Journal of Biomolecular NMR, 9 (1997) 25–34
ESCOM

J-Bio NMR 408

Application of a genetic algorithm in the conformational analysis of methylene-acetal-linked thymine dimers in DNA: Comparison with distance geometry calculations

Mischa L.M. Beckers[a], Lutgarde M.C. Buydens[a,*], Jeroen A. Pikkemaat[b,**] and Cornelis Altona[b]

[a]*Laboratory for Analytical Chemistry, Catholic University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands*
[b]*Leiden Institute of Chemistry, Gorlaeus Laboratories, P.O. Box 9502, 2300 RA Leiden, The Netherlands*

Received 26 August 1996
Accepted 16 October 1996

Keywords: Genetic algorithm; Distance geometry; Methylene-acetal linkage

Summary

The three-dimensional spatial structure of a methylene-acetal-linked thymine dimer present in a 10 base-pair (bp) sense–antisense DNA duplex was studied with a genetic algorithm designed to interpret NOE distance restraints. Trial solutions were represented by torsion angles. This means that bond angles for the dimer trial structures are kept fixed during the genetic algorithm optimization. Bond angle values were extracted from a 10 bp sense–antisense duplex model that was subjected to energy minimization by means of a modified AMBER force field. A set of 63 proton–proton distance restraints defining the methylene-acetal-linked thymine dimer was available. The genetic algorithm minimizes the difference between distances in the trial structures and distance restraints. A large conformational search space could be covered in the genetic algorithm optimization by allowing a wide range of torsion angles. The genetic algorithm optimization in all cases led to one family of structures. This family of the methylene-acetal-linked thymine dimer in the duplex differs from the family that was suggested from distance geometry calculations. It is demonstrated that the bond angle geometry around the methylene-acetal linkage plays an important role in the optimization.

Introduction

Genetic algorithms belong to the class of global optimization algorithms (Holland, 1975; Goldberg, 1989). A population of trial solutions is iteratively manipulated by a series of genetic operators, such as selective reproduction, recombination and mutation, to satisfy an objective function. These algorithms receive more and more attention in the field of conformational analysis of biomacromolecules, such as proteins and nucleic acids. Basically, two paths are followed. In the first one the genetic algorithm search is guided by an energy criterion supplied by an implemented molecular force field (McGarrah and Judson, 1993; Brodmeier and Pretsch, 1994; Sun, 1995). The second path directs the search with the use of experimental data (Blommers et al., 1992; van Kampen et al., 1996). Usually, these experimental data are based on nuclear magnetic resonance experiments (Wüthrich, 1986). One of the most widely used techniques in determining the three-dimensional structure of a molecule is multidimensional nuclear Overhauser enhancement (NOE) spectroscopy (Jeener et al., 1979; Macura and Ernst, 1980; Macura et al., 1981). NOE peaks provide information about the spatial arrangement of protons in the molecule.

In this paper, the conformational analysis of a thymine dimer containing a methylene-acetal linkage, O3'-CH_2-O5', instead of the regular phosphodiester linkage, O3'-PO_2^--O5', is presented. Methylene-acetal-linked nucleotides provide interesting test cases for conformational analysis techniques, since their backbone conformation is relatively well defined, owing to the additional NOEs of the methylene-acetal protons. The initial interest in the methylene-acetal linkage was focussed on its potential application as an antisense DNA oligonucleotide to in-

*To whom correspondence should be addressed.
**Present address: Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, U.S.A.

0925-2738/$ 6.00 + 1.00 © 1997 ESCOM Science Publishers B.V.

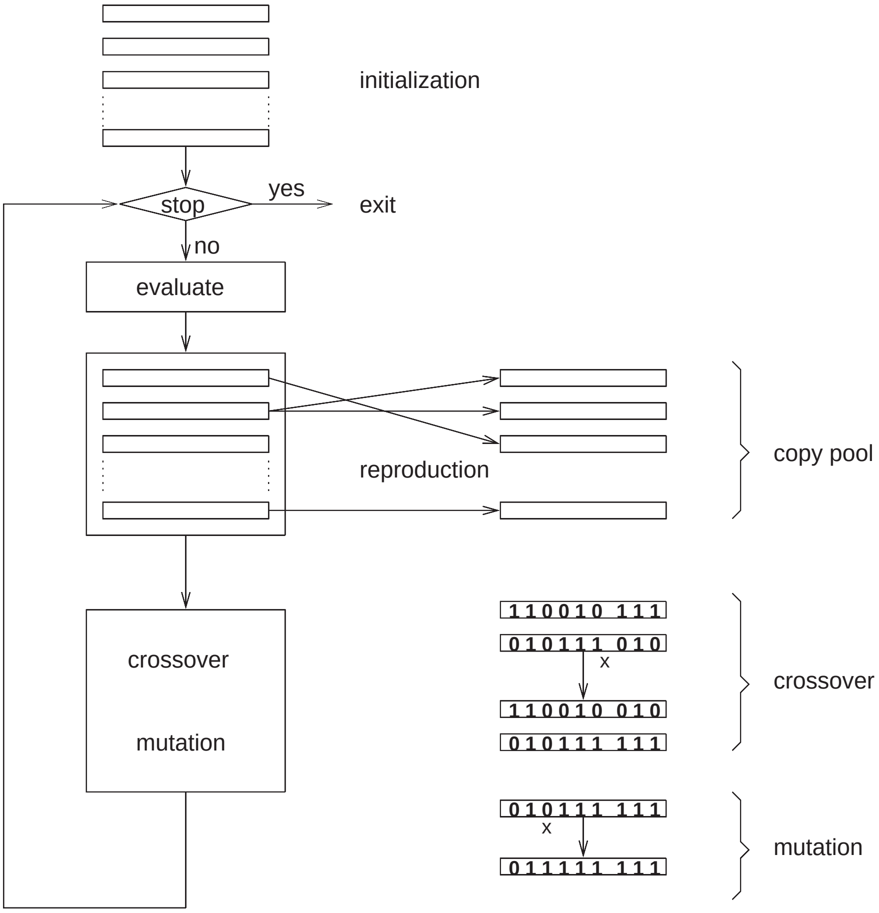

Fig. 1. Flow chart of a genetic algorithm. In the crossover block, x indicates a breakpoint. Parts of the bit strings after this point (one-point crossover) are exchanged. In the mutation block, x denotes the bit that will be changed. For the sake of simplicity, these basic forms of crossover and mutation are depicted in the figure.

hibit the expression of selective genes. In order to study the affinity of the antisense nucleotide for the sense (unmodified complementary) nucleotide, the methylene-acetal-linked thymine dimer, T^T, was built in decamer duplexes. Comparative NMR studies of the modified duplexes and the corresponding unmodified duplex suggested regular B-DNA structures (Gao et al., 1992; Quaedflieg et al., 1993). The two T^T dimers in the modified duplex 5'-d(GCGT^TTT^TGCG)•d(CGCAAAACGC)-3' were also studied separately and in more detail. For the T4^T5 dimer, 63 proton–proton distance restraints were used in a distance geometry calculation (Havel and Wüthrich, 1984). The structures that were found could be classified in one family. On the basis of torsion angles, it was concluded that this family occurred in a regular B(I)-DNA conformation. However, the ε, ζ and β torsion angles were somewhat biased towards the less common B(II)-DNA conformation. The indication of DNA families mentioned in this study corresponds with the definition by Privé et al. (1987), i.e., B(I) has ε (t) and ζ (g^-), B(II) has ε (g^-) and ζ (t) and, in addition, the β of B(II) is somewhat smaller than the β of B(I).

In this study a genetic algorithm is used to minimize the violations of the available distance restraints. The variables to be optimized, in this case torsion angles, are allowed to vary in a wide range in order to get a good impression of the conformational space that is spanned by the distance restraints. These ranges cover all B-DNA-type rotamers, as well as other, less common, rotamers. In the genetic algorithm optimization, bond angle geometries can be kept fixed, while in the distance geometry calculations the embed algorithm can distort these geometries. It is demonstrated that the bond angle geometry has an important effect on the resulting conformation.

Materials and Methods

Genetic algorithms

In genetic algorithm optimization, trial solutions are encoded on bit strings. The parameters to be optimized are assigned to bit fields on the bit string. The first stage of a genetic algorithm run is the initialization. A population of randomly initiated trial solutions is created. The parameters on the bit string receive a value between the lower and upper bound of the allowed range. The initia-

TABLE 1
BOND ANGLES (°) OF THE METHYLENE-ACETAL-LINKED THYMINE DIMER DETERMINED IN VARIOUS WAYS

Bond angle	GA	DG	DGII	DGII (10%)
C3'-O3'-CM	116.3	120.5 (1.1)	115.6 (0.3)	112.5 (4.4)
O3'-CM-O5'	110.4	108.5 (0.7)	109.6 (0.1)	111.1 (2.4)
O3'-CM-HMA	111.7	118.0 (0.5)	123.5 (1.1)	108.1 (4.6)
O3'-CM-HMB	108.3	102.6 (0.7)	105.2 (0.3)	111.3 (7.0)
HMA-CM-HMB	108.1	107.0 (0.6)	103.8 (0.5)	107.4 (4.4)
O5'-CM-HMA	109.6	105.9 (1.0)	104.3 (0.4)	111.1 (6.3)
O5'-CM-HMB	109.2	115.4 (1.0)	110.0 (0.1)	107.9 (4.3)
CM-O5'-C5'	112.4	114.3 (0.9)	108.0 (0.2)	111.2 (4.3)
O5'-C5'-H5'	110.4	106.0 (1.0)	109.0 (0.2)	107.8 (2.8)
O5'-C5'-H5"	110.8	111.5 (0.9)	109.2 (0.2)	110.7 (4.9)
H5'-C5'-H5"	109.0	109.2 (1.0)	109.3 (0.2)	107.3 (2.2)
O5'-C5'-C4'	110.4	111.5 (1.0)	109.6 (0.1)	109.7 (2.1)

GA: genetic algorithm; DG and DGII: average bond angles (with the standard deviation in parentheses) of the 10 best structures resulting from initial distance geometry calculations and from calculations with the second-generation distance geometry package; DG (10%): the results of DGII calculations for the restraints set in which the restraints were relaxed by 10%.

lization stage is led by the random seed value of a random generator. Each of the bit strings in the population receives a quality value which, in genetic algorithm terminology, is called fitness. This is the evaluation stage. The fitness, which should be maximized, is calculated in an objective function. The next stage in the optimization is selection. Here a new population, called a copy pool, is created by allowing only strings that fulfill a certain selection criterion. Usually, the probability of a bit string to be selected is proportional to its fitness. Selection *exploits* the information content of bit strings. As soon as the copy pool contains the same number of bit strings as the original population, the bit strings in the copy pool are subjected to the crossover operator. This operator exchanges parts of bit strings or bit fields between (randomly) selected pairs of bit strings. Crossover takes place with a certain probability. By spreading high-quality parts of bit strings through the population, important information is preserved. Crossover is followed by mutation, which swaps the value of single bits with a certain probability. By mutating bit strings, new information can be introduced in the population. Hence, crossover and mutation *explore* the information that is present in the search space. After crossover and mutation the copy pool replaces the initial population. The new population can be subjected to a new cycle of evaluation, selection, crossover and mutation. Such a cycle is called a generation in genetic algorithm terminology. Usually, in genetic algorithm

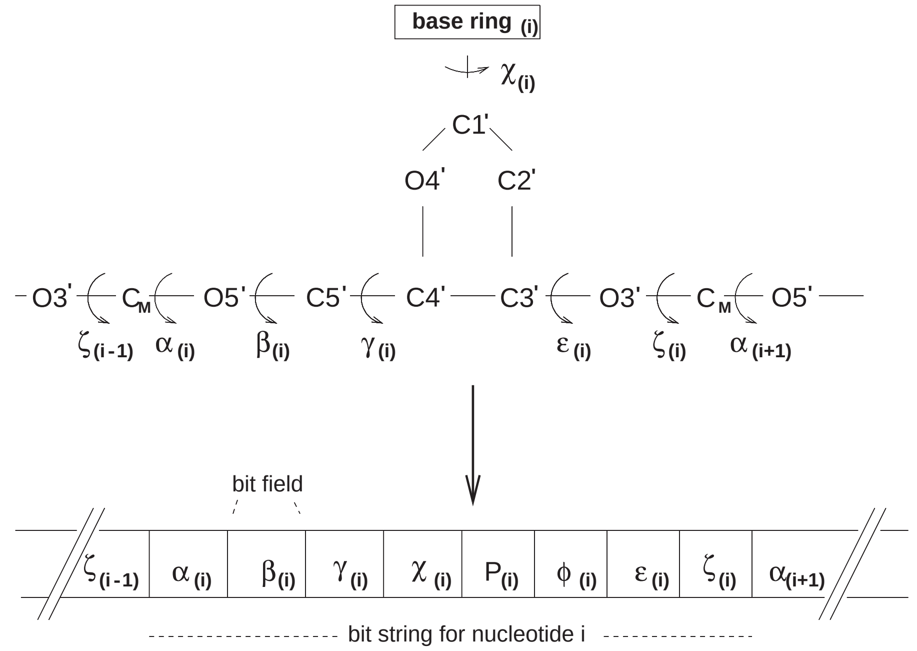

Fig. 2. Torsion angle representation of a methylene-acetal-linked thymine dimer. The protons attached to the carbon atoms are not shown. CM is the carbon that replaces the original phosphorus atom. CM has two protons (HMA and HMB) attached to it.

TABLE 2
PROTON–PROTON DISTANCE RESTRAINTS FOR THE METHYLENE-ACETAL-LINKED THYMINE DIMER

Proton 1	Proton 2	Upper bound (Å)	Lower bound (Å)
T4 H6	T5 H6	4.549	9.990
T5 HB	T4 H6	3.876	9.990
T4 H3'	T4 H6	3.263	3.465
T4 H5'	T4 H6	3.775	6.011
T4 H4'	T4 H6	3.640	7.431
T4 H5"	T4 H6	3.538	4.320
T4 H2"	T4 H6	3.101	3.566
T4 H2'	T4 H6	2.327	2.618
T4 H6	T5 H7	3.189	3.446
T4 H6	T4 H7	2.824	2.910
T4 H1'	T5 H6	3.356	3.976
T5 HB	T5 H6	3.919	4.023
T4 H3'	T5 H6	3.639	4.193
T5 H4'	T5 H6	3.979	4.913
T5 H5'	T5 H6	3.938	4.936
T5 H5"	T5 H6	4.017	4.835
T4 H2"	T5 H6	2.463	2.750
T5 H6	T5 H2"	2.796	3.089
T5 H6	T5 H2'	2.234	2.513
T4 H2'	T5 H6	2.940	3.154
T5 H6	T5 H7	2.967	3.071
T5 HB	T4 H1'	3.566	9.990
T5 HA	T4 H1'	3.547	8.958
T4 H1'	T4 H3'	3.223	4.036
T4 H1'	T5 H5"	3.808	4.403
T4 H1'	T4 H2"	2.471	2.647
T4 H1'	T4 H2'	2.892	2.986
T4 H1'	T5 H7	3.898	4.985
T5 H1'	T5 H3'	3.661	4.675
T5 H1'	T5 H2"	2.473	2.587
T5 H1'	T5 H2'	2.898	3.033
T5 H1'	T5 H7	4.815	6.082
T5 HB	T4 H3'	2.351	2.619
T2 HA	T4 H3'	2.869	2.932
T2 H4'	T5 H5"	2.331	2.673
T4 H4'	T4 H2'	3.243	4.865
T4 H3'	T4 H5'	2.989	9.990
T4 H3'	T4 H4'	2.757	2.854
T4 H3'	T4 H5"	2.660	2.822
T4 H3'	T5 H5"	4.248	4.335
T4 H3'	T4 H2"	2.664	3.000
T4 H3'	T4 H2'	2.452	2.658
T4 H3'	T5 H7	3.878	4.919
T4 H3'	T4 H7	4.943	9.990
T5 HB	T5 H5'	3.293	3.512
T5 HA	T5 H5"	2.568	2.638
T5 HB	T5 H5"	2.987	3.279
T5 HB	T4 H2'	3.326	4.181
T5 HB	T5 H7	4.590	9.990
T4 H7	T4 H2"	4.603	6.202
T5 H7	T4 H2"	3.153	3.498
T5 H7	T5 H2"	5.319	9.990
T5 H7	T5 H2'	4.850	9.990
T5 H7	T4 H2'	3.035	3.070
T4 H7	T4 H2'	3.835	5.958
T5 H7	T4 H7	4.371	9.990
T5 H3	T5 H3	3.000	5.000
T5 HB	T5 H2"	4.000	20.000
T5 HB	T5 H2'	4.000	20.000
T5 HB	T4 H2"	3.500	7.000
T5 HB	T4 H2'	2.700	6.000
T5 HA	T4 H2"	3.500	7.000
T5 HA	T4 H2'	2.700	6.000

optimization the iteration through generations is done until no further improvement of the trial solutions is observed. Figure 1 shows a flow chart of a genetic algorithm.

Representation

In the present study trial structures for the genetic algorithm are represented by torsion angles. This means that bond lengths and bond angles are kept fixed. The backbone is represented by the torsion angles α, β, γ, ε and ζ. The sugar ring is represented by the pucker amplitude ϕ and a pucker phase angle P (Altona and Sundaralingam, 1972; De Leeuw et al., 1980). The orientation of the base ring with respect to the sugar is given by the torsion angle χ. Hence, a single nucleotide is represented by eight variables. These are the parameters to be optimized by means of the genetic algorithm (see Fig. 2).

Evaluation

In the present study the objective function takes the form

$$V_{ij} = \begin{cases} \dfrac{(lb_{ij} - r_{ij})^2}{(lb_{ij})^2} & \text{when } r_{ij} < lb_{ij} \\ 0 & \text{when } lb_{ij} \le r_{ij} \le ub_{ij} \\ \dfrac{(r_{ij} - ub_{ij})^2}{(ub_{ij})^2} & \text{when } r_{ij} > ub_{ij} \end{cases} \tag{1}$$

$$\text{rmsd} = \sqrt{\frac{\sum V_{ij}}{N}} \tag{2}$$

where r_{ij} is the distance between protons i and j in the trial structure, lb_{ij} is the lower bound of the proton–proton distance restraint, ub_{ij} is the upper bound of the proton–proton distance restraint, rmsd is the root-mean-square difference and N is the number of proton–proton distance restraints. This objective function must be minimized. Hence, the fitness is the reciprocal of the objective function.

Implementation

The genetic algorithm used in this study was developed

TABLE 3

CONFIGURATIONAL SETTINGS FOR THE GENETIC ALGORITHM OPTIMIZATION OF THE METHYLENE-ACETAL-LINKED THYMINE DIMER

Torsion angle ranges (°)	
ε	160–270
ζ	30–330
α	30–330
β	120–240
γ	20–100
χ	90–270
ϕ	32–44
P	100–200
Population	
Size	100
Fitness scaling	
Mode	Linear static
Fitness offset	0.0
Scale factor	1.01
Selection	
Mode	Threshold
Elitist fraction	0.05
Threshold fraction	0.25
Crossover	
Mode	Uniform
Probability	0.90
Swaps	0.16
Mutation	
Mode	Distributed
Probability	0.04

with the toolbox GATES (Genetic Algorithm Toolbox for Evolutionary Search) (Lucasius and Kateman, 1993, 1994a,b). A large variety of genetic operators are available in GATES. The parsing procedure from torsion angles to atomic coordinates (constant bond lengths and bond angles), that is needed for the objective function, was taken from the DENISE (Dna Evolutionary Noe Interpretation system for Structure Elucidation) program (Lucasius et al., 1991). This procedure was adapted to allow a methylene-acetal linkage in the thymine dimer. Equilibrium bond lengths and bond angles are well defined in most force fields or literature on nucleic acids. However, these values are to be used explicitly with their specific force constants and force field. Therefore, they cannot be used directly to constrain the geometry of the modified thymine dimer trial structures in a genetic algorithm optimization. It is possible to subject the complete 5'-d(GCGT^TT^TGCG)•d(CGCAAAACGC)-3' duplex to energy minimization by means of a force field and extract bond length and bond angle values from the minimized structure. Not all the bond angles in the methylene-acetal moiety are defined in the literature. A reasonable approximation of these bond angles can be deduced by an energy minimization of the complete duplex by means of a modified force field. We used the AMBER force field (Weiner et al., 1986) with additional parameters for the methylene-acetal moiety in the energy minimization of the duplex. The bond angles around the methylene-acetal linkage were measured in the minimized structure. They are summarized in column 2 of Table 1. Because the bond lengths did not differ significantly from standard values, they are not indicated in Table 1.

TABLE 4

CONFIGURATIONAL SETTINGS FOR THE DGII CALCULATIONS OF THE METHYLENE-ACETAL-LINKED THYMINE DIMER

Smooth	
Triangle smoothing	On
Triangle violation tolerance	0.01
Tetrangle strategy	None
Embed	
Uniform probability density	On
Probability coefficient	0.5
Eigenvalue iteration	100
Eigenvalue iteration	0.001
Metrization	Prospective
Embed dimension	4
Majorize	
Guttman transform	10
Linear conjugate gradient transform	100
Linear conjugate gradient criterion	0.001
Scale centroid	Off
Calculate Moore–Penrose inverse	On
Moore–Penrose inversion criterion	0.001
Weighting scheme	Constant
Overwrite structures	On
Optimize	
Dimension weight	0.20
Chirality weight	0.1
Lower maximum	10.0
Contact maximum	1.00
Dimension scaling	0.30
Upper weight limit	1.00
Error function form	Full matrix
Extra radii	1.00
Simulated annealing	
Initial temperature	1.00
Maximum heating	2.00
Maximum number of steps	500
Calculate initial energy	Off
Initial energy	1000.0
Maximum temperature	200.0
Fail level	1.00
Atom mass	1000
Step size	2e–13
Conjugate gradient	
Maximum iterations	250
Rms gradient	0.001
Global setup	
Generate database	On
Number of structures	75
Omega wobble	10
Increment files	On

Data set

The NOE buildup data obtained for the duplex were analyzed by means of an iterative relaxation matrix approach (IRMA) (Boelens et al., 1988,1989). Experimental NOE data from a series of NOESY spectra taken at mixing times of 50, 75, 105 and 145 ms were included for approximately 300 proton–proton pairs. The coordinates of a structure with B-DNA model geometry that was subjected to energy minimization with the all-atom version of the AMBER force field were used as a starting structure in the first IRMA cycle. The resulting upper and lower bounds were relaxed by 5% to allow for various sources of errors in the distance determinations. With the use of distance geometry calculations, 50 candidate structures were generated fulfilling as closely as possible these restraints. Special restraints were added to keep the Watson–Crick base pairs intact. The five candidate structures with the lowest number of distance restraints violations were averaged. This averaged structure was refined by energy minimization and subsequently used for the next IRMA cycle. Convergence was reached after three cycles. This procedure resulted in 56 IRMA refined proton–proton distance restraints for the methylene-acetal-linked T4^T5 dimer. This set was extended to 63 restraints by adding additional restraints to exclude potential rotamers that were contradictory to the NOE data. The restraints are depicted in Table 2.

Experimental

Configuration

Table 3 gives the configurational settings of the genetic algorithm that is used in this study. The first entries concern the torsion angle ranges. From the NOE data it could be deduced that the γ torsion angles were restricted to the *gauche plus* domain, the χ torsion angles to the *anti* domain and the sugar rings in a *South* conformation. The torsion angles were encoded by Gray coding (Caruana and Schaffer, 1988). The genetic algorithm population consisted of 100 trial structures. A threshold selection criterion was used with an elitist fraction of 5% and a threshold value of 25%. This means that the best 5% of all the structures in the population are always selected in the copy pool. The copy pool is filled further with bit strings from the best 25% of the strings from the population. Uniform crossover with a probability of 90% and a swap rate of 16% was used. A mutation operator also based on this principle was used with a probability of 4%. The reader is referred to Lucasius and Kateman (1994a,b) for a detailed reading on genetic algorithm configuration.

Initially the distance geometry calculations for T4^T5 were performed with a first-generation algorithm. The calculations were repeated with a modern distance geometry algorithm, i.e., a DGII package by Biosym (Biosym Technologies, San Diego, CA, U.S.A., 1993). Like in the initial distance geometry calculations, 75 structures were generated. In the DGII case they were refined by 500 steps of simulated annealing. More than 500 steps of simulated annealing did not lead to, e.g., lower maximum violations and mean violations of the restraints. The configurational settings for the DGII calculations are depicted in Table 4.

Convergence

A genetic algorithm run is usually terminated when no significant improvement of the bit strings is observed. The simplest way is to monitor a ‘bestever’ structure or a ‘best of the current population’ structure. When one of these structures does not show any improvement during a (large) number of generations, the algorithm is said to be converged. In this case each genetic algorithm run delivers a single (best) structure. However, this does not mean that the complete population will resemble one of the best structures. There can still be a large diversity in the population (largely due to the mutation operator). In a distance geometry calculation an ensemble of structures is produced. Usually a number of structures that fulfill the distance restraints the best are superimposed to get an impression of the conformational space spanned by the available restraints. To compare the results of the genetic

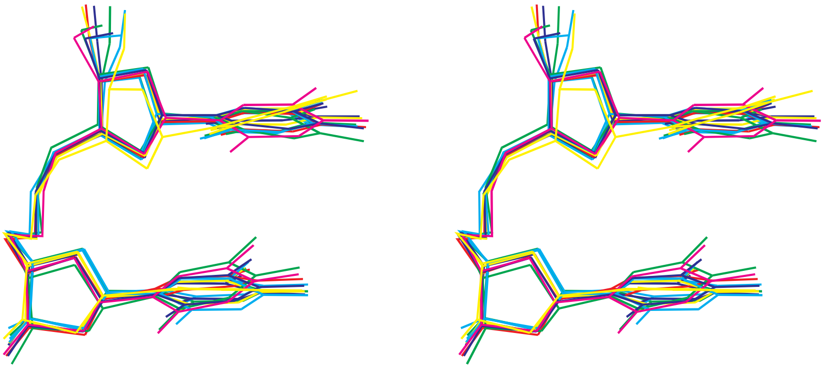

Fig. 3. Stereoscopic view of 10 superimposed structures of the methylene-acetal-linked dimer generated during initial distance geometry calculations showing the least number of violations (hydrogens are not shown).

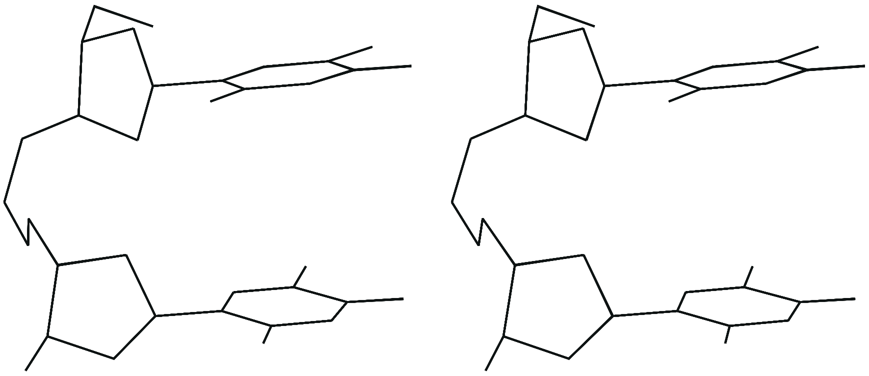

Fig. 4. Stereoscopic view of 10 superimposed ('bestever') structures of the methylene-acetal-linked dimer generated by a genetic algorithm using 10 different random seeds (hydrogens are not shown).

algorithm runs with the distance geometry calculations, we followed two strategies. First, we started several genetic algorithms with identical configurational setting files. However, they were initialized with a different random seed value in the initialization stage. The 'bestever' structures resulting from these runs were superimposed. Here the influence of the starting position in the search space was investigated. Second, we started a single genetic algorithm and let it run until it converged. Then we selected bit strings from the population that had an objective function value below a certain threshold. These structures were also superimposed. Here we investigated the diversity among structures that had a low objective function value. Because T4^T5 is so well defined by the distance restraints, we expected little diversity in bit strings that had low objective function values. In other words, only one solution was expected for different initializations of the genetic algorithm. To circumvent this we defined two additional experiments. In the first additional experiment we relaxed the available restraints by 10%. Then 10 structures (with low objective function values) of a population that had converged were superimposed. In the second additional experiment we randomly removed restraints from the data set step-by-step. Here 10 structures that had converged to such a reduced restraints set were superimposed.

Hardware

Genetic algorithm versions for conformational analysis are available for SUN Sparc^{TM}, Silicon Graphics and PC platforms. The experiments in this study were performed on a SUN Sparc Ultra workstation. Convergence was typically reached within 1000 generations, which took approximately 55 CPU seconds. The DGII calculations were performed on a Silicon Graphics Indigo R4600 with the INSIGHT II package from Biosym/Molecular Simulations.

Results and Discussion

Figure 3 shows a stereo plot of the 10 best structures from the initial distance geometry calculations. The average backbone torsion angles (plus a standard deviation) for these structures are shown in column 2 of Table 5. Ten 'bestever' structures that resulted from 10 genetic algorithm runs, started with different random seeds, are depicted in the stereo plot of Fig. 4.

As expected, these structures are virtually the same. Column 4 of Table 5 shows the average backbone torsion angles of these genetic algorithm structures. Columns 5 and 6 of Table 5 give the average backbone torsion angles of the regular B(I)-DNA rotamers and the less common B(II)-DNA rotamers. They are calculated from the crystal

TABLE 5
AVERAGE BACKBONE TORSION ANGLES (°) OF THE METHYLENE-ACETAL-LINKED THYMINE DIMER (DETERMINED IN VARIOUS WAYS) AND OF TWO DIFFERENT B-DNA ROTAMER FAMILIES

Torsion angle	DG	DGII	GA[a]	B(I)-DNA	B(II)-DNA
ε	205 (5)	209 (1)	199	180 (13)	246 (17)
ζ	248 (4)	247 (2)	265	267 (12)	175 (14)
α	302 (3)	300 (1)	323	301 (12)	298 (18)
β	148 (7)	156 (1)	177	179 (10)	144 (10)
γ	65 (5)	56 (2)	33[b]	49 (9)	45 (11)
$\varepsilon-\zeta$	–43 (8)	–38 (2)	–66	–87 (17)	71 (22)

GA: genetic algorithm; DG and DGII: distance geometry and second-generation distance geometry.

[a] Because of the very small standard deviations on the torsion angles of the genetic algorithm structures, they are not depicted in the table.

[b] The average γ torsion angle value of T5 is given here; the average γ torsion angle value of T4 was 60.0.

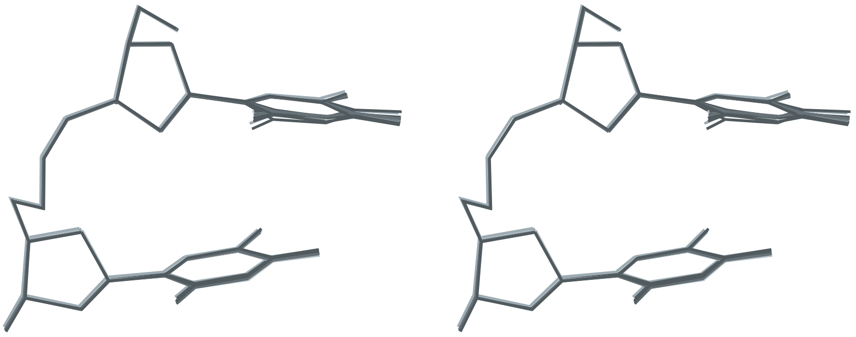

Fig. 5. Stereoscopic view of 10 superimposed structures of the methylene-acetal-linked dimer generated during DGII distance geometry calculations showing the least number of violations (hydrogens are not shown).

structures of 20 selected B-DNA decamers and dodecamers that are present in the Nucleic Acid Database Project (Berman et al., 1992). Besides the individual backbone torsion angles, the ε−ζ difference is also shown as an indicator of the B-DNA rotamer family.

The torsion angles that resulted from the distance geometry calculations suggest that T4^T5 belongs to the regular B(I)-DNA family of rotamers. However, the ε, ζ and β torsion angles are somewhat biased towards a less common B(II)-DNA conformation. The results of the genetic algorithm optimization show a more pronounced regular B(I)-DNA conformation.The average rmsd of the 10 genetic algorithm structures was 0.0026. The close resemblance of these structures (standard deviation 1.0e–05, see Fig. 4) is an indication that the genetic algorithm is not dependent on the starting point in the search space spanned by the torsion angle ranges and available distance restraints. However, the 10 best initial distance geometry structures all had lower rmsd values than the genetic algorithm structures and hence fulfilled the distance restraints better. We were surprised that the genetic algorithm did not find one of these structures, although the torsion angle ranges used in the genetic algorithm configurational settings included the torsion angles that were found by the initial distance geometry calculations.

Therefore, an experiment was set up in which the torsion angle ranges for the genetic algorithm were constrained to allow only structures that fell in the initial distance geometry structures category. Hence, the torsion angle ranges for the genetic algorithm trial structures were defined by the average torsion angles of the 10 best distance geometry structures and their respective standard deviations (see Table 5). Under these circumstances, the ‘bestever’ genetic algorithm structure had an rmsd of 0.0037. This suggests that there is a difference in geometry between distance geometry structures and genetic algorithm structures other than torsion angles. To verify this, bond lengths and bond angles of the best initial distance geometry structures were measured. The measured bond angles are shown in column 3 of Table 1. These values clearly differ from those used in the genetic algorithm optimization. It seems that during distance geometry calculations, the bond angle geometry is somewhat distorted to fulfill the distance restraints. To verify this, the best distance geometry structures were subjected to energy minimization to see whether the distorted geometry would hold. It could be seen that in the first steps of the minimization, the bond angles were relaxed back to the original values. Simultaneously, however, the rmsd increased during the minimization. The minimized distance geometry structures showed a higher degree of violations than the genetic algorithm structures. It dem-

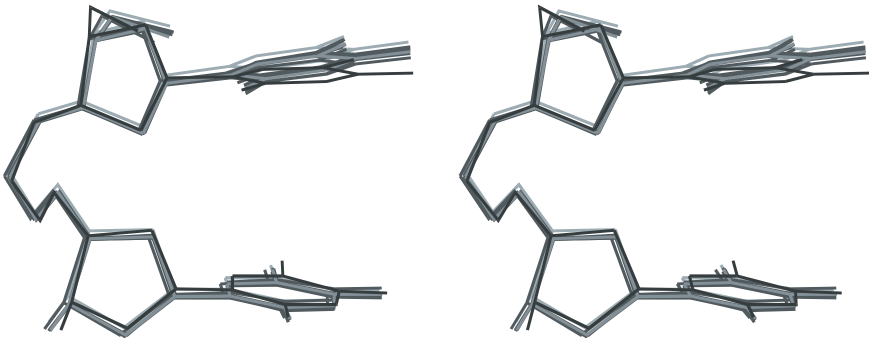

Fig. 6. Stereoscopic view showing 10 superimposed (‘bestever’) structures of the methylene-acetal-linked dimer generated by a genetic algorithm using 10 different random seeds with the original restraints (black line, compare with Fig. 4) and 10 superimposed structures generated by a genetic algorithm with low objective function values when the restraints were relaxed by 10% (other lines).

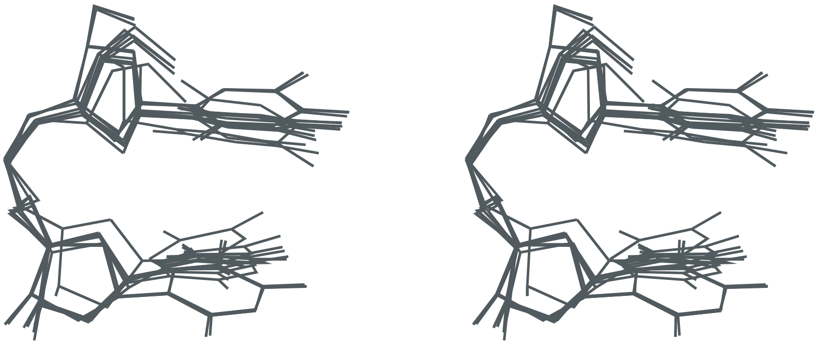

Fig. 7. Stereoscopic view showing 10 superimposed ('bestever') structures of the methylene-acetal-linked dimer generated by a genetic algorithm when reduced restraints sets were used. Up to 40 restraints were removed step-by-step in a random fashion. For restraints sets where more than 35 restraints were removed, sometimes ill-defined structures resulted. Therefore, optimized structures for restraints sets where 1, 2, 4, 8, 12, 16, 20, 24, 28 and 32 restraints, respectively, were removed are depicted.

onstrates that, although distance geometry calculations of the modified dimer can produce structures with small distance restraints violations, the bond angles are biased towards the restraints, which is not desirable in most cases.

Because first-generation distance geometry packages are known to produce structures of poor quality with respect to bond lengths and bond angles, the calculations were repeated with the DGII package. Ten structures that fulfill the restraints well are superimposed in Fig. 5. They resemble the initially found distance geometry structures (see column 3 of Table 5). Also the rmsd values are comparable with the initially found structures. However, as can be seen in Table 1 even after 500 steps of simulated annealing some extreme values for bond angles around the methylene-acetal linkage are found. The genetic algorithm does not have this drawback. However, the problem with the genetic algorithm is the extreme convergence to a specific solution.

Hence, an additional point that is addressed is the sampling behavior of the genetic algorithm optimization. Within a population the diversity among structures can be quite high. We drew up an inventory of the individual bit strings in the population and selected bit strings that had an rmsd < 0.0030. In this way we could see whether, among the structures with a low rmsd, different conformations were present. The structures with rmsd < 0.0030 all resembled to a high degree the structures depicted in Fig. 4. Obviously, the available distance restraints force the genetic algorithm to converge to a family of structures that are virtually the same. When the restraints are relaxed by 10%, some variability is introduced in genetic algorithm structures as can be seen in Fig. 6. However, the variability among structures with a comparable low objective function value is not large. This may be attributed to the rather tight original restraints, which on relaxation by only 10% remain rather tight. However, it is clear that there is variability between the group of structures that were optimized with the original restraints and the group of structures that were optimized with the restraints that were relaxed by 10% (the pairwise rmsd for atom positions was greater than 0.35 in all cases). Obviously, DGII calculations with the restraints that were relaxed resulted in a larger variability in structures, but it also resulted in a larger variability in bond angles. As can be seen in column 5 of Table 1, the mean bond angles are acceptable but the standard deviation is large. This means that bond angles in some of the structures had rather extreme values. Finally, in Fig. 7 a superposition of 10 'bestever' genetic algorithm structures optimized for reduced sets of restraints is depicted. Here it can be seen that when, by chance, a 'tight restraint' is removed from the restraints set, a change in 'bestever' structure occurs.

Conclusions and Outlook

We compared distance geometry calculations and genetic algorithm optimization in the structure determination of a methylene-acetal-linked thymine dimer with NMR-derived distance restraints. The geometry around the bond angles plays an important role. Distance geometry calculations produce structures that fulfill the restraints to a reasonable degree, but unreliable bond angles are found. Especially the bond angles around the central carbon in the methylene-acetal linkage differ from the expected tetrahedral geometry. In this study, it appears that the genetic algorithm optimization of torsion angles with fixed bond angles taken from an energy-minimized duplex yields more reliable results. It has to be stressed that in comparing structures produced by both genetic algorithm optimization and distance geometry calculations, only structures that fulfill the restraints well are superimposed. Obviously, the genetic algorithm structures showed little variability. Selecting 10 structures that fulfill restraints the best from an ensemble of 75 distance geometry structures also leads to a set with little variability. However, it is a proper way to study the bond angle geometry of the optimized structures, especially the geometry around the methylene-acetal linkage.

The genetic algorithm optimization in torsion angle space suggests a three-dimensional spatial structure of the methylene-acetal-linked thymine dimer that is in the regular B(I)-DNA rotamer domain. These structures show slightly larger violations of the distance restraints than the distance geometry structures. However, the user now has influence on the choice of the bond angle geometry of, e.g., the methylene-acetal linkage. A comparison of the violations by the distance geometry structure and the violations by the genetic algorithm structure might lead to the detection of possible inconsistencies in the distance restraints file. The fact that all genetic algorithm structures converge to a similar family of conformations of the regular B(I)-DNA rotamer might also give a clue on how to relax the IRMA-derived restraints more. For example, in the genetic algorithm structures α always converges to a rather high value, while γ of T5 always converges to a rather low value. Optionally, it seems interesting to try and optimize the bond angles that define the geometry around the central carbon atom in the methylene-acetal linkage by means of the genetic algorithm. For this purpose the bond angles can be taken as additional parameters on the bit strings. Under the assumption of a tetrahedral geometry, the sum of the six bond angles around the central carbon of the methylene-acetal linkage must add up to 6·arccos(−1/3). Preliminary results using this approach are promising.

Acknowledgements

The authors thank Kees Huige for providing the force field parameters of the methylene-acetal group and Antoine van Kampen for his aid in the DGII calculations. Bart Rutjes is acknowledged for his help in implementing the genetic algorithm used in this study.

References

Altona, C. and Sundaralingam, M.J. (1972) *J. Am. Chem. Soc.*, **94**, 8205–8212.

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

Blommers, M.J.J., Lucasius, C.B., Kateman, G. and Kaptein, R. (1992) *Biopolymers*, **32**, 45–52.

Boelens, R., Koning, T.M.G. and Kaptein, R. (1988) *J. Mol. Struct.*, **173**, 299–311.

Boelens, R., Koning, T.M.G., Van der Marel, G.A., Van Boom, J.H. and Kaptein, R. (1989) *J. Magn. Reson.*, **82**, 290–308.

Brodmeier, T. and Pretsch, E. (1994) *J. Comput. Chem.*, **15**, 588–595.

Caruana, R.A. and Schaffer, J.D. (1988) In *Proceedings of the Fifth International Conference on Machine Learning*, San Mateo, CA, U.S.A. (Ed., Laird, J.), Morgan Kaufmann, Los Altos, CA, U.S.A., p. 153.

De Leeuw, H.P.M., Haasnoot, C.A.G. and Altona, C. (1980) *Isr. J. Chem.*, **20**, 108–126.

Gao, X., Brown, F.K., Jeffs, P., Bischofberger, N., Lin, K.-Y., Pipe, A.J. and Noble, S.A. (1992) *Biochemistry*, **31**, 6228–6236.

Goldberg, D.E. (1989) *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley, Reading, MA, U.S.A.

Havel, T. and Wüthrich, K. (1984) *Bull. Math. Biol.*, **46**, 673–698.

Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI, U.S.A; revised: (1992) MIT Press, Cambridge, MA, U.S.A.

Jeener, J., Meier, B.H., Bachmann, P. and Ernst, R.R. (1979) *J. Chem. Phys.*, **71**, 4546–4553.

Lucasius, C.B., Werten, S., Van Aert, A.H.J.M., Blommers, M.J.J. and Kateman, G. (1991) In *Parallel Problem Solving from Nature 1* (Eds., Schwefel, H.P. and Männer, R.), Springer, Berlin, Germany, pp. 90–97.

Lucasius, C.B. and Kateman, G. (1993) *Chemometr. Intell. Lab. Syst.*, **19**, 1–33.

Lucasius, C.B. and Kateman, G. (1994a) *Comput. Chem.*, **18**, 127–136.

Lucasius, C.B. and Kateman, G. (1994b) *Comput. Chem.*, **18**, 137–156.

Macura, S. and Ernst, R.R. (1980) *J. Mol. Phys.*, **41**, 95–117.

Macura, S., Huang, Y., Suter, D. and Ernst, R.R. (1981) *J. Magn. Reson.*, **43**, 259–281.

McGarrah, D.B. and Judson, R.S. (1993) *J. Comput. Chem.*, **14**, 1385–1395.

Privé, G.G., Heinemann, U., Chandrasegaran, S., Kan, L.S., Kopka, M.L. and Dickerson, R.E. (1987) *Science*, **238**, 498–504.

Quaedflieg, P.J.L.M., Pikkemaat, J.A., Van der Marel, G.A., Kuyl-Yeheskiely, E., Altona, C. and Van Boom, J.H. (1993) *Recl. Trav. Chim. Pays-Bas*, **112**, 15–21.

Sun, S. (1995) *Biophys. J.*, **69**, 340–355.

Van Kampen, A.H.C., Buydens, L.M.C., Lucasius, C.B. and Blommers, M.J.J. (1996) *J. Biomol. NMR*, **7**, 214–224.

Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A. (1986) *J. Comput. Chem.*, **7**, 230–252.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY, U.S.A.