa, Tt	1.127	0.653	0.465	0.009
at, Ta	1.217	0.738	0.475	0.004
ag, Ga	1.438	0.887	0.455	0.097
aatt	1.141	0.684	0.451	0.006
attcg	1.109	0.662	0.429	0.018
cgaattcg	1.149	0.684	0.433	0.032

One type of two-body interaction is included in this energy force field. It is between each phosphate and neighboring oxygen atom of the second strand only, both of them are responsible for the bonding of the neighboring bases. The bond length between these atoms in the first strand remains unchanged and no energy interaction is needed. The prediction of the equilibrium bond distance R_0, for that bond is based on the structural data found from the previous X-ray derived DNA molecules. In order to form the phosphate–oxygen bond, a large force must be applied to the B strand of the DNA. This is the complementary B strand where the bases are apart and bonded only with hydrogen bonds, to the bases of the opposite A strand. To achieve this, the energy of this interaction is increased by a factor of 10^{10}. The bond energy is then found by the formula:

$$E_{BND} = \sum_{n=0}^{\text{total bases}} (R - R_0)^2$$

where R is the new distance between phosphate and oxygen and R_0 is the equilibrium distance defined at 1.60 Å.

There are two types of three-body interaction taken into account in the force field. The first takes place between two oxygen atoms and the phosphate and the second between the oxygen, the phosphate and the carbon atom of two neighboring bases, which are allowed to rotate in three-dimensional space. The energy of every three-body interaction is calculated from the formula:

$$E_{\text{ANG}} = \sum_{n=0}^{\text{total bases}} [K \sin(\vartheta - \vartheta_0)]^2$$

Where under the usual approximation $(\vartheta - \vartheta_0) \ll 1 \Rightarrow \sin(\vartheta - \vartheta_0) \cong (\vartheta - \vartheta_0)$ the E_{ANG} becomes:

$$E_{\mathrm{ANG}} = \sum_{n=0}^{\text{total bases}} K(\vartheta - \vartheta_0)^2,$$

$$K = 100(\mathrm{kcal/mol})/\mathrm{degrees}^2$$

where K is the force constant for all angle bend interactions; ϑ (in degrees), is the new angle between the two bonds and ϑ_0 is the equilibrium angle. The equilibrium angle ϑ_0 was found from a high resolution X-ray file, from the Brookhaven databases [42] and is 108° for the first and 121° for the second interaction. However, both strands in the DNA molecule are also allowed to move by four body interactions; the energies of these interactions were not included in the force field. This is due to the large number of different torsion angles found in the real DNA structures for each one of these and to the difficulty and uncertainty in predicting the constants for the various torsion interactions.

In the non-bonded energies the hydrogen bond is not included. However, if the second strand of the helical DNA is added to the total structure allowing structural changes between the base pairs (not allowed in this method), the hydrogen bond must be incorporated. Even though the phosphate groups in the DNA backbone are heavily charged, the electrostatic interactions are not incorporated in the current model. This is due to the nature of these charges, which are balanced by the presence of magnesium ions associated with the DNA, which are not taken into account in the simulation.

The third non-bonded interaction is the van der Waals interaction, which is described by the Lennard–Jones expression (see Fig. 3). This expression describes the potential energy of two non-bonded molecules or atoms. For short distances, the nuclear and electronic repulsions are acting and the rising kinetic energy begins to dominate the attractive forces. The repulsions increase steeply with decreasing separation in a way that can be deduced only by very extensive, complicated molecular structure calculations.

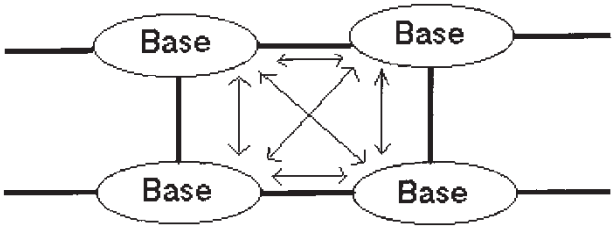

FIGURE 3 The Lennard–Jones Interactions Between Two Adjacent Base Pairs. In Total Eight Different Interactions Are Calculated for the Double Strands and Two for Single strands.

TABLE V The Van Der Waals Parameters Used to Calculate the Potential energy

Atom	r_0 (Å)	C (*kcal/mol*)
H	3.195	0.0152
C	3.8983	0.0951
N	3.6621	0.0774
O	3.4046	0.0957
P	4.1500	0.3200

The sum of the repulsive and the attractive interaction is here approximate by the Lennard–Jones (n, 6)-potential:

$$V = \frac{C_n}{r^n} - \frac{C_6}{r^6}$$

As usual for mathematical reasons it is convenient to select $n = 12$. The new equation can be written as:

$$V_{\mathrm{vdw}} = \sum_{n=0}^{\text{total bases}} C\left\{\left(\frac{r_0}{r_n}\right)^{12} - \left(\frac{r_0}{r_n}\right)^{6}\right\}$$

where C is a constant parameter, r_0 is the equilibrium distance between the two atoms or molecules, and r_n is the new distance between them (see Table V). To truncate large numbers of Lennard–Jones calculations, the energy between neighbor base pairs is calculated and a cut off distance of 10 Å was used.

For single stranded DNA structures, only angle bend and Lennard–Jones potentials are used. The angle bend potential is used to calculate the A_1 and A_2 angles (Fig. 1) whereas the Lennard–Jones potential is calculated between adjacent bases. For the double strands all three potentials (bond stretch, bond angle and Lennard–Jones) are used. The bond stretch is calculated only for the second strand where the sugars attached to the bases are not connected. The bond angle for both strands and the Lennard–Jones interaction between adjacent and neighbouring bases are measured and are recorded in Table IV.

Monte-Carlo Annealing Simulation

In order to identify the global minimum of the total energy (in arbitrary energy units), an algorithm that combines Metropolis MC sampling with a simulated annealing [45] procedure is used. The starting conformation is always a linear strand in vacuum with only three types of interaction taken into account, two bonded and one non-bonded as explained in the "Force Field" section. During the minimization, the algorithm searches for the minimum energy structure at a given temperature. For every temperature a large number, N, of attempts at minimizing the energy is undertaken. These N attempts are called one "interval". Each attempt is also called a "stage" or "step". The number N of steps depends on the size of the sequence and is

normally taken as $N = 200 \times$ (size of sequence). The annealing process usually finishes after 200 (double chain) or 300 (single chain) intervals (or temperature grading) while during the processes of the minimization only $\approx 50\%$ of the generated structures are accepted. During this process at every 20 stages, the new structure and the values of the energies, angles and torsions of the new strands are saved. At each step, the strand is allowed to move randomly in the configuration space, by changing angles or torsions of a random base–base pair, in the range of $\pm 1.0°$. During the simulation, the algorithm keeps track of two particular strand conformations, the current and the trial. The energy is then calculated for the new conformations saving the most favorable ones.

The lower energy structures are automatically accepted and those of higher energy are accepted on the basis of the Boltzman factor of the energy increase. The new conformations are accepted with probability p given by:

$$p = \begin{cases} 1 & \text{if } \Delta E < 0 \\ e^{-(\Delta E \times 1000)/T)} & \text{if } \Delta E > 0 \end{cases}$$

This means that the new structures are accepted only if the random number between 1 and 0 is smaller than the quality $e^{-(\Delta E \times 1000)/T)}$, where T is the current temperature of the system and ΔE is the energy difference $E_{\text{new}} - E_{\text{old}}$. The value 1000 in the expression guarantees that 40–60% of the moves among all trials are accepted. This value works only with the current energy scale. Changing the scale of the force field would change the acceptance ratio. The system is then allowed to approach an equilibrium distribution at a given starting temperature, T_0. The starting temperature is obtained from the expression $E_0 + 100$, where E_0 is the starting energy and 100 is chosen by experiment. The temperature is reduced by an accelerated cooling procedure where its current value T is lowered by a factor of 97% during the 200 intervals. The cooling process is allowed to continue only if the energy E_{old} is larger or equal to 1 unit (see Table IV) or the temperature T is lower than 1. Otherwise the minimization process is terminated, saving the energy, torsion, angle and bond values and the coordinates of the equilibrated structure at this point.

SINGLE STRANDED B-DNA

The simulation in all cases starts with linear (straight) strands. To test the algorithm, two different cases have been considered the first using only the bond angle potential and the second using the bond angle and the Lennard–Jones potential.

Using Only the Bond Angle Potential

For single stranded DNA several test trials were performed using only the bond angle potential E_{ANG}. Figure 4 shows the minimized structures of four different single stranded DNA sequences, the homogeneous *cc* strand of 20 base pairs (I), the random *tcaaatgggc cccgaaatcg* 20 base pairs sequence (II), the homogeneous *cc* strand of 40 base pairs (III) and the real coding *tgagaacgaa aagctgcgcc gggaggttga agaactgcgg* 40 base pairs DNA sequence (IV) (see Table AII).

The average values of bending angles for the first structure are $A_1 = 107.99°$ and $A_2 = 122.01°$, for the second $A_1 = 107.991°$ and $A_2 = 121.99°$, the third $A_1 = 108.00°$ and $A_2 = 121.99°$ and for the fourth $A_1 = 108.00°$ and $A_2 = 122.01°$. In all cases the values are very close to the equilibrium values of 108.0 and 122.0°, respectively. The algorithm has converged and the maximum simulation time was 70 h. This test simulation was performed in order to investigate how fast and how accurately the annealing method works.

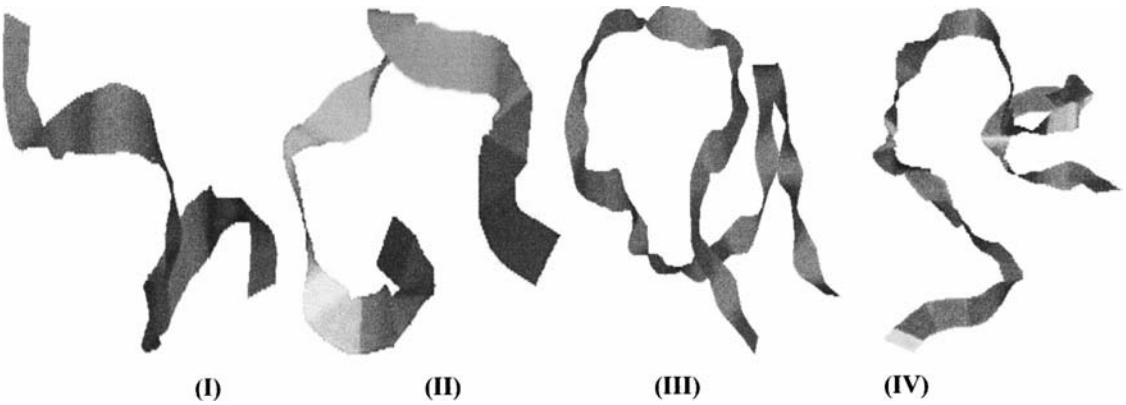

FIGURE 4 Display of the Backbone (ribbon Representation) for Four Different DNA Sequences After the Energy Minimization. (I) DC_{20} Base Pairs, (II) Random *Tcaaatgggccccgaaatcg*, (III) DC_{40} Base Pairs, (IV) Coding *Tgagaacgaa Aagctgcgcc Gggaggttgaagaactgcgg*.

Using the Bond Angle and the Lennard–Jones Potential

For single stranded DNA sequences energy minimization was again performed using both bond angle and Lennard–Jones potentials. The same four single stranded sequences were used and their bending angles A_1 and A_2 were also recorded (see Table VI).

Again the average values for the bending angles for the first structure are $A_1 = 107.99°$ and $A_2 = 122.01°$, for the second $A_1 = 108.01°$ and $A_2 = 122.00°$, the third $A_1 = 108.01°$ and $A_2 = 122.01°$ and for the fourth $A_1 = 108.01°$ and $A_2 = 122.00°$. As expected in all cases the values are very close to the equilibrium values of 108.0 and 122.0°, respectively. This type of simulation was performed in order to investigate how single DNA strands behave in vacuum and the structure they form under simple force fields. To find out the helical twist per base for each one of the four previous structures, the torsion angle X (Fig. 1) was used. All torsion angles per structure were summed and were divided by the total number of bases in order to find out the average torsion, which is then subtracted from 180°. The resulting angle is the helical twist per base denoted by h. If h is multiplied by 10 (or 20, 40 …etc., depending on the length of the strand), it gives the average total rotation per strand similar to the base pairs per helical turn.

In all cases the torsion angles are very close to 142.2° with helical twist per base $h = \sim 37.8°$ (for (I) $h = 37.78°$, for (II) $h = 37.76°$, for (III) $h = 37.79°$ and for (IV) $h = 37.75°$), close to the 36° of the double stranded B-DNA structure. In all different DNA sequences it is observed that the average helical twist is independent of the size of the strand and the type (coding or non-coding, homogeneous or heterogeneous) and the average angles A_1 and A_2 are very close to the equilibrium values with average difference of $\pm 0.1°$. In Figs. 4 and 5 all the structures contain random helical turns, whereas in the second case, Fig. 5, the side chains, repel each other as expected, in order to minimize the Lennard–Jones potential. It is evident from the calculations in the current section that because the side chains in a single strand repel each other, they will be distributed in a spiral manner around the backbone of the single strand. As a consequence, when (in the next section) the two strands join together, the resulting equilibrium structure will be the double helix.

TABLE VI The Starting and the Final Energy Values of the Minimized Structures for the Case 3.2, B-DNA strands

		Initial Energy (arbitrary Units)				*Final Energy (Arbitary Units)*			
Type	*––––––10––––––20––––––30––––––40*	*Total*	*VDW*	*ANG*	*BND*	*Total*	*VDW*	*ANG*	*BND*
Homogeneous	cccc	78312.9	5.188	78307.8	0.0	2.383	2.332	0.052	0.0
Dyadic	cgcgcgcgcg	78323.5	15.748	78307.8	0.0	2.082	1.979	0.103	0.0
Dyadic	cgcgcgcgcgcgcgc	78324.7	16.957	78307.8	0.0	2.671	2.449	0.223	0.0
Real	cgcgaattcgcg	78322.6	14.863	78307.8	0.0	2.531	2.301	0.200	0.0
Homogeneous	ccccccccc Ccccccccc	78312.9	5.188	78307.8	0.0	2.707	2.579	0.128	0.0
Homogeneous	ccccccccc Ccccccccc Cccc	78312.9	5.200	78307.8	0.0	3.371	3.189	0.182	0.0
Homogeneous	ccccccccc Ccccccccc Ccccccccc	78312.9	5.188	78307.8	0.0	3.212	2.861	0.351	0.0
Homogeneous	ccccccccc Ccccccccc Ccccccccc Ccccccccc	78312.9	5.188	78307.8	0.0	3.161	2.893	0.268	0.0
Real-coding	tgagaacgaa Aagctgcgcc	78323.6	15.813	78307.8	0.0	2.756	2.612	0.144	0.0
Random	tcaatgggc Cccgaaatcg	78320.9	13.111	78307.8	0.0	2.719	2.580	0.140	0.0
Real-coding	actgataccg Cggttgctga	78331.4	23.665	78307.8	0.0	2.563	2.372	0.191	0.0
Real-coding	tgagaacgaa Aagctgcgcc Gggaggttga Agaactgcgg	78328.3	20.552	78307.8	0.0	3.197	2.878	0.319	0.0

DOUBLE STRANDED B-DNA

For double stranded DNA all interactions have been used as explained in the "Force Field" section and the same sequences as before are used (Table AIII).

The method starts with the two strands (Table AIII) A and B, where strand A is the one with the sugars initially connected, while strand B is the complementary strand with the sugars initially disconnected and the nucleic acids of B (together with their sugar) attached to strand A via a hydrogen bond (Fig. 2). This method is called the "sugar disconnected method".

Figures 6 and 7 show a real double stranded B-DNA structure with the nucleotides sequence of strand A *cgcgaattcgcg*. The structure was minimized, starting from a linear, high energy conformation, with the sugars of the complementary strands not bonded (see Fig. 2). This molecular structure was saved every 20 intervals, starting from the initial structure (a), to the 20, 40, 80, 100, 120, 140, 160, 180 intervals and the final structure (j) at 200 intervals. To simplify the minimization algorithm the hydrogen bonds remain unchanged with no rotation between sugars and the side chains. To force the molecule to fold, the bond energy is increased by a large factor of 10^{10}. In this way the non-bonded strand is forced to form the "Phosphate–Oxygen" bond, giving a final structure close to the natural B-DNA. Other DNA sequences were generated giving similar results, including homogeneous or heterogeneous sequences and real or random base sequences.

During the minimization, the temperature of the system was decreased during each interval, allowing the structure to relax to the equilibrium position. In total, 19 different B-DNA structures were minimized, which consisted of the minimum two pairs to the maximum 40 pairs of DNA. Larger strands were not used because of the extended simulation time needed. From all the sequences only the sequence *cgcgaattcgcg* was compared with the 166D.pdb DNA obtained from the Brookhaven database because it contains the same sequence of DNA pairs. This sequence is used as a test heteropolymer to examine the validity of our method. The quantitative structural results for this sequence are in good agreement with X-ray findings. It was not possible to compare the RMS deviation for all the minimized structures because there are no suitable X-ray structures in the Brookhaven database.

Structures of DNA larger than 12 bases are not available from the database, since only 10 bases are needed to make a complete rotation round the helical axis in B-DNA. The minimized structure obtained from the simulations (see Figs. 6 and 7) is superimposed on the initial B-DNA strand found from Brookhaven database, using the *pdbfit* program [43], Fig. 8. It is found that the RMS between the two structures is 1.605 Å.

Our algorithm gives correct predictions of the structure of the known *cgcgaattcgcg* structure, including the angles (Fig. 9) and the bond lengths (Fig. 10) and thus can be used to predict the structure of other DNA sequences of coding and/or non-coding origin in order to explore differences in their spatial structure. Fluctuations around the equilibrium distance are of the order of ± 0.1 Å and around the mean angle are of the order of $\pm 40°$ in all generated homogeneous or non-homogeneous DNA stands. In strand A, the observed natural equilibrium values of the angle fluctuations are higher than the equilibrium, in contrast with the complementary strand B where the values oscillates around the equilibrium value.

From Fig. 11, for the first few intervals the energy is reduced rapidly during the minimization, whereas afterwards it is decreased almost logarithmically with large fluctuations during the process. The largest part of the energy is from the bond between phosphate and oxygen of the second strand of the DNA. This bond energy is decreased rapidly in the beginning of the process making it the most important part of the force field, whereas the other energy components are more important during the later stages of the process (see Table AIII).

The average helical twist, h, of the *pdb* structure is calculated as 35.7° for strand A and 35.1° for

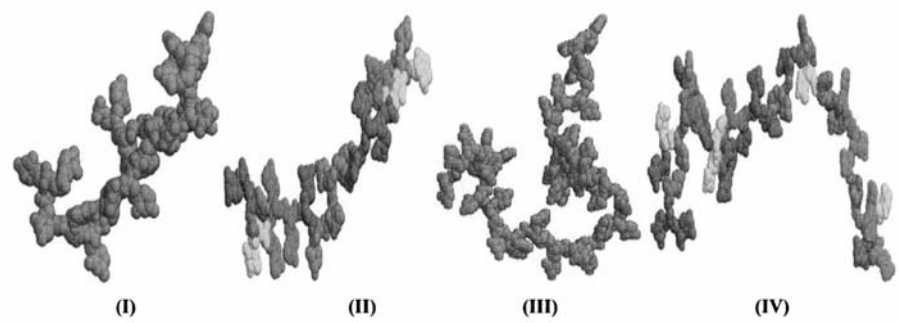

FIGURE 5 Display of the Backbone (all Atom Representation) for Four Different DNA Sequences After the Energy Minimization. Sequences (I)–(IV) Are the Same As in Fig. 4.

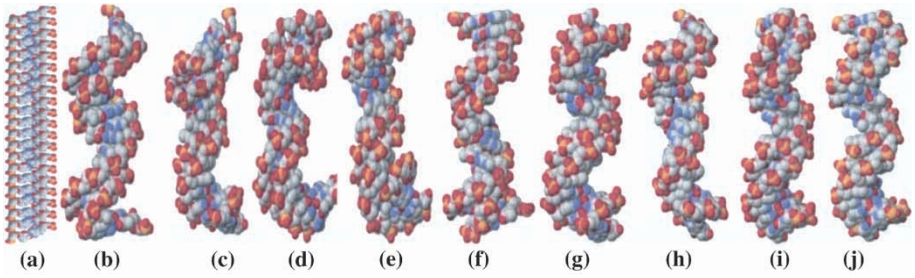

FIGURE 6 The Evolution of Double Strand B-DNA Structure of the Dickerson Dodecamer As Represented by Ball and Stick Drawings Viewed Perpendicular to the Helix axis. (Colour version available online.)

the complementary strand B and the average angles A_1 and A_2 are 99.25° and 121.08° for strand A and 97.70° and 121.21° for strand B. The difference in the value of h between strands A and B may be attributed to distortion due to the presence of an organic molecule, which was attached to the structure. From the simulations the helical twist was measured as $h = 37.90°$ and $h = 45.46°$, for both strands A and B, respectively, where the average angles A_1 and A_2 was 74.05°, 126.97° and 129.30°, 151.14°. A factor preventing exact overlap of the two structures is the organic molecule intercalated into the structure in the *pdb* file. This is not taken into account in the simulation, although organic molecules (such as drugs) in the minor groove typically have little effect on the helical conformation.

In the simulations, the difference in the value of h in the two strands is attributed to the different initial conformations of the two strands A and B. It is worth remembering that in strand A, the sugars attached to the bases are interconnected, whereas they were not in strand B. This probably gives rise to different final values of h in the two strands. Modification of the method so that the two strands have similar initial conformations are expected to give better agreement for the values of h, A_1 and A_2 between the two strands.

DIFFERENCES AND SIMILARITIES BETWEEN CODING AND NON-CODING DNA'S

To explore the differences and stress the similarities between coding and non-coding DNA sequences we have simulated the double stranded DNA sequences presented in the "Using the Bond Angle and the Lennard–Jones Potential" section (Appendix A, Table AIII) following the procedures described in the previous section. In Fig. 12 the two largest sequences are presented. The first sequence was homogeneous containing 40 cg pairs while the second had the pattern *tgagaacgaa aagctgcgcc gggaggttga agaactgcgg* in strand A and the complementary pattern *actcttgctt ttcgacgcgg ccctccaact tcttgacgcc* in strand B. This is a coding sequence obtained from the *Lambda* virus.

From the four structures in Fig. 5 (same sequence but with the double strands), the average values of the bending angles A_1 and A_2, of strand A and B are calculated and displayed in Table VII. In all cases the values are very close to the equilibrium values of

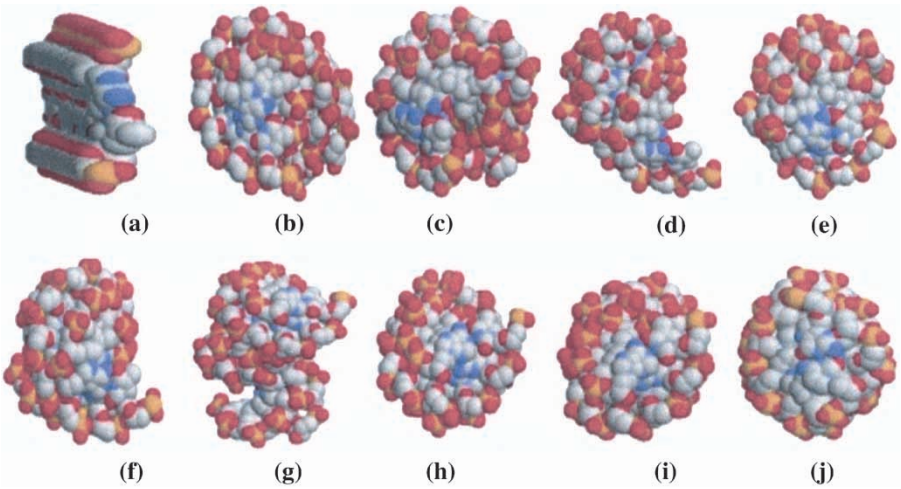

FIGURE 7 The Evolution of Double Strand B-DNA Structure With the Same Sequence As in Fig. 6, As Represented by Ball Drawings Viewed Down the Helix axis. (Colour version available online.)

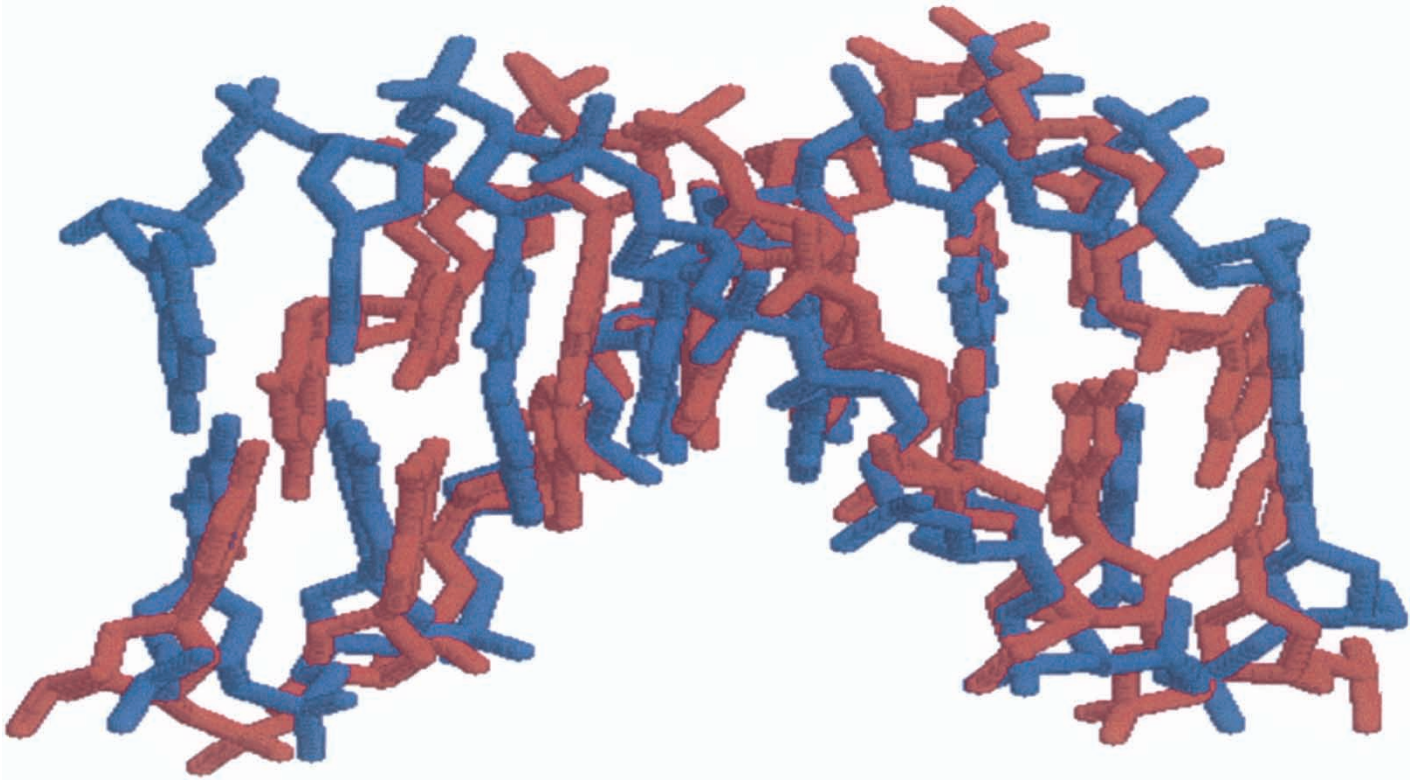

FIGURE 8 The Superposition of the DNA Dickerson Dodecamer Minimized Using Monte Carlo Annealing, With the Same Sequence Found in Brookhaven Database. Red is the Minimized Structure and Blue is the Real Structure from the Pdb file. (Colour version available online.)

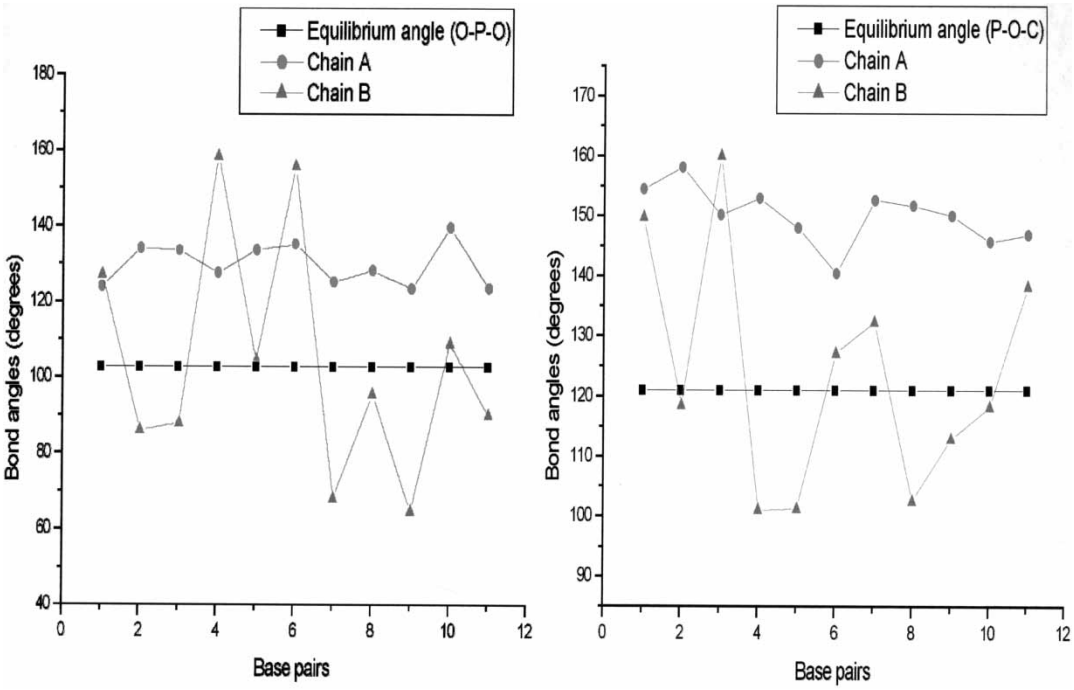

FIGURE 9 Values of the Angles $A_1 = \mathrm{O-P-O}$ And $A_2 = \mathrm{P-O-C}$ In Degrees, for Sequence *Cgcgaattcgcg* And *Gcgcttaagcgc*. For Comparison the Values of the Equilibrium Angle is Denoted by squares.

(Table VII) 108.0 and 122.0°, respectively, with the values in complementary strand B in better agreement with the equilibrium values.

The average helical twist for strand A in (I) is 36.26° and for strand B is 47.20° and for (II) is 36.03 and 45.68°, respectively. For (III) the values are 35.65 and 47.10°, and for (IV), 35.10 and 48.30° (Fig. 5). In all cases we note that strand B has relatively larger helical twist h than strand A. This overestimation is probably due to the different treatment of strands A and B since the sugars in strand B are initially disconnected. If the two strands are initially treated equivalently the same helical twist h will be obtained in both A and B strands. Another reason for this

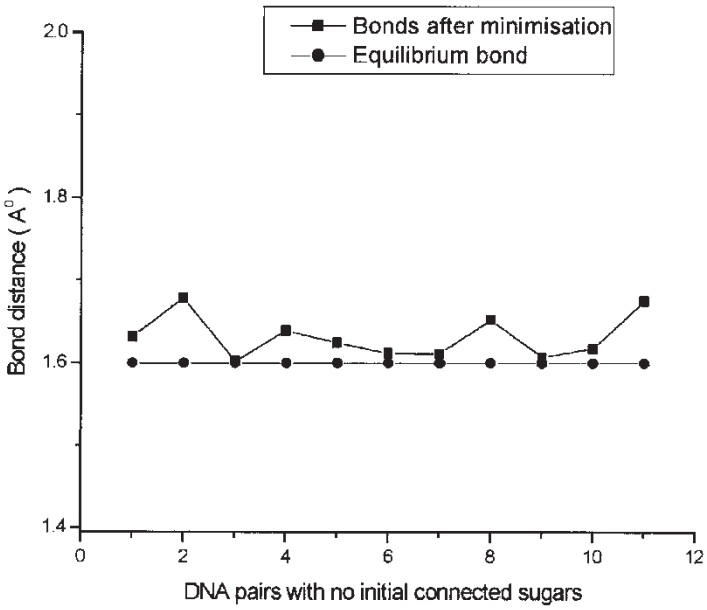

FIGURE 10 Values of the Bond Distance P–O in Angstroms, for the Dickerson Dodecamer Strand (squares), for the Final Generated Structure of the Initial Non-bonded Strand, Compared With the Equilibrium Distance (circles).

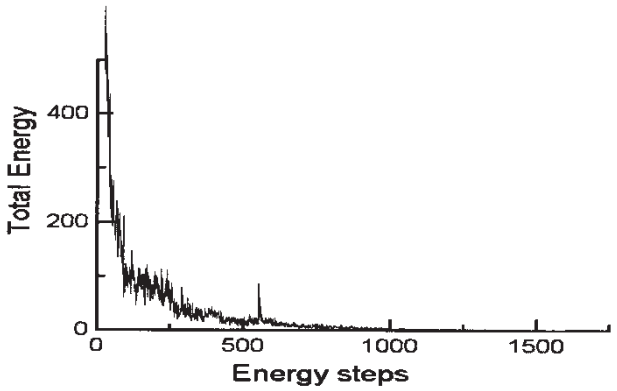

FIGURE 11 The Total Energy Progress During the Minimization of the Dickerson Dodecamer sequence.

discrepancy is due to the non-flexible hydrogen bonds and T8 torsions. All the structures contain the helical turn with the side chains almost parallel to one another. Small deviations exist only for the first and last base pairs as expected. Figure 13 shows the distribution of the first four torsion angles across the strand, of four B-DNA structures with different sequences. The first (I) (black line) contains a homogeneous 20 cc base pairs, the second (II) (red line) a random 20 base pairs, the third (III) (green line) a homogeneous 40 cc base pair sequence and the fourth (IV) a coding 40 base pair sequence. The fluctuations around the mean value are similar in all cases independently of the size of the sequence. The other three torsion angles T_5, T_6 and T_7 have similar distributions.

From comparison of the value of h, A_1 and A_2 between functionally different sequences (coding, non-coding, artificial random) no quantitative differences or similarities can be stressed at this level. It is possible that the discrepancy between the values of h, A_1 and A_2 in the two strands hide structural tendencies in homogeneous and non-homogeneous chains. Modified initial conditions for the strands may reveal differences in the structural characteristics of strands with different functionality.

CONCLUSIONS

In this study the folding of single and double stranded DNA molecules were studied, starting from an initial extended chain sequence. Our MC annealing method was shown to quickly fold the structure into recognizable helical shapes with helical twist angles close to its experimentally observed values. Our force field contained only a minimum of terms, indicating the importance of a small number of interactions in the folding process.

The MC simulated annealing method was used with parameters from the Dreiding force field [42] and showed that minimizing the energy of chains with different or the same nucleotides gives torsion and twist angles with a variety of values. The variation of the angles between adjacent nucleotides may be an artifact of the model due to the limited freedom that the strand has to move at this level (Non-flexible T8 torsion and Hydrogen bonds). For single strands of 20–40 base pairs, the structure contains twists and bends while double strands of similar length contain small bends independent of the sequence and the length of the structure. In the current study the initial structures had the sugars connected in strand A while in strand B the sugars were disconnected between oxygen and phosphate. For single strands it was found that the bond angles are $A_1 = 107 \pm 1°$ and $A_2 = 122 \pm 1°$, while the helical twist $h = 37.8 \pm 0.1$. For double DNA strands our model predicts the helical twist $h = 35.5 \pm 2°$ well in the strand A, while the prediction is less good $h = 47 \pm 2°$ in strand B. The average values of A_1 and A_2 was $130 \pm 1°$ and $150 \pm 3°$ for strand A and $102 \pm 4°$ and $123 \pm 5°$ for strand B. These larger deviations are due to the different treatment of the two strands.

The single stranded calculations obtained by the MC annealing method could account for RNA structures consisting of one strand, a linear arrangement of the bases adenine (A), uracil (U), cytosine (C)

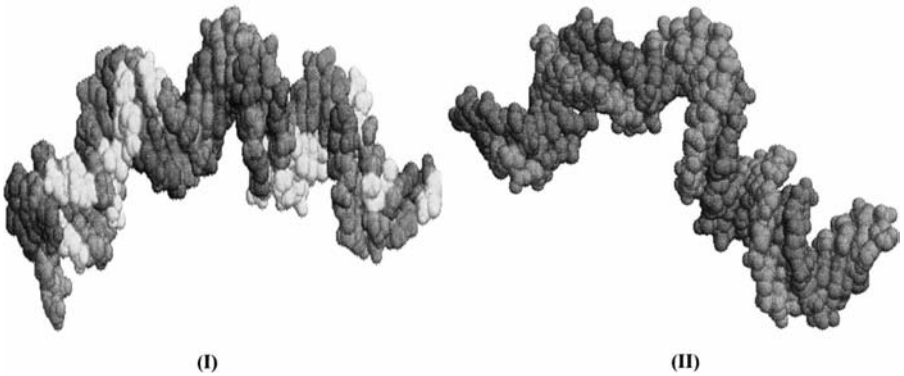

FIGURE 12 Two Final Double Stranded DNA Structures After the Minimization. Sequence (I) is the Homogeneous $DC_{40}dG_{40}$ And (II) is Coding, Heterogeneous and Contains 40 Base Pairs With the Sequence of Strand a As *Tgagaacga Aagctgcgc Gggaggttga Agaactgcgg*.

TABLE VII The Average Values for the Angles A_1 And A_2 Of the Four Structures I–IV

	I		*II*		*III*		*IV*	
Structure	*A1*	*A2*	*A1*	*A2*	*A1*	*A2*	*A1*	*A2*
Strand A	130.42°	150.74°	131.42°	151.62°	131.89°	152.35°	131.92°	150.72°
Strand B	99.64°	123.25°	104.48°	123.10°	02.90°	121.09°	104.03°	128.20°

and guanine (G). Thus the calculations included in this study may represent the dynamics of double DNA strands or of RNA strand where *T* is substituted for Uracil. At this level of approximation the geometrical difference between thymine and uracil is not important.

In the current version of the model many assumptions and restrictions were taken in order to construct the DNA model. Some of them were imposed in order to reduce the time of the minimization process and others to make the model simpler to handle. In later versions, a more accurate and realistic force field will be taken into account in order to better simulate the environment and the forces existing in the real DNA environment. Additional energy functions need to be added which should account for torsion and further angle changes, and for coulombic and hydrogen bond interactions. With these additions, the force field should mimic more closely the actual environmental forces.

In the case of double DNA strands, the initial configuration was such that the sugars were connected in strand A but they were disconnected in the complementary strand B. Due to that difference, deviations in h between the two strands were observed. Modifications of the initial conditions so that the two strands will start from similar conformations should lead to more accurate predictions of h as well as bond angle values A_1 and A_2 in both strands.

To be able to predict a more detailed structure of the DNA strand, more degrees of freedom in the DNA structure must be added by including more torsion dependent rotations. At the moment this model includes only two angle and seven torsion rotations leading to many side-chain—side-chain constraints. Due to computer limitations it was impossible to run the strands for longer periods of time increasing the number of intervals and stages and decreasing the temperature *T*. In the next version the algorithm should be able to predict Z-DNA structures and predict more accurately A- and B-DNA. Options for minimizing small organic molecules together with the DNA strands should be included together with the addition of a medium, e.g. solvent, such as water.

This algorithm attempts to predict a general structure of different types of DNA starting from a linear strand with very high energy. It shows that

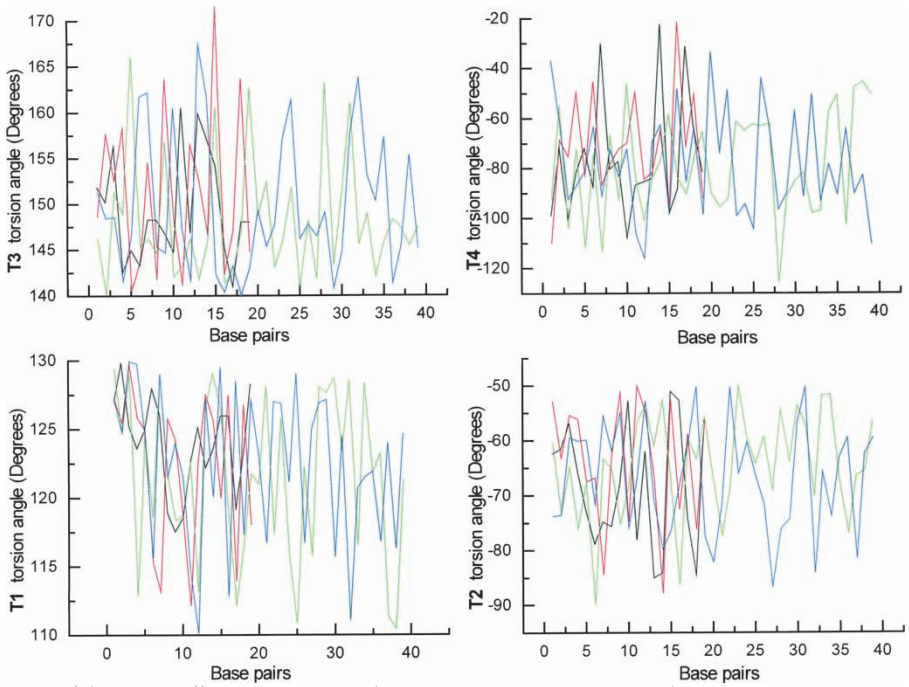

FIGURE 13 Distribution of the Four Different Torsion Angles in Degrees, Moving Across the Sequences in Fig. 12. They Are Taken from the Final Generated Structure of Strand a After the minimization. (Colour version available online.)

the most important factors for folding the DNA to the helical structure are the internal forces (hydrogen bond, angle, torsions, etc.…) leaving the environmental forces responsible for the interconnection of the bases (P–O bond) and the folding and unfolding of the DNA strand. At this stage it is not able to predict accurately large enough segments of DNA, answering the question, if a hetero or homogeneous strand is more curved. The later version of the algorithm might be used to predict the structure of DNA strands including small organic molecules.

Acknowledgements

This work is supported by a SIMU-ESF travel grant. The authors would like to thank Dr Y. Almirantis for his help in understanding the functionality of different DNA sequences and Dr N. Zacharopoulos for helpful discussions.

APPENDIX A

I. Initial Structures

TABLE AI The Values for the Bonds and Angles for the B-DNA Strand from the Pdb File 166D.pdb

		Angles (*degrees*)			
		Strand A		*Strand B*	
Pair Sequence	*Bonds* (*Å*) *Strand B*	*A1*	*A2*	*A1*	*A2*
cg	1.556	99.431	119.633	100.851	121.693
gc	1.608	91.351	121.134	95.327	119.034
ag	1.624	93.019	122.008	95.400	122.455
aa	1.607	98.744	118.724	99.589	125.903
ta	1.593	96.731	118.459	99.808	120.381
tt	1.609	100.776	117.534	100.214	116.611
ct	1.569	103.802	125.117	103.272	125.592
gc	1.621	97.097	126.144	102.447	123.011
cg	1.553	96.549	117.381	95.536	108.223
gc	1.663	100.903	122.330	100.351	127.826

II. Single Strand

TABLE AII The Starting and the Final Energy Values of the Minimized Structures for Case 3.1, B-DNA strands

		Initial Energy (Arbitrary Units)				*Final Energy (Arbitary Units)*			
Type	*––––––10––––––20––––––30––––––40*	*Total*	*VDW*	*ANG*	*BND*	*Total*	*VDW*	*ANG*	*BND*
Homogeneous	ccccc	78312.9	5.188	78307.8	0.0	3.528	3.368	0.159	0.0
Dyadic	cgcgcgcgcg	78323.5	15.748	78307.8	0.0	4.942	4.612	0.330	0.0
Dyadic	cgcgcgcgcgcgcgc	78324.7	16.957	78307.8	0.0	4.032	3.745	0.287	0.0
Real––––––	cgcgaattcgcg	78322.6	14.863	78307.8	0.0	3.924	2.623	0.301	0.0
Homogeneous	ccccccccc Ccccccccc	78312.9	5.188	78307.8	0.0	4.445	4.258	0.188	0.0
Homogeneous	ccccccccc Ccccccccc Ccccc	78312.9	5.188	78307.8	0.0	4.079	3.821	0.258	0.0
Homogeneous	ccccccccc Ccccccccc Ccccccccc	78312.9	5.188	78307.8	0.0	3.709	3.353	0.356	0.0
Homogeneous	ccccccccc Ccccccccc Ccccccccc Ccccccccc	78312.9	5.188	78307.8	0.0	4.064	3.647	0.417	0.0
Real-coding	tgagaacgaa Aagctgcgcc	78323.6	15.813	78307.8	0.0	4.268	3.960	0.308	0.0
Random	tcaaaatgggc Cccgaaatcg	78320.9	13.111	78307.8	0.0	3.687	3.491	0.196	0.0
Real-coding	actgataccg Gggttgctga	78331.4	23.665	78307.8	0.0	5.225	4.967	0.258	0.0
Real-coding	tgagaacgaa Aagctgcgcc Gggaggttga Agaactgcgg	78328.3	20.552	78307.8	0.0	3.902	3.450	0.452	0.0

III. Double Strand

TABLE AIII The Starting and the Final Energy Values of the Minimized Double B-DNA strands

		Initial Energy (Arbitrary Units)				*Final Energy (Arbitary Units)*			
Type	*−−−−−−10−−−−−−20−−−−−−30−−−−−−40*	*Total*	*VDW*	*ANG*	*BND*	*Total*	*VDW*	*ANG*	*BND*
Homogeneous	cc	8277.7	0.001	0.856	8276.8	1.169	0.608	0.512	0.049
Homogeneous	aa	8316.3	38.6	0.856	8276.8	1.200	0.600	0.500	0.000
Real	cg	8277.7	0.014	0.856	8276.8	1.178	0.630	0.508	0.041
Real	ta	8316.3	38.6	0.856	8276.8	1.180	0.668	0.506	0.007
Real	ag	8277.9	0.014	0.856	8276.8	1.260	0.700	0.500	0.000
Real	aatt	8290.5	12.8	0.856	8276.8	1.900	0.400	0.500	1.800
Homogeneous	ggggg	8277.6	0.006	0.856	8276.8	1.089	0.596	0.438	0.054
Homogeneous	ccccc	8277.6	0.006	0.856	8276.8	1.101	0.623	0.426	0.053
Dyadic	cgcgcgcgcg	8277.7	0.007	0.856	8276.8	1.080	0.614	0.413	0.054
Dyadic	cgcgcgcgcgcgcgc	8277.7	0.007	0.856	8276.8	1.130	0.636	0.452	0.052
Real	cgcgaattcgcg	8284.7	7.025	0.856	8276.8	1.308	0.654	0.488	0.166
Homogeneous	ccccccccc Ccccccccc	8277.7	0.001	0.856	8276.8	1.400	0.700	0.500	0.200
Homogeneous	ccccccccc Ccccccccc Ccccc	8277.7	0.001	0.856	8276.8	1.700	1.000	0.500	0.300
Homogeneous	ccccccccc Ccccccccc Ccccccccc	8277.7	0.001	0.856	8276.8	1.700	0.800	0.500	0.300
Homogeneous	ccccccccc Ccccccccc Ccccccccc Ccccccccc	8277.7	0.001	0.856	8276.8	1.600	0.800	0.500	0.300
Real-coding	tgagaacgaa Aagctgcgcc	8291.9	14.228	0.856	8276.8	1.300	0.700	0.500	0.200
Random	tcaaatgggc Cccgaaatcg	8289.9	12.195	0.856	8276.8	1.500	0.700	0.500	0.300
Real-coding	actgataccg Gggttgctga	8283.8	6.101	0.856	8276.8	1.500	0.600	0.500	0.300
Real-coding	tgagaacgaa Aagctgcgcc Gggaggttga Agaactgcgg	8289.6	11.885	0.856	8276.8	1.800	0.900	0.500	0.300

References

[1] Albert, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1994) *Molecular Biology of the Cell*, 3rd Ed. (Garland Publishing, New York).

[2] Almirantis, Y. and Provota, A. (1999) "Long- and short-range correlations in genome organization", *J. Stat. Phys.* **97**(1/2), 233–262.

[3] Schlick, T., Bread, D.A., Huang, J., Strahs, D.A. and Qian, X. (2000) "Computational challenges in simulating large DNA over long times", *Comput. Sci. Eng.*, 38–51.

[4] Ayadi, L., Coulombeau, C. and Lavery, R. (1999) "Abasic sites in duplex DNA: molecular modeling of sequence-dependent effects on conformation", *Biophys. J.* **77**, 63218–63226.

[5] Bostock, L. and Chandler, S. (1979) *Pure Mathematics 2* (Stanley Thornes, U.K.).

[6] Creighton, T.E. (1993) *Structures and Molecular Properties*, 2nd Ed. (Proteins, W.H. Freeman, New York).

[7] Goodfellow, J.M. and Moss, D.S. (1992) *Computer Modeling of Bimolecular Processes* (Ellis Horwood, UK).

[8] Bostock, L. and Chandler, S. (1978) *Pure Mathematics 1* (Stanley Thornes, U.K.).

[9] Louise-May, S., Auffinger, P. and Westhof, E. (1996) "Calculations of nucleic acid conformations", *Curr. Opin. Struct. Biol.* **6**, 289–298.

[10] Swamminathan, S., Ravishaanker, G. and Beveridge, D.L. (1991) "Molecular dynamics of B-DNA including water and counterions: a 140-ps trajectory for d(CGCGAATTCGCG) based on the GROMOS force field", *J. Am. Chem. Soc.* **113**, 5027–5040.

[11] Drukker, K. and Schatz, G.C. (2000) "A model for dynamics of DNA denaturation", *J. Phys. Chem. B* **104**(26), 6108–6111.

[12] Bruant, N., Flatters, D., Lavery, R. and Genest, D. (1999) "From atomic to mesoscopic descriptions of the internal dynamics of DNA", *Biophys. J.* **77**(5), 2366–2376.

[13] Jian, H. and Vologodski, A.V. (1997) "A combined wormlike-chain and bead model for dynamics simulations of long linear DNA", *J. Comput. Phys.* **136**, 168–179.

[14] Lavery, R. and Lebrun, A. (1999) "Modelling DNA stretching for physics and biology", *Genetica* **106**(1–2), 75–84.

[15] Olson, W.K. and Zhurkin, V.B. (2000) "Modeling DNA deformations", *Curr. Opin. Struct. Biol.* **10**(3), 286–297.

[16] Vlahovicek, K. and Pongor, S. (2000) "Model.it. building three dimensional DNA models from sequence data", *Bioinformatics* **16**(11), 1044–1045.

[17] Treger, M. and Westhof, E. (1987) "An interactive modeling program for DNA", *J. Mol. Graph.* **5**, 178–183.

[18] Srinivasan, A.R. and Olson, W.K. (1987) "Nucleic acid model building: the multiple backbone solutions associated with a given base morphology", *J. Biomol. Struct. Dyn.* **4**, 895–938.

[19] Schlick, T. (1995) "Modeling superhelical DNA: recent analytical and dynamic approaches", *Curr. Opin. Struct. Biol.* **5**, 245–262.

[20] MacKerrell, Jr. A.D., Banavali, N. and Foloppe, N. (2000) "Development and current status of the CHARM force field for nucleic acids", *Biopolymers* **56**, 257–265.

[21] Schneider, B., Neidle, S. and Berman, H.M. (1997) "Conformations of sugar-phosphate in helical DNA crystal structures", *Biopolymers* **42**, 113–124.

[22] Duan, Y., Wilkosz, P. and Crowley, M. (1997) "Molecular dynamics study of DNA dodecamer d(CGCGAATTCGCG)", In: Rosenberg, J.M., ed, *Solution: Conformation and Hydration* Journal of Molecular Biology, **272**, (4), pp 553–572.

[23] Sherer, E.C., Harris, S.A., Soliva, R., Orozco, M. and Laughton, C.A. (1999) "Molecular dynamics studies of DNA A-tract structure and flexibility", *J. Am. Chem. Soc.* **121**, 5981–5991.

[24] Ezaz-Nikpay, K. and Verdine, G.L. (1992) "Aberrantly methylated DNA: site-specific introduction of N7-Methyl-2′-deoxyguanosine into the Dickerson/Drew dodecamer", *J. Am. Chem. Soc.* **114**, 6562–6563.

[25] Tereshko, V., Minasov, G. and Egli, M. (1999) "The Dickerson-Drew B-DNA dodecamer revisited at atomic resolution", *J. Am. Chem. Soc.* **121**, 470–471.

[26] Tereshko, V., Minasov, G. and Egli, M. (1999) "Additions and corrections", *J. Am. Chem. Soc.* **121**(29), 6970.

[27] Mazur, A.K. (2001) "Molecular dynamics of minimal B-DNA", *J. Comput. Chem.* **22**(4), 457–467.

[28] Westcott, T.P., Tobias, I. and Olson, W.K. (1997) "Modeling self-contact forces in the elastic theory of DNA supercoiling", *J. Chem. Phys.* **107**(10), 3967–3980.

[29] Ramachandran, G. and Scchlick, T. (1995) "Solvent effects on supercoiled DNA dynamics explored by Langevin dynamics simulations", *Phys. Rev. E* **51**(6), 6188–6203.

[30] Schlick, T. and Olson, W.K. (1992) "Supercoiled DNA energetics and dynamics by computer simulation", *J. Mol. Biol.* **223**, 1089–1119.

[31] Schlick, T. and Olson, W.K. (1992) "Trefoil knotting revealed by molecular dynamics simulations of supercoiled DNA", *Science* **257**, 1110–1115.

[32] Lavery, R., Zakrzewska, K. and Sklenar, H. (1995) "JUMNA (Junction minimisation of nucleic acids)", *Comp. Phys. Commun.* **91**, 135–158.

[33] Lafontaine, L. and Lavery, R. (2000) "Optimization of nucleic acid sequences", *Biophys. J.* **79**(2), 680–685.

[34] Gabb, H.A., Prevost, C., Bertucat, G., Robert, C.H. and Lavery, R. (1997) "Collective-variable Monte Carlo simulation of DNA", *J. Comput. Chem.* **18**(16), 2001–2011.

[35] Provata, A., Tsakiroglou, M. and Almirantis, Y. "Non randomness and non linearities in DNA". *Proceedings to 1st Interdisciplinary Symposium on Non-Linear Problems Athens*, January (2000).

[36] Olson, W.K., Srinivasan, A.R., Hao, H.H. and Nauss, J.L. (1988) "Structure and expression 3", In: Olson, W.K., Sarma, M.H., Sarma, R.H. and Sundaralingman, M., eds, *DNA Bending and Curvature* (Adenine Press, New York), pp 225–242.

[37] Schurr, J.M. (1985) "Effect of anisotropic bending rigidity and finite twisting rigidity on statistical properties of DNA model filaments", *Biopolymers* **24**, 1233–1246.

[38] Selepova, P. and Kypr, J. (1985) "Computer simulation of DNA supercoiling in a simple elastomechanical approximation", *Biopolymers* **24**, 867–882.

[39] Shimada, J. and Yamakawa, H. (1985) "Statistical mechanics of DNA topoisomers. The helical worm-like chain", *J. Mol. Biol.* **184**, 319–329.

[40] Tsuru, H. and Wadati, M. (1986) "Elastic model of highly supercoiled DNA", *Biopolymers* **25**, 2083–2096.

[41] Tan, R.K. and Harvey, S.C. (1989) "Molecular mechanics model of supercoiled DNA", *J. Mol. Biol.* **205**, 573–591.

[42] Mayo, S.L., Olafson, B.D. and Goddard, W.A. (1990) "DREIDING: a force field for molecular simulations", *J. Phys. Chem.* **94**, 8897–8909.

[43] Sayle, R. Rasmol home page (1994) World Wide Web: http://www.rasmol.com.

[44] Guex, N. and Peitsch, M.C. Swiss-PdbViewer, v3.6b3, (1997) Glaxo Wellcome Experimental Research (http://www.expasy.ch/s-pdbv/mainpage.html).

[45] Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992) *Numerical Recipes in Fortran. The Art of Scientific Computing*, 2nd Ed. (Cambridge University Press, Cambridge).

[46] Voet, D. and Voet, J.G. (1995) *Biochemistry,* 2nd Ed. (Wiley, New York).

[47] Zimmerman, S.B. (1982) "The three-dimensional structure of DNA", *Ann. Rev. Biochem.* **51**, 395–427.

[48] Nunn, C.M., Jenkins, T.C. and Neidle, S.E. (1994) "Crystal structure of gamma-oxapentamide complexed with d(CGCGAATTCGCG)$_2$. The effects of drug structural change on DNA minor-groove recognition", *Eur. J. Biochem.* **226**, 953–961.

[49] Protein Data Bank Contents Guide. Atomic Co-ordinates Entry Format Description, World Wide Web: http://www.rcsb.org.

[50] Denisov, A.Y., Zamaratski, E.V., Maltseva, T.V., Sandstrom, A., Bekiroglou, S., Altmann, K.-H., Egli, M. and Chattopadhyaya, J. (1998) "The solution conformation of a carbocyclic analog of the Dickerson-Drew dodecamer: comparison with its own X-ray structure and that of the NMR structure of the native counterpart", *J. Biomol. Struct. Dyn.* **16**(3), 547–568.

[51] Maltseva, T.V., Altmann, K.-H., Egli, M. and Chattopadhyaya, J. (1998) "The residence time of the bound water in the

hydrophobic minor groove of the carbocyclic-nucleoside analogs of Dickerson-Drew dodecamers", *J. Biomol. Struct. Dyn.* **16**(3), 569–578.

[52] Ikeda, H., Fermandez, R., Wilk, A., Barchi, Jr., J.J., Huang, X. and Marquez, V.E. (1998) "The effect of two antipodal fluorine-induced sugar puckers on the conformation and stability of the Dickerson-Drew dodecamer duplex [d(CGC-GAATTCGCG)]$_2$", *Nucleic Acid Res.* **26**(9), 2237–2244.

[53] Gelasco, A. and Lippard, S.J. (1998) "NMR solution structure of a DNA Dodecamer Duplex containing a *cis*-Diammineplatinum(II) d(GpG) intrastrand cross-link, the major adduct of the anticancer drug cisplatin", *Biochemistry* **37**(26), 9230–9239.

[54] IUPAC: International Union of Pure and Applied Chemistry, World Wide Web: http://www.chem.qmul.ac.uk/iupac.

[55] Gillespie, R.J., Hargittai, I. (1991) *The VSEPR Model of Molecular Geometry* (Allyn and Bacon, Boston, MA).