Journal of Biomolecular NMR, **16:** 147–164, 2000.
KLUWER/ESCOM

Determination of the populations and structures of multiple conformers in an ensemble from NMR data: Multiple-copy refinement of nucleic acid structures using floating weights

Adrian Görler, Nikolai B. Ulyanov & Thomas L. James*
Department of Pharmaceutical Chemistry, University of California, 513 Parnassus Ave., 926 Medical Science, San Francisco, CA 94143-0466, U.S.A.

Received 19 August 1999; Accepted 6 December 1999

Key words: A-DNA, B-DNA, conformational variability, ensemble averaging, Monte Carlo simulations, nucleic acids, relaxation rates, structure refinement

Abstract

A new algorithm is presented for determination of structural conformers and their populations based on NMR data. Restrained Metropolis Monte Carlo simulations or restrained energy minimizations are performed for several copies of a molecule simultaneously. The calculations are restrained with dipolar relaxation rates derived from measured NOE intensities via complete relaxation matrix analysis. The novel feature of the algorithm is that the weights of individual conformers are determined in every refinement step, by the quadratic programming algorithm, in such a way that the restraint energy is minimized. Its design ensures that the calculated populations of the individual conformers are based only on experimental restraints. Presence of internally inconsistent restraints is the driving force for determination of distinct multiple conformers. The method is applied to various simulated test systems. Conformational calculations on nucleic acids are carried out using generalized helical parameters with the program DNAminiCarlo. From different mixtures of A- and B-DNA, minor fractions as low as 10% could be determined with restrained energy minimization. For B-DNA with three local conformers (C2′-*endo*, O4′-*exo*, C3′-*endo*), the minor O4′-*exo* conformer could not be reliably determined using NOE data typically measured for DNA. The other two conformers, C2′-*endo* and C3′-*endo*, could be reproduced by Metropolis Monte Carlo simulated annealing. The behavior of the algorithm in various situations is analyzed, and a number of refinement protocols are discussed. Prior to application of this algorithm to each experimental system, it is suggested that the presence of internal inconsistencies in experimental data be ascertained. In addition, because the performance of the algorithm depends on the type of conformers involved and experimental data available, it is advisable to carry out test calculations with simulated data modeling each experimental system studied.

Abbreviations: DNA, deoxyribonucleic acid; MD, molecular dynamics; MDtar, time-average molecular dynamics; NOE, nuclear Overhauser effect.

Introduction

It is a well acknowledged fact that biomolecules such as proteins, nucleic acid fragments or polysaccharides are flexible in solution and can adopt multiple conformations. Therefore, data obtained by NMR measurements must be regarded as averages over the measurement time and the molecules in the sample. However, in conventional methods of structure determination by NMR, a structure is accepted only if it satisfies all or most of the experimental restraints. Such approaches often result in an ensemble of structures tightly distributed around their mean and showing little evidence of conformational variability. Furthermore, the experimental restraints might contain

*To whom correspondence should be addressed. E-mail: james@picasso.nmr.ucsf.edu inherent contradictions, which cannot be resolved by a single structure alone. In the past few years, different approaches have been proposed to tackle this problem and take into account the time- and ensemble-average nature of NMR restraints.

Torda et al. (1990) proposed a method, commonly referred to as time-average molecular dynamics (MDtar). Rather than enforcing distance restraints at each time step, MDtar only requires that they are satisfied for appropriately averaged distances over the course of a molecular dynamics (MD) trajectory. For systems with multiple conformers, MDtar is efficient in exploring conformational space and able to generate ensembles of structures, which satisfy experimental NMR data better than any single structure (González et al., 1995; Yao et al., 1997). In general, a successful determination of structural ensembles with distinct conformers depends on having a sufficient number of mutually inconsistent experimental restraints, i.e., those which cannot be simultaneously satisfied by a single structure. MDtar resolves such inconsistencies by enabling the molecule to switch back and forth between various conformational minima, satisfying experimental restraints on a time-average basis (Schmitz et al., 1996). The rates of these conformational transitions are not realistic, of course; they are accelerated. This very property of MDtar is the source of its fundamental limitations. Indeed, if the molecule jumps quickly between different minima during the simulation, it must also spend a significant amount of time on the top of energy barriers. Consequently, many snapshots of MDtar trajectories have compromised conformational energy. Another methodological limitation of the MDtar approach lies in the fact that it is only practical to use a relatively small time interval for the averaging, significantly limiting the range of dynamic excursions.

In a different approach, three algorithms for the ensemble-average MD (multiple-copy MD) were presented roughly simultaneously (Bonvin and Brünger, 1995; Fennen et al., 1995; Kemmink and Scheek, 1995). Differing in details, they all refine an ensemble of structures simultaneously. At each time step, appropriately averaged distances are calculated using all structures in the ensemble. The averages are enforced to be in agreement with NOE-derived distance restraints. Because multiple-copy MD employs ensemble averaging rather than time averaging, it does not require frequent transitions past energy barriers, so it is devoid of one major limitation of MDtar. However, this approach must deal with the question of relative weights of individual members of the ensemble. In the algorithms of Bonvin and Brünger and of Kemmink and Scheek, it is assumed that all structures in the ensemble have the same weight. Clearly, such algorithms are ill-suited for the determination of relative populations of solution conformers (Bonvin and Brünger, 1996). In the method of Fennen et al., individual structures are weighted with the Boltzmann factors calculated from their conformational energies. In this method, the resulting populations would, therefore, depend more on the force field used than on experimental data. Furthermore, in such an approach, only the depth of conformational minima, but not their width, is taken into account.

Here, we present a new method to perform a restrained multiple-copy refinement of nucleic acid structures. The basis of our method is similar to ensemble-average MD, but we approach differently the question of relative weights of individual structures. Relative populations of conformers are determined at each step of the simulation based on experimental data, using the PDQPRO algorithm (Ulyanov et al., 1995). The implementation of this method, multiple-copy refinement with floating weights, is based on extension and combination of three existing programs: (1) The DNAminiCarlo program (Ulyanov et al., 1989), which performs conformational calculations in the internal coordinate space, is used as the refinement engine. (2) PDQPRO (Ulyanov et al., 1995) calculates optimal populations of conformers at each refinement step by a quadratic programming algorithm. (3) RELAX (Görler and Kalbitzer, 1997) computes proton-proton dipolar cross-relaxation rates for the ensemble; the rates are used to calculate the penalty function, which is optimized during the refinement. The new method is applied to a number of simulated systems. The possibilities and limitations of the algorithm are assessed.

Theory

Relaxation rates and fast exchange

All methods to determine structures from NMR data have a common feature: they minimize the total energy E_{total} of the system, which is the sum of the empirical force field-derived conformational energy and an artificial energy term E_{NMR}. E_{NMR} is designed to rise as violations of experimental data increase. Commonly, distance restraints are used to incorporate the information derived from NOE data into this penalty function while torsion angle restraints (or sometimes, scalar coupling constant restraints directly) take care of the information derived from J-coupling data.

The matrix of NOE intensities $\mathbf{A}(\tau_m)$ is related to the matrix of dipolar relaxation $\mathbf{R}$ by the generalized Solomon equation:

$$\mathbf{A}(\tau_m) = \exp(-\mathbf{R}\tau_m) \tag{1}$$

with the mixing time τ_m (Keepers and James, 1984). For each structure α, the cross-relaxation rates R_{ij}^{α} (the off-diagonal elements of the matrix $\mathbf{R}$) are inversely proportional to the sixth power of the corresponding interproton distances r_{ij}^{α}:

$$R_{ij}^{\alpha} \propto (r_{ij}^{\alpha})^{-6} \tag{2}$$

If m different conformers $\alpha = 1...m$ are in fast (on the NMR time scale) exchange with each other, effective relaxation rates R_{ij} can be calculated as linear averages of the relaxation rates for individual conformers (Landy and Rao, 1989):

$$R_{ij} = \sum_{\alpha=1}^{m} p^{\alpha} R_{ij}^{\alpha}, \tag{3}$$

where p^{α} is the normalized population of conformer α. Analogously, one can calculate the effective relaxation rates from r^{-6}-averaged distances:

$$R_{ij} \propto \left(\sum_{\alpha=1}^{m} p^{\alpha} (r_{ij}^{\alpha})^{-6}\right) \tag{4}$$

Such r^{-6}-averaged distances are used in conventional multiple-copy refinement or in MDtar to calculate the penalty function.

In the approach presented here, instead of using distance restraints to incorporate NOE-derived information, we use experimentally determined dipolar relaxation rates as restraints directly to calculate a penalty function Q^r:

$$Q^r = \sum_{observed(ij)} \left(\sum_{\alpha=1}^{m} p^{\alpha} R_{ij}^{\alpha} - X_{ij}\right)^2, \tag{5}$$

where X_{ij} are experimentally determined relaxation rates corresponding to the model-derived relaxation rates R_{ij}^{α} and p^{α} is the population of the copy α.

Dipolar relaxation rates can be determined only indirectly from experimental NOE data. A feasible approach to obtain dipolar relaxation rates is to employ the MARDIGRAS procedure (Borgias et al., 1990).

PDQPRO algorithm

Strictly speaking, Equation 5 can only be evaluated if the populations p^{α} of the individual conformers α are known, which they are generally not. However, Q^r takes a minimal value if the populations p^{α} are chosen correctly. In this approach, we take advantage of this fact and treat the populations p^{α} as floating, forcing them to adopt the values for which Q^r is minimal. We do so by calculating Q^r using the algorithm PDQPRO (Probability Distribution by Quadratic PROgramming, Ulyanov et al., 1995).

Generally, PDQPRO calculates the set of probabilities $\{p^1 ... p^m\}$ that minimizes the quadratic function

$$Q^r\left(\left\{p^1 \ldots p^m\right\}\right) = \sum_{k=1}^{n} w_k (T_k - E_k)^2 \tag{6}$$

Herein E_k stands for the experimental value of an observable parameter k, and T_k and w_k are a theoretical value and a weight of the same parameter. T_k is calculated as population-weighted linear average

$$T_k = \sum_{\alpha=1}^{m} p^{\alpha} t_k^{\alpha} \tag{7}$$

of the values t_k^{α}, which are calculated for each theoretical conformer α. The populations p^{α} are normalized:

$$\sum_{\alpha=1}^{m} p^{\alpha} = 1 \tag{8}$$

and non-negative:

$$p^{\alpha} \geq 0, \alpha = 1 \ldots m \tag{9}$$

Weights w_k can be used to reflect differential experimental uncertainty of individual observed parameters or to equalize the contribution of parameters of different type (such as dipolar relaxation rates and scalar coupling constants). However, in all simulations presented here, we made use of only dipolar relaxation rates ($E_k = X_{ij}$ and $t_k^{\alpha} = R_{ij}^{\alpha}$), and we assumed all weights w_k equal.

In the n-dimensional vector space spanned by n observable parameters, the quadratic function Q^r (Equation 6) has a geometric interpretation, which is easy to visualize (Figure 1). The value of Q^r is a squared distance between the shaded polyhedron and the open circle. The circle represents a point in the n-dimensional space corresponding to the experimental values of observable parameters ($E_1 ... E_n$). The polyhedron is defined by m vertices given by the vectors of theoretical values for the observable parameters,

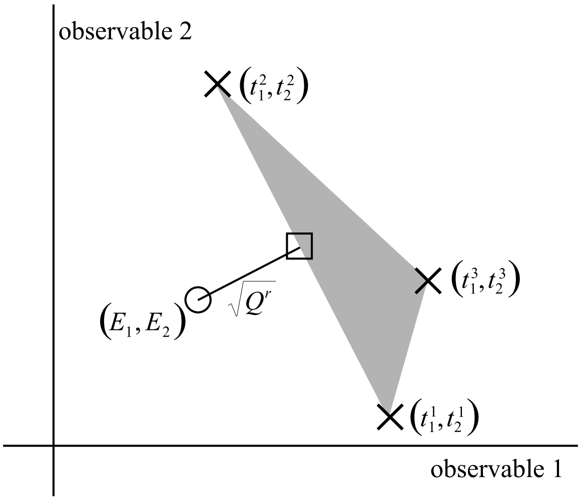

Figure 1. Geometrical interpretation of the objective function Q^r of PDQPRO is illustrated in a two-dimensional space of observable parameters. The open circle (○) represents measured values of observable parameters (E_1, E_2). Black crosses (×) represent the corresponding theoretical values for individual conformers (three, in this example). The shaded polyhedron (a triangle) occupies a region in space with all possible theoretical values of observable parameters (T_1, T_2) for the ensemble (see text). The objective function Q^r is the squared distance between the polyhedron and the open circle. The open square (□) represents a point within the polyhedron which is closest to the experimental point (E_1, E_2); this defines optimal populations of each individual conformer. Practically, this point is found by minimization of Q^r with the quadratic programming algorithm (see text).

$\mathbf{t}_\alpha = (t_1^\alpha \ldots t_n^\alpha)$, $\alpha = 1 \ldots m$; each vertex corresponds to one of the m current conformers. The polyhedron represents all possible values of the theoretical parameters ($T_1 \ldots T_n$) for the ensemble, when m current conformers are fixed and their populations span all possible values (Equation 7). In PDQPRO, this distance is determined by minimization of Q^r with the quadratic programming algorithm (Fletcher, 1981) which ensures that the global minimum is found as solution.

Restraint energy

In single-copy refinement, it is assumed that a single conformation contributes to the NMR signal and that experimental restraints contain no intrinsic contradictions. Ideally, the conformational energy and the NMR-derived penalty function have a common minimum in the conformational space. Then, a sufficiently large number of perfect restraints alone could drive a refinement to this minimum, just as this would be possible with an ideal force field alone.

In multiple-copy refinement, the situation is fundamentally different, even for a perfect force field and exactly measured restraints. Here, a solution ensemble cannot be calculated from experimental restraints alone. Conformational space of ensembles of m structures has m times as many dimensions as the conformational space of a single structure. Consequently, a set of experimental restraints could be explained by a variety of ensembles of structures, but the individual structures of these ensembles would not necessarily have low conformational energy. To successfully find a solution ensemble with multiple-copy refinement, both conformational energy and restraint energy are required. The two energy terms must be well balanced during the refinement. If E_{NMR} dominates, the refinement could result in structures with unreasonably high conformational energy. If E_{conf} dominates, significant conformational minima might be missed.

Any structure calculation, by energy minimization, by molecular dynamics or by Metropolis Monte Carlo simulations, is driven by differences in the total energy or, equivalently, by the gradient of the energy. The penalty function Q^r, as given by Equation 5, is of purely quadratic nature so that its derivative is a linear function of Q^r. To avoid dominance of the restraint energy E_{NMR} when Q^r is large, we do not usually apply Q^r directly as the energy term E_{NMR}, but rather transform it into the flat-well function

$$E_{NMR}(Q^r) = k_{NMR} \begin{cases} 0 & : \sqrt{Q^r} < q_0 \\ (\sqrt{Q^r} - q_0)^2 & : q_0 \leq \sqrt{Q^r} < q_1 \\ s(\sqrt{Q^r} - q_0) + b & : q_1 \leq \sqrt{Q^r} \end{cases} \qquad (10)$$

This form of penalty function is reminiscent of how distance restraints are often used in restrained MD calculations. In Equation 10, E_{NMR} is defined in segments: E_{NMR} is zero for $\sqrt{Q^r}$ smaller than a value q_0, E_{NMR} is proportional to Q^r for values of $\sqrt{Q^r}$ between q_0 and q_1, and it is proportional to $\sqrt{Q^r}$ for values of $\sqrt{Q^r}$ larger than q_1. The error bound q_0 and the slope s are defined by the user, and q_1 and b are adjusted automatically so that $E_{NMR}(Q^r)$ has a continuous derivative. By adjusting s and q_0 for each refinement problem, one can ensure that E_{NMR} and E_{conf} are well balanced during the course of the refinement.

Restraint energy in the space of observable parameters

When one examines the shape of the restraint energy in the space of observable parameters, a fundamental difference between conventional multiple-copy re-

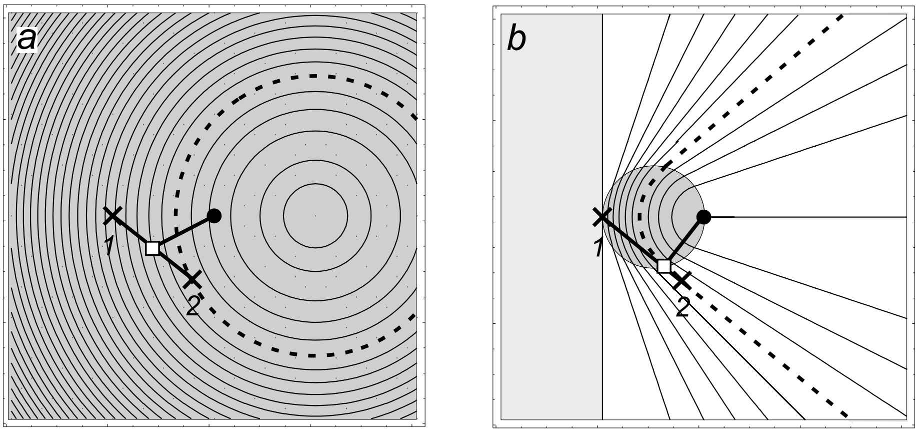

Figure 2. The potential of restraint energy E_{NMR} in space of two observable parameters is shown in the case of multiple-copy refinement with fixed equal populations (a) and in the case of multiple-copy refinement with floating populations (b). Points representing a pair of theoretical values for two different conformers are shown by black crosses, and the point representing experimental values is shown by a filled circle. In this case, the polyhedron of Figure 1 is a segment connecting conformers 1 and 2. The figure shows the restraint energy of the system, when copy 1 is fixed and copy 2 is variable. In multiple-copy refinement with equal weights (a), the potential is purely harmonic. In multiple-copy refinement with floating weights, the potential is divided into three regions. (1) If copy 2 is within the region shaded in light gray, then the restraint energy is constant, because the population of copy 2 is zero. (2) If copy 2 is within the white region, the equipotential lines are straight lines originating from the location of copy 1 (both copies have non-zero populations). (3) If copy 2 is within the circular region, the potential is harmonic and does not depend on the position of copy 1 (because the population of copy 1 is zero). The dashed line limits the area that can be reached by copy 2 without raising the restraint energy of the system. In multiple-copy refinement with equal weights (a), this copy can move only within a closed circular region without raising its restraint energy; in multiple-copy refinement with floating weights (b), this region is cone-shaped, bounded by straight lines and open (unlimited).

finement and multiple-copy refinement with floating weights becomes obvious. Figure 2 illustrates this difference for the case of two observable parameters and two copies, but the extension to more dimensions or more copies is straightforward. In multiple-copy refinement with equal weights, the restraint energy experienced by one copy is purely harmonic, if all other copies remain fixed. In multiple-copy refinement with floating weights, the situation is different, however. The space of observable parameters is separated into three areas (Figure 2b). Only if the moveable copy has a population of 1.0, then the restraint energy of the system is harmonic. If it has zero population, the restraint energy does not depend on this copy at all. If the population is intermediate, then the equipotential lines are linear.

The two potentials have a different topology. In conventional multiple-copy refinement, the equipotential lines are closed circles. Consequently, a copy can sample only a closed area in the conformational space without raising the restraint energy of the system. When one copy is fixed, the refinement forces the other copy towards a single point in the space of observable parameters where E_{NMR} is zero. In multiple-copy refinement with floating weights, the equipotential lines are open, so that a copy can sample an unbounded region of the conformational space without raising E_{NMR}. The refinement drives the moveable copy towards a line in the space of observable parameters along which E_{NMR} is zero. This open topology of the restraint potential in multiple-copy refinement has two positive effects. (1) It allows for better sampling of the conformational space. (2) It enables a structure to follow a valley of low conformational energy, a path that might not be accessible in conventional multiple-copy refinement due to high barriers in the restraint energy.

Figure 2 shows another interesting aspect of multiple-copy refinement with floating weights: in the space of observable parameters, the gradient of the restraint energy is perpendicular to the line connecting both copies, if both copies have nonzero probabilities. This means that the restraint energy does not impose a direct force to change the populations of the copies. Instead, it is the molecular force field of the system which acts on the individual structures, so that conformations and their populations in the ensemble change. The experimental restraints do not contain the populations of the conformers as intrinsic information.

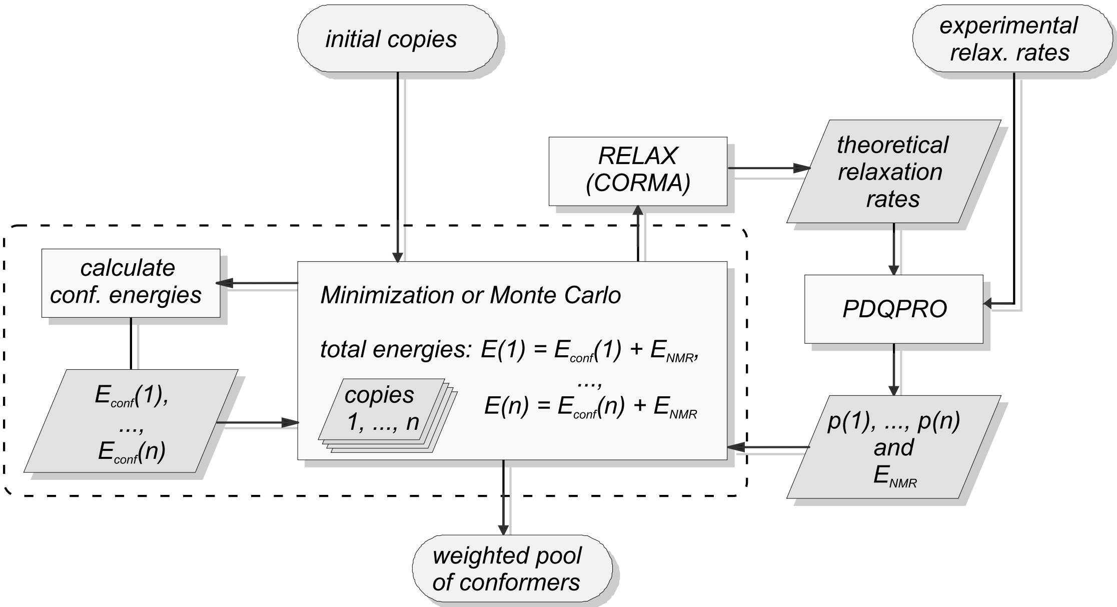
initial copies
experimental relax. rates
RELAX (CORMA)
theoretical relaxation rates
calculate conf. energies
Minimization or Monte Carlo
total energies: E(1) = E_conf(1) + E_NMR,
...,
E(n) = E_conf(n) + E_NMR
copies 1, ..., n
PDQPRO
E_conf(1), ..., E_conf(n)
p(1), ..., p(n) and E_NMR
weighted pool of conformers

Figure 3. Flow chart of multiple-copy refinement with DNAminiCarlo. Starting with an ensemble of initial structures, multiple copies are refined in parallel. The total energy $E_{total}(i)$ of a copy is calculated as the sum of the conformational energy $E_{conf}(i)$ of this copy and restraint energy E_{NMR}. E_{NMR} is evaluated for the ensemble of all copies simultaneously. At each step, PDQPRO determines optimal (floating) populations for the copies in the ensemble, so that the restraint energy is minimal.

Refinement engine

The new algorithm has been implemented by modifying the program DNAminiCarlo (Ulyanov et al., 1989, 1993; Gorin et al., 1990; Zhurkin et al., 1991). The program is capable of performing energy minimization as well as Metropolis Monte Carlo simulations (Metropolis et al., 1953) of nucleic acid structures. The program does not perform structure refinement in Cartesian coordinates but rather uses a set of internal coordinates: generalized helical parameters (Dickerson et al., 1989) for bases and a pseudorotational representation for sugar moieties (Altona and Sundaralingam, 1972). The geometry of the sugar-phosphate backbone is calculated using a special backbone closure algorithm (Zhurkin et al., 1978). This parameterization combines two advantages: (1) It uses idealized geometry for the bases, treating them as rigid bodies, and it fixes bond lengths and most bond angles to their ideal values. (2) It reduces the number of degrees of freedom by about one order of magnitude. As in many MD programs, such as AMBER or Xplor, the structure refinement in DNAminiCarlo is driven by a user-specified refinement protocol. The protocol consists of a sequence of commands, which invoke refinement steps or control the flow of the protocol. The program has recently been extended so that it is applicable not only to DNA but to RNA and hybrid structures as well. Additionally, it underwent numerous changes such as additions to the command language and performance enhancements. A detailed description of the program is in preparation. The program will be available through the authors, as soon as this description is published.

Summary of the algorithm

The overall algorithm is illustrated in Figure 3. An ensemble of starting structures, which can be identical or different, is created by conventional MD, distance geometry, modeling or any other suitable method. With this ensemble, a parallel structure refinement is performed by executing a sequence of refinement steps (energy minimization or Monte Carlo) according to a user-specified refinement protocol. To perform an individual refinement step, the program cycles through all copies in the ensemble, executing the refinement step on all (selected) individual copies sequentially. At each step, the total energy $E_{total}(i)$ of an individual copy i is calculated as the sum of the conformational energy $E_{conf}(i)$ of the copy and the restraint energy E_{NMR}. The conformational energy is calculated for all copies individually so that they do not interact with each other. E_{NMR} is equal for all copies in the ensemble and it is calculated for the whole ensemble simultaneously. For this purpose, all observed dipolar relaxation rates for all individual copies are calculated by invoking routines of the program RELAX (Görler and Kalbitzer, 1997) incorporated in the algorithm. RELAX calculates the relaxation matrix in the same manner as the previously described CORMA (Borgias and James, 1989). The deviation of calculated rates for the ensemble from experimental relaxation rate restraints (which are given in the RELAX input file) is determined with floating populations using PDQPRO and transformed via Equation 10 to E_{NMR}. The process is repeated until all refinement steps specified in the protocol are executed.

Materials and methods

Generation of test structures

To allow an in-depth analysis of the behavior of the new algorithm, all calculations were carried out with simulated data. Two test molecules were used: a DNA dimer with the sequence d(GG):d(CC) and DNA pentamer d(GGGGG):d(CCCCC).

For the dimer sequence, two regular (i.e., with the helical symmetry) target structures were generated with DNAminiCarlo, one in the A-family of forms and another in the B-family. Target A structure had helical parameters similar to the fiber X-ray structure of A-DNA (Chandrasekaran et al., 1989). The target B conformation is close to the ‘average-sequence’ B-DNA in solution as determined by high-resolution NMR (Ulyanov and James, 1995). Both structures are not optimal energetically (Table 1) relative to the force field used (Zhurkin et al., 1981). The reason why we used sub-optimal conformers as target structures was to take into account the fact that the theoretical force fields are never perfect. A number of target ensembles involving these two discrete conformations, A and B, were prepared as described below. All refinements for this test molecule started with the ‘starting A structure’, which was significantly different from both target A and energetically optimal A conformations (Table 1).

For the pentamer sequence, a continuous set of reference structures was determined in the following way. Starting from a regular B-form structure, an initial structure was generated by energy minimization. Then, the sugar pseudorotation angle P_{G3} of G3 was incrementally changed in steps of 0.1°, in the range between 0° and 200°. At each step, the modified structures were energy-minimized keeping P_{G3} fixed. In addition, the geometry of terminal base pairs was also fixed during the minimization. The resulting energy profile has three minima, at $P_{G3} = 28.0°$, 83.5° and 150.0°; we will refer to the corresponding structures as S_{28}, S_{83} and S_{150}, respectively. Although it is beyond the scope of this work to discuss the nature of these three minima, it is consistent with our previous conformational calculations and MDtar simulations (Gorin et al., 1989; Schmitz et al., 1995). The structures S_{28}, S_{83} and S_{150} will be used to construct various three-member ensembles with artificially chosen populations (see below). In addition, a continuous ensemble of 40 structures, with P_{G3} changing from 0° to 200° in steps of 5°, was used in some calculations. The energy profile of these 40 structures is shown in Figure 4.

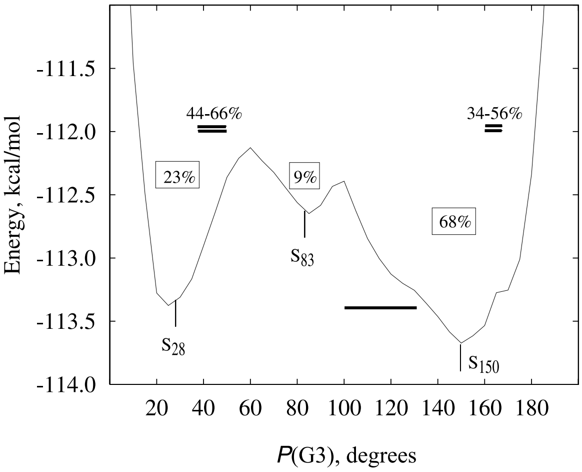

Figure 4. The conformational energy $E_{conf}(P_{G3})$ of the d(GGGGG):d(CCCCC) pentamer is shown as a function of P_{G3}. The energy profile has three minima represented by three discrete structures S_{28} (C3′-*endo*), S_{83} (O4′-*exo*) and S_{150} (C2′-*endo*). The three continuous minima have total populations of 23, 9 and 68%, respectively (the values were obtained by integration between the maxima of energy). The single horizontal line shows a range of structures refined under the assumption of a single copy. Double horizontal lines show a range of refined ensembles with two copies; the range of populations is indicated above.

Simulation of relaxation rates

To simulate experimental NMR data, dipolar cross-relaxation rates were calculated from the target structures described above. Using the program RELAX (Görler, 1997), all ‘essential’ relaxation rates were calculated (Table 2) assuming isotropic tumbling with an overall correlation time of 2.0 ns. These are the rates that correspond to NOE cross peaks typically observed by high-resolution NMR. Contacts that do not vary significantly upon conformational change

Table 1. Helical parameters of the d(GG):d(CC) dimer

	Ω	τ	ρ	Dx	Dy	Dz	ω	κ	σ	Sx	Sy	Sz	P_G	χ_G	P_C	χ_C	E_{conf}
Target structure																	
A	30.1	0.5	12.6	−0.02	−1.43	3.32	−3.0	−0.8	−0.3	0.14	0.03	0.00	28.0	204.0	30.0	204.9	−26.4
B	35.1	0.9	4.5	0.08	−0.35	3.21	−5.9	1.2	−0.9	0.15	0.02	−0.03	145.0	249.4	135.0	238.8	−21.3
Minimal conformational energy																	
A	31.4	−0.6	3.6	−0.13	−1.05	3.54	−8.8	−3.6	−0.3	0.12	0.01	−0.02	21.2	202.6	29.8	209.2	−29.4
B	41.1	0.4	−3.3	−0.02	0.28	3.33	−5.6	3.8	0.0	0.15	0.01	−0.17	166.6	255.1	157.2	245.7	−26.9
Starting structure																	
A	33.1	1.6	−13.5	0.39	−0.51	3.66	−23.9	8.1	−2.8	0.30	0.07	−0.98	18.2	220.6	30.0	201.6	−25.0

Helical parameters (twist Ω, tilt τ, roll ρ, shift Dx, slide Dy, rise Dz, propeller twist ω, buckle κ, opening σ, shear Sx, stretch Sy, stagger Sz, sugar pseudorotation angle P and glycosydic angle χ) are listed for the initial structure and for the target and minimum-energy A- and B-conformations. Definitions of the helical parameters conform to the guidelines of the ‘Cambridge Convention’ (Dickerson et al., 1989). The exact definitions of this particular set of parameters are given by, e.g., Ulyanov and James (1995). Linear parameters are given in angstroms, angular parameters in degrees, and energy in kcal/mol.

Table 2. Observable parameters

G_i - G_i	C_j - C_j	G_i - G_{i+1}	C_j - C_{j+1}
H2′ - H8	H2′ - H6	H2″ - H8	H2′ - H6
H3′ - H8	H2″ - H4′	H2′ - H8	H2″ - H6
H2″ - H4′	H1′ - H4′	H1′ - H8	H2′ - H5
H1′ - H4′	H3′ - H6	H3′ - H8	H2″ - H5
H2″ - H8	H1′ - H6	H8 - H8	H3′ - H6
H1′ - H8	H2″ - H6		H1′ - H6
H4′ - H8	H4′ - H6		H6 - H5
	H2′ - H5		H5 - H5
			H3′ - H5
			H6 - H6
			H1′ - H5

Intra-residue and sequential proton pairs used in the calculations of relaxation rates for both test molecules are listed. No inter-strand restraints were used.

of DNA, such as H1′-H2′ or H1′-H2″, were not included. Overall, 46 relaxation rates were calculated for the d(GG):d(CC) dimer and 143 for the d(GGGGG):d(CCCCC) pentamer. We will use capital Greek letters Δ and Π to denote sets of relaxation rates for the dimer and pentamer, respectively.

For the dimer, 11 sets of relaxation rate restraints Δ_0, $\Delta_1 \ldots \Delta_{10}$ were generated by averaging the essential rates calculated for the A-form and the B-form dimer with the ratios p(A):p(B) = {0.0:1.0, 0.1:0.9,. . ., 0.9:0.1, 1.0:0.0}, respectively. For the pentamer, three sets of relaxation rates (Π_{333}, Π_{226} and Π_{316}) were calculated in the following way. The relaxation rates from the structures S_{28}, S_{83} and S_{150} were determined and averaged, weighted with three discrete sets of populations $p(S_{28})$:$p(S_{83})$:$p(S_{150})$ = {1:1:1; 0.2:0.2:0.6; 0.3:0.1:0.6}, respectively. In addition, two nondiscrete ensembles were considered. For the first one, an artificial potential of the conformational energy was created to simulate a single continuous broad minimum. $E_{conf}(P_{G3})$ was approximated in the range from 140° to 160° by a quadratic parabola $E'_{conf}(P_{G3})$. This ensemble included 20 actual structures calculated as described above with P_{G3} evenly spaced between 120° and 180°. However, its relaxation rates (set Π_{broad}) were averaged with Boltzmann factors based on the artificial quadratic energy $E'_{conf}(P_{G3})$ at temperature $T = 300$ K:

$$p(P_\alpha) = \frac{\exp(-E'_{conf}(P_\alpha)/k_B T)}{\sum_{\alpha'} \exp(-E'_{conf}(P_{\alpha'})/k_B T} \qquad (11)$$

$E'_{conf}(P_{G3})$ has a minimum at $P_{G3} = 150.41°$. The second non-discrete ensemble included 40 structures with P_{G3} between 0° and 200°; its relaxation rates (set Π_{full}) were calculated using Boltzmann factors based on the actual energy $E_{conf}(P_{G3})$ shown in Figure 4. This energy profile can be subdivided in three wide minima: $P_{G3} < 60°$ with total population of 23%, $60° < P_{G3} < 100°$ (population of 9%), and $100° < P_{G3}$ (68%).

Number of observable parameters and relaxation rate errors

The influence of the number of observed parameters was investigated using the Δ_5 target ensemble, which included target A and B conformations of the d(GG):d(CC) dimer with populations of 50% each. The original set of observed parameters included 46 relaxation rates (Table 2), or 11.5 per residue. Four additional series of parameter sets were constructed by randomly removing 10, 20, 30, and 40% of the parameters. Each of these four series consisted of 20 random sets.

Analogously, additional sets of parameters were constructed by adding random errors to the relaxation rates of the original Δ_5 set. Five series of sets had maximum relative errors of 5, 10, 20, 40, and 60%, respectively; each series consisted of 20 random sets of relaxation rates.

Results and discussion

Mixtures of A-form and B-form dimers

The algorithm presented here has been designed to simultaneously determine the location of multiple minima in the conformational space as well as their individual populations. However, its convergence properties and its discriminating power must be investigated, and its limitations assessed. To this purpose, a series of restrained energy minimizations has been carried out with the d(GG):d(CC) dimer. The target ensembles were defined by the sets of relaxation rate restraints $D_0 \ldots D_{10}$, described above. All calculations started with an ensemble of two identical, not energy-minimized, structures in A-form (Table 1). The refinement protocol consisted of two phases of multiple-copy energy minimization. Phase I comprised 20 cycles of energy minimization with the purpose of enabling the ensemble to overcome barriers in the conformational energy. During this phase, the user-defined parameters of the restraint energy term in Equation 10 were kept constant with the values $k_{NMR} = 40$, $s = 1.0$ and $q_0 = 0.0$, if not otherwise stated. In phase II, the five cycles of energy minimization were meant to force the ensemble closer to the target. This was achieved by raising the weight of the restraint energy term, k_{NMR}, to a value of 200. Table 3 summarizes the results of these calculations.

Single-structure ensembles

For the target ensembles Δ_0 and Δ_{10}, the calculated restraints originate from a single structure (pure B or pure A) and do not contain any intrinsic contradictions. Therefore, they should not cause any difficulties for a conventional single-copy refinement. However, it remained to be proven that the algorithm presented here is capable of solving this problem as well. As shown in Table 3, both target ensembles could be found with 25 cycles of multiple-copy energy minimizations. In the case of $p(\mathrm{A}) = 1.0, p(\mathrm{B}) = 0.0$ (Δ_{10}), an ensemble consisting of two A-form-like structures was calculated. The first copy had probability of 1.0 and conformational energy of −28.4 kcal/mol; the second copy had probability of 0.0 and energy of −29.3 kcal/mol. Because the zero-probability copy does not contribute to E_{NMR}, it was essentially subjected to free energy minimization during the refinement; in fact, it is very close to the structure with minimum conformational energy. The refined copy with probability 1.0 is close to the target structure; its helical parameters and energy are intermediate between target A and optimal A structures (Tables 1 and 3).

With a pure B-form dimer as target, the ensemble obtained by multiple-copy minimization consisted of two different structures: an A-like and a B-like structure with the A-like structure having a population of 0.0. The outcome of both calculations, which differs from the result that one expects from multiple-copy refinement with equal weights, is easy to explain. The algorithm found that one copy is sufficient to minimize the restraint energy of the system. The second structure got a population of zero and remained close to the initial A-form-like structure. The minimization affected therefore only the conformational energy of this structure; it essentially underwent an unrestrained energy minimization.

A 50:50 mixture of A- and B-DNA dimer

In our calculations, a multiple-copy energy minimization against the target restraint set Δ_5 was immediately successful, resulting in an ensemble of an A-like structure and a B-like structure with populations $p(\mathrm{A}) = 51\%$ and $p(\mathrm{B}) = 49\%$. Figure 5 shows the values of populations of the two copies, sugar pseudorotation angle and energy terms in the course of the calculation. The plot of the sugar pseudorotation angles (Figure 5a) illustrates that the two copies are separated in the conformational space immediately in the first cycle of energy minimizations. While copy 2 remains in the A-family of forms, copy 1 adopts a conformation intermediate between A- and B-DNA already after the first cycle of minimization, with the sugar pseudorotation angle about 90° for both dG and dC residues. This conformation provides a relatively reasonable compromise for the ‘experimental’ data for the mixture of A- and B-forms; therefore, its population stayed close to 1.0 for 20 cycles of minimization (Figure 5b). The objective function Q^r (Figure 5d) and

Table 3. Refinement against relaxation rate sets $\Delta_0 - \Delta_{10}$

Set	$p(1)$	$p(2)$	$P_G(1)$	$P_C(1)$	$\Omega(1)$	$P_G(2)$	$P_C(2)$	$\Omega(2)$	$E_{conf}(1)$	$E_{conf}(2)$	Q^r	k_{NMR}	s
Δ_0	0.01	0.99	21.9	30.0	33.9	152.9	142.0	38.0	−29.3	−24.8	$2.2 \cdot 10^{-3}$	40	1.00
Δ_1	0.14	0.86	22.0	30.0	33.9	155.3	142.3	38.2	−29.0	−24.7	$3.2 \cdot 10^{-3}$	40	1.00
Δ_2	0.23	0.77	22.8	30.0	34.1	155.7	143.2	38.3	−29.0	−24.8	$2.5 \cdot 10^{-3}$	40	1.00
Δ_3	0.33	0.67	23.1	30.0	33.4	156.9	143.8	38.6	−28.8	−25.0	$2.7 \cdot 10^{-3}$	40	1.00
Δ_4	0.42	0.58	23.3	30.0	33.3	157.1	144.8	38.8	−28.66	−25.1	$2.7 \cdot 10^{-3}$	40	1.00
Δ_5	0.51	0.49	23.2	30.0	33.3	151.3	144.3	38.1	−28.5	−24.9	$1.6 \cdot 10^{-3}$	40	1.00
Δ_6	0.60	0.40	23.3	29.6	33.3	156.4	146.4	38.9	−28.4	−25.5	$2.4 \cdot 10^{-3}$	60	1.00
Δ_6	0.53	0.47	21.9	28.4	34.0	107.6	136.7	33.1	−27.4	−18.3	$1.1 \cdot 10^{-2}$	40	1.00
Δ_7	0.66	0.34	22.5	28.7	33.3	151.8	147.3	35.6	−28.1	−25.1	$2.0 \cdot 10^{-3}$	80	1.00
Δ_7	0.47	0.53	20.6	25.6	33.9	93.3	97.9	23.8	−27.8	−17.3	$3.1 \cdot 10^{-2}$	40	1.00
Δ_8	0.77	0.23	23.2	28.9	33.3	153.9	149.7	36.4	−28.3	−25.3	$1.7 \cdot 10^{-3}$	200	0.50
Δ_8	0.53	0.47	20.3	26.4	33.9	86.1	87.9	28.2	−28.2	−23.1	$3.8 \cdot 10^{-2}$	40	1.00
Δ_9	0.88	0.12	24.3	29.2	33.4	150.0	150.1	35.4	−28.3	−25.4	$1.1 \cdot 10^{-3}$	15 000	0.07
Δ_9	0.00	1.00	21.8	30.0	33.4	31.0	37.9	30.3	−29.3	−27.7	$4.7 \cdot 10^{-2}$	40	1.00
Δ_{10}	1.00	0.00	24.9	29.4	33.4	21.9	30.0	33.4	−28.4	−29.3	$7.0 \cdot 10^{-4}$	80	1.00
Δ_{10}	1.00	0.00	24.8	29.3	30.7	21.9	30.0	33.4	−28.4	−29.3	$7.4 \cdot 10^{-4}$	40	1.00

The results are shown for the two-copy refinement of structures against target ensembles calculated for different ratios of A- and B-DNA. Listed are sugar pseudorotation angles P_{G1} ($=P_{G2}$) and P_{C3} ($=P_{C4}$), helical twist Ω and conformational energy E_{conf} for both resulting conformers, and also the PDQPRO objective function Q^r (in Hz^2) for the ensemble. The structures were calculated by 25 cycles of energy minimizations. During the first 20 cycles, k_{NMR} and s were set to the values listed in the table; during the last five cycles, they were set to $k_{NMR} = 200$ and $s = 1.0$. Index '1' in parentheses refers to the first conformer, and index '2' to the second conformer.

restraint energy E_{NMR} (Figure 5c) decreased significantly for this intermediate ensemble compared to the pure A-form. At the same time, the conformational energy of the first copy increased only ca. 5 kcal/mol (Figure 5c). This is apparently a quite stable local minimum at $k_{NMR} = 40$; both copies change very little after the first cycle of minimization. The restraint force constant k_{NMR} had to be increased to 200 (note a discontinuity at minimization cycle 20, Figure 5) to force further refinement of the ensemble. The first copy surpassed a conformational barrier of ca. 10 kcal/mol, and then it rapidly relaxed to a minimum corresponding to the B-form with population close to 50%. Figure 5c illustrates that the conformational energies of both copies remain negative throughout the calculations. None of the copies had to surpass a large energy barrier in the course of the refinement. Similar to the case of refinement of pure B- or A-forms (see previous section), the resulting structures are intermediate between the target and minimum-energy conformations (Table 3). The reasons for this are discussed in the next section.

Weight of the restraint energy

To investigate the role of the weight of restraint energy relative to conformational energy, additional calculations were carried out using restraint set Δ_5. Starting with the ensemble refined as described in the previous section, five additional cycles of minimization were performed with k_{NMR} systematically varied in the range of 0 through 100 000 (Table 4). The starting ensemble already contained two structures in the A-DNA and B-DNA families, so no overcoming of energy barriers was required. Calculations with $k_{NMR} = 0.0$ are equivalent to unrestrained energy minimization of two non-interacting copies, A- and B-DNA. The resulting structures are very close to the minimum-energy conformations, and the populations fitting 'experimental' data best are p(A) = 62% and p(B) = 38%. Note that even though the refinement is no longer driven by experimental data with k_{NMR} = 0, the optimum populations of current structures are still calculated with PDQPRO based on experimental data. The 62-38 distribution could be obtained by unrestrained energy minimization of A- and B-DNA, and then applying a stand-alone PDQPRO program (Ulyanov et al., 1995). Any improvement over that is due to integration of PDQPRO with the refinement engine.

Increasing the weight of the restraint energy results in several effects. Objective function Q^r decreases monotonically; populations p(A) and p(B) approach the 50-50-target distribution; the resulting conforma-

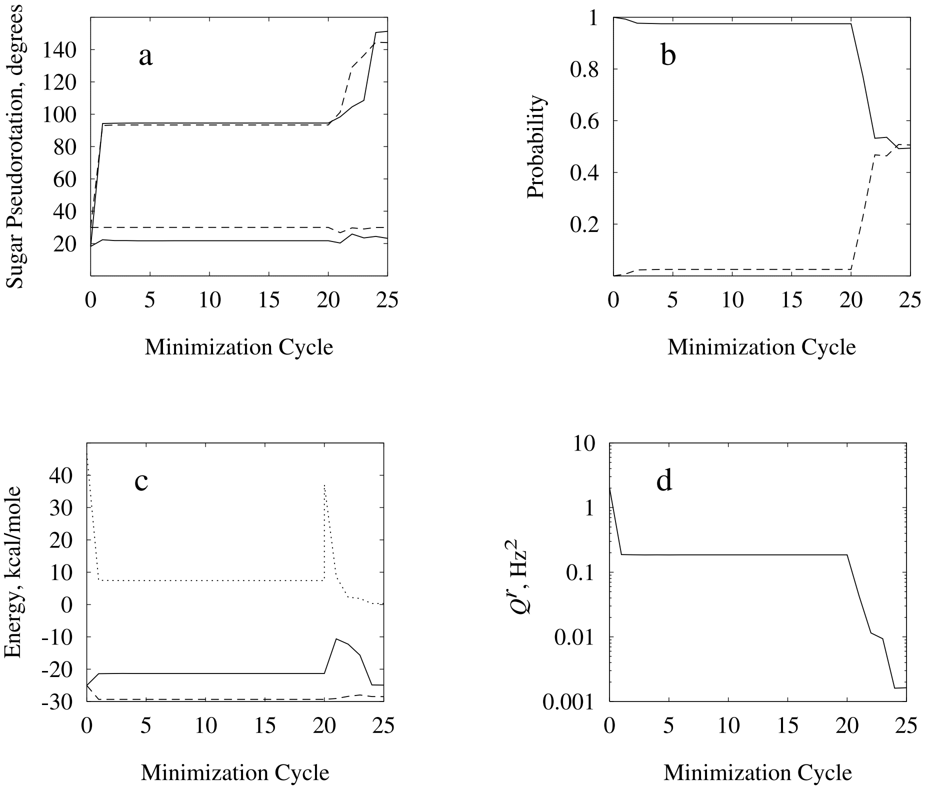

Figure 5. Key parameters of two copies of DNA dimer during restrained energy minimization. (a) Sugar pseudorotation angles P_G (solid lines) and P_C (dashed lines). (b) Population p for copy 1 (solid line) and copy 2 (dashed line). (c) Conformational energy E_{conf} for copy 1 (solid line) and copy 2 (dashed line), and restraint energy E_{NMR} for the ensemble (dotted line). (d) Objective function Q^r. The target ensemble consisted of a 1:1 mixture of A- and B-DNA. Note a discontinuity in E_{NMR} at minimization cycle 20 caused by a fivefold increase of force constant k_{NMR} at this step (see text).

tions approach the target structures; conformational energy moderately increases for both copies (Table 4).

Unavoidable discrepancies between the force field used in the calculations and the true potential experienced by a molecule in solution can have a severe impact on the outcome of the calculations. In practical situations, k_{NMR} should be selected high enough to minimize bias due to the empirical force field used. The upper limit on k_{NMR} depends largely on quality and consistency of experimental data, and it should be determined in each individual case. In the case of simulated ensemble Δ_5 with perfect ‘experimental’ data, it is possible to raise k_{NMR} infinitely high without compromising the quality of the resulting structures. With a relatively low value of $k_{NMR} = 100$, the target values for the sugar pseudorotation angle are reproduced within 10°, and for the helical twist within 3.5° (Table 4). At high values of k_{NMR}, pseudorotation angles could be reproduced almost exactly, but helical twist of the B-conformer was still off by ca. 1°. Such residual bias due to the force field is expected for the helical parameters which are not restrained directly by experimental data (Metzler et al., 1990; Ulyanov et al., 1992). Indeed, in the case of d(GG):d(CC) dimer, there are no cross-strand restraints in the list of observable parameters (Table 2). Helical parameters such as helical twist are expected to be defined better for the mixed sequences, due to the presence of adenine H2 protons in the minor groove of DNA (these protons participate in a number of cross-strand NOE cross peaks). In the case of conventional single-copy refinements of mixed-sequence DNA duplexes, bias due to the force field is not great; it has been shown that using different force fields results in very similar structures (Ulyanov et al., 1993; Tonelli et al., 1998). A similar result may be expected for multiple-copy refinements as well, but additional investigations are necessary, of course.

Table 4. Influence of the force constant k_{NMR}

$p(1)$	$p(2)$	$P_G(1)$	$P_C(1)$	$\Omega(1)$	$P_G(2)$	$P_C(2)$	$\Omega(2)$	$E_{conf}(1)$	$E_{conf}(2)$	Q^r	k_{NMR}
0.62	0.38	21.3	30.0	31.1	167.4	158.4	41.5	−29.4	−26.9	$4.4 \cdot 10^{-1}$	0
0.53	0.47	21.4	30.0	30.9	164.4	152.3	40.4	−29.0	−26.4	$5.7 \cdot 10^{-2}$	10
0.51	0.49	22.5	30.0	31.0	154.2	145.8	38.5	−28.6	−25.3	$3.2 \cdot 10^{-3}$	10^2
0.50	0.50	25.7	30.0	30.8	151.0	141.0	37.7	−28.3	−24.7	$5.0 \cdot 10^{-4}$	10^3
0.50	0.50	28.0	30.1	30.6	146.9	136.7	36.8	−28.1	−23.9	$5.3 \cdot 10^{-5}$	10^4
0.50	0.50	28.1	30.0	30.6	145.2	135.5	36.0	−27.5	−23.3	$8.5 \cdot 10^{-6}$	10^5

Five cycles of restrained minimization were performed against the restraint set Δ_5 with k_{NMR} listed in the table and $s = 1.0$ (see text).

Other mixtures of A- and B-DNA

It must be expected that convergence or failure of a given refinement protocol depends on starting structures. If both starting copies are close to the minor fraction of the target ensemble, the restraints can easily push one of the structures near to the major fraction. This problem is very much similar to the refinement of the single structure starting with the 'wrong' conformation. However, if both initial structures are near the major fraction, the restraint energy E_{NMR} is already relatively low. In this case, the restraints push the copies only in the general direction of the minor fraction, and the refinement is more likely to be trapped in a local minimum of the conformational energy.

Starting from an A-like structure, a target ensemble with the A-form dimer as the minor fraction (sets Δ_1–Δ_4) could be determined without changing the protocol described above (Table 3). The calculations essentially showed a behavior similar to the refinement against restraint set Δ_5.

However, with the same refinement protocol, it was impossible to reproduce target ensembles with the B-form as the minor fraction (sets Δ_6–Δ_9). Most often, the refinement resulted in the second copy having a conformation intermediate between A- and B-DNA. The values of the restraint weight k_{NMR} which were sufficiently high to overcome the conformational energy barrier in the case of Δ_5 (Figure 5) were not effective for the sets Δ_6–Δ_9. Indeed, the same value of k_{NMR} resulted in much lower values of restraint energies E_{NMR}, because the first copy was already close to the major conformer (A-form). It is important to mention that no a priori knowledge of the target structure was required to tell that the refinement failed to reproduce the target ensemble. This could be recognized by the mere fact that either the objective function Q^r or the conformational energies were high for the resulting structures. To achieve a successful convergence, the slope s (Equation 10) of the restraint energy term E_{NMR} had to be decreased and k_{NMR} had to be increased simultaneously (Table 3). With parameters of E_{NMR} thus modified, minor fractions of B-DNA in the target ensemble as low as 10% could be reproduced.

In all cases, the resulting structures have conformations intermediate between the target and minimum-energy conformations (Table 3). Additional minimizations with increased weight k_{NMR} had a similar effect on the refinement of the Δ_1 set, as described above for the Δ_5 set. Namely, the target structures were reproduced more closely (data not shown).

For the reasons discussed above, successful determination of minor fractions could not have been expected in advance. To succeed, the algorithm must find a valley in the conformational energy landscape that connects the two conformers, and the restraints push the structure gently through this valley. A better than expected performance of the algorithm can be explained by two reasons. We assume that the open shape of the restraint potential (Figure 2) helps find a path between conformers. In addition, the fact that structures are parameterized via helical parameters reduces the number of local minima in the conformational space significantly. This parameterization may also be important for successful convergence.

Calculations with a wrong number of copies

To investigate if the algorithm depends on a priori knowledge of the number of conformers in the target ensemble, additional energy minimizations were performed using the target ensemble Δ_5. Calculations followed the same protocol as described above but were carried out with a single copy only, as well as with three copies (Table 5). Not surprisingly, the result of the single copy refinement was a compromise structure between A- and B-form with sugar pseudorotation angle of ca. 95° for both dG and dC residues. A

Table 5. Influence of wrong number of copies in ensemble

Ensemble	Copy number	p	P_G	P_C	Ω	E_{conf}	Q^r
One-copy	1	1.00	94.1	96.5	28.2	-13.8	$8.8 \cdot 10^{-2}$
Three-copy	1	0.50	154.7	145.6	35.7	-24.3	$1.9 \cdot 10^{-3}$
	2	0.50	23.9	30.0	30.8	-28.5	
	3	0.00	21.9	30.0	33.3	-29.3	

The results are shown for one- and three-copy refinements against the restraints set Δ_5. Parameters listed are defined in the legend to Table 3.

high conformational energy of -13.8 kcal/mol and a high PDQPRO objective function $Q^r = 0.1$ Hz^2 imply that unresolved contradictions between 'experimental' restraints and the conformation of the structure still exist. A refinement with three copies produced two structures in A-form and one structure in B-form. From the two A-forms, the structure with the lower conformational energy had a population of zero, indicating that it was not essential for solving the problem and could be omitted.

Number of observable parameters and relaxation rate errors

The presence of conflicting restraints is the only experimental information used by our algorithm to calculate more than one distinct conformer. In a typical NOESY spectrum, two sets of interproton distances are measured, which can help distinguish between A-form, B-form, and their mixture (Ulyanov et al., 1998). One set includes distances which are normally short for B-DNA, such as intra-residue H2′-H6/H8 and H2″-H6/H8 and sequential H2″-H6/H8. The other set includes intra-residue and sequential H3′-H6/H8 and intra-residue H4′-H2″; they are short in A-DNA. In no single structure can all of these distances be short simultaneously, without seriously compromising its conformational energy. However, all these distances must appear short for the mixture of A- and B-forms, due to r^{-6}-averaging. If the list of observable parameters lacks some or all of these conflicting restraints, it is unlikely that the algorithm would successfully recover the contributing conformations. Indeed, this was confirmed by test calculations with the Δ_5 set that had a certain number of restraints randomly removed. Among 20 randomly prepared subsets with 10% of the restraints removed, only 70% of the subsets led to successful refinement. However, when 40% of the restraints were removed (7.0 restraints per residue left), only 35% of the refinements were successful (data not shown). For this purpose, we define a refinement as successful if it resulted in two distinct conformers, one in the B-family of forms, another in the A-family. The same refinement protocol was used as before (Table 3). It is conceivable that the rate of successful refinements could be improved by modifying the protocol, but the trend is likely to remain.

In a different test, we kept constant the number of restraints in the Δ_5 set (11.5 restraints per residue), but added random errors to the 'observed' relaxation rates. Increase in the relative random errors caused a systematic decrease in the rate of successful refinements (Figure 6a). When the 'observed' parameters had errors of up to 60%, only half of the random data sets led to successful determination of A- and B-conformers. Furthermore, even among successful refinements, the precision of the resulting conformers deteriorates when the amount of errors increases (Figure 6). Nevertheless, the calculated structures, on average, still have conformations intermediate between the target and minimum-energy conformations, similar to the case of error-free data (see above).

Three discrete minima

The behavior of the algorithm in the case of three minima was investigated using d(GGGGG):d(CCCCC) as test molecule. The pentamer was refined against the restraint sets Π_{333}, Π_{226} and Π_{316}, calculated from target structures S_{28}, S_{83} and S_{150} described above (see Methods). In contrast to the studies on the dimer d(GG):d(CC), where *global* A- and B-conformations were considered, here we deal with the *local* conformers of the G3 residue (C3′-*endo*, O4′-*exo*, and C2′-*endo* for the three conformers, respectively). A number of different refinement protocols were tested, all being started with an ensemble of identical copies of structure S_{28}.

At first, we tested a simple restrained minimization protocol similar to the one used for the refinement of

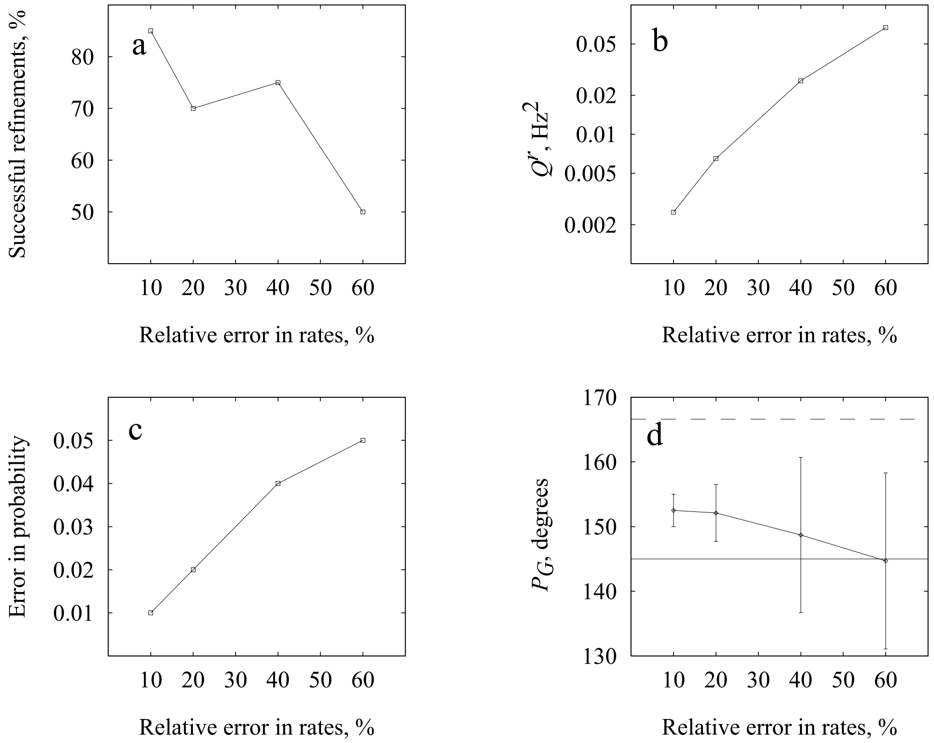

Figure 6. Influence of random errors in observed parameters on two-copy refinement of the Δ_5 set. (a) Ratio of successful refinements (see text). (b) Average objective function Q^r. (c) Average difference between calculated and target probabilities. (d) Average values and standard deviations for the pseudorotation angle P_G in the B-conformer. In addition, the value of this parameter is shown for the target structure (solid line) and minimum-energy structure (dashed line).

the dimer (sets Δ_0–Δ_{10}, Table 3). However, in this case, three copies instead of two were refined simultaneously. For each of the three restraint sets, however, the population of 0.0 was calculated for one of the copies in the refined ensemble (Table 6, protocol ‘a’). For the Π_{226} and Π_{316} restraint sets, the two remaining copies reproduced target structures S_{28} and S_{150} and the population of the major conformer (S_{150}). For the Π_{333} set, the conformer S_{150} was not found at all; instead, target structures S_{28} and S_{83} were reproduced with populations 0.39:0.61 (Table 6, protocol ‘a’).

In another test, a more sophisticated protocol was used, which included both restrained minimization and Metropolis Monte Carlo simulated annealing. At first, 25 cycles of restrained minimization were carried out, with k_{NMR} being exponentially increased from 200 to 10 000. These were followed by 10 000 Monte Carlo steps when the temperature gradually increased from 300 to 600 K, 10 000 steps at 600 K, 10 000 steps when the system was cooled back to 300 K, and another 10 000 steps at 300 K. During the last 10 000 steps, the copies were averaged based on helical parameters (Ulyanov et al., 1993) and restraint-minimized with k_{NMR} being decreased back from 10 000 to 200. The calculations started with initial ensembles, similar to the ones obtained as described above, with the first copy duplicated. During the Monte Carlo phases, the first copy was not shaken but was taken into account as a member of the ensemble with its population floating. The results of these calculations are summarized in Table 6 (protocol ‘b’). In short, only in the case of the Π_{333} restraint set were the three target structures reproduced with roughly correct populations. Refinements against restraint sets Π_{226} and Π_{316} did not produce the ‘intermediate’ conformer S_{83}.

In all these cases, the residual objective function Q^r is relatively small (Table 6). Obviously, two copies are sufficient to resolve most of the major contradictions in the restraints. The remaining contradictions are insufficient to push the third copy over the barriers in conformational energy using the restrained minimization protocol. The situation is similar, in a way, to refinement of the Δ_9 set, when the major conformer alone could explain ‘experimental’ data reasonably

Table 6. Refinement of pentamer with three discrete minima

Refinement protocol	Restraint set	$p(1)$	$p(2)$	$p(3)$	$P_{G3}(1)$	$P_{G3}(2)$	$P_{G3}(3)$	Q^r
a	Π_{333}	0.00	0.39	0.61	27.3	27.6	93.8	$7.2 \cdot 10^{-3}$
	Π_{226}	0.00	0.33	0.67	27.9	27.9	142.0	$1.2 \cdot 10^{-3}$
	Π_{316}	0.00	0.37	0.63	27.6	27.6	146.3	$5.8 \cdot 10^{-4}$
b	Π_{333}	0.42	0.29	0.29	29.6	90.1	150.0	$7.4 \cdot 10^{-4}$
	Π_{226}	0.32	0.52	0.16	30.6	145.6	149.6	$8.1 \cdot 10^{-4}$
	Π_{316}	0.37	0.49	0.14	28.0	144.8	150.0	$3.3 \cdot 10^{-4}$

Protocol ‘a’ consisted of restrained minimization; protocol ‘b’ consisted of a combination of restrained minimization and Monte Carlo simulated annealing (see text).

well. In that case, the force constant k_{NMR} had to be significantly increased for successful refinement (Table 3). It is very likely that by increasing k_{NMR} even further or modifying the refinement protocol in some other way, it would be possible to reproduce all target structures for the error-free Π_{226} and Π_{316} data sets. For example, systematically varying sugar conformations in each individual residue is a very efficient search method (Gorin et al., 1990; Ulyanov et al., 1995). It is clear, however, that in real situations with experimental errors, the set of observable parameters used (Table 2) is not sufficient to define uniquely all three hypothetical conformers. Using additional experimental data may change this situation. Such additional data may include proton-proton J-scalar couplings for the sugar moieties (Rinkel and Altona, 1987; Mujeeb et al., 1992; Conte et al., 1996), phosphorus-proton couplings for the backbone (Gorenstein, 1994; McAteer et al., 1998; Tisne et al., 1999), or even NOE data involving stereo-specifically assigned H5′ and H5″ protons (Hines et al., 1993).

Single continuous minimum

All calculations described so far entailed the target structures at discrete positions in conformational space. In real situations, however, it is likely that conformations have continuous distributions. For example, experimental H1′-H6/H8 NOE data measured for a number of DNA duplexes have been explained in terms of restricted diffusion, which involved correlated sugar repuckering and glycosidic torsion angle χ rotation (Tonelli and James, 1998). To study how the present algorithm performs with restraints derived from a single continuous minimum, energy minimizations were performed against the restraint set Π_{broad}, generated for the pentamer. The target ensemble included conformations near structure S_{150}. The target probabilities were calculated using an artificial harmonic potential, which approximated the energy profile (Figure 4) in the corresponding local minimum (see Materials and methods).

For the calculations, the same refinement protocol was used as for the dimer with $k_{NMR} = 200$ and $s = 1.0$. Starting from the structure S_{28}, restrained energy minimizations were initiated with one copy as well as with three copies. With one copy, the resulting structure had helical parameters close to the expectation values of the system. With three copies, only one copy near the S_{150} minimum resulted. The other two copies remained in the neighborhood of the starting structure, but their populations were below 1%. To investigate if the result depends on the starting structure, a three-copy energy minimization was carried out with S_{150} as the starting structure. Again, a single copy was sufficient to explain most of the data. Although calculated from multiple structures, the restraint set Π_{broad} is essentially free of internal contradictions.

Three continuous minima

The last test system studied combined the complications of a continuous distribution of structures with three distinct conformations. The set of restraints Π_{full} (see Materials and methods) was based on the actual energy profile shown in Figure 4. Although the details of this energy profile are probably incorrect, we still consider it the most realistic model of the conformational distribution of DNA in solution. We used the same combination of restrained minimization and Monte Carlo simulated annealing for the refinement against this data set, as earlier for the refinement of three discrete minima. The results are summarized in Figure 4.

At first, we carried out the refinement with the single copy, starting with either global A- or global B-conformation. The two refinements converged; the resulting structure is B-DNA with sugar pseudorotation P_{G3} between 101° and 130° (shown by a single horizontal line in Figure 4). The structures have conformational energy between -109.2 and -106.4 kcal/mol and the residual Q^r is between 0.035 and 0.049.

At the second step, the refinement was repeated with two copies, starting with four different initial ensembles. Each starting ensemble had two identical copies of the pentamer. In two of them, the two structures calculated at the previous step were duplicated. In two other, the refinement started with two identical global A- or two global B-conformations. Refinement that started with the ensemble of two A-conformations was trapped in a high-energy local minimum. The remaining three refinements converged, producing very similar ensembles (shown by double horizontal lines in Figure 4). Residual Q^r decreased by an order of magnitude (0.0036–0.0054), and the conformational energy decreased to the range of -118.7 to -113.4, for each copy. (Note that the resulting conformational energies are even lower than energies of the target distribution. This is because the target distribution shown in Figure 4 was calculated with the terminal base pairs of the pentamer fixed; see Materials and methods. However, the refinement was carried out with fully flexible pentamer, and certain improvement in energy came because of optimization of terminal base pairs.)

Additional refinements with three copies failed to produce an intermediate conformation with the O4′-*exo* sugar pucker for the G3 residue. In short, the results are similar to the case of three discrete minima discussed above. However, in this case, the population of the major conformer is reproduced less accurately: it can be off by as much as a factor of two (Figure 4).

Conclusions

The algorithm presented here, multiple-copy refinement with floating populations, is designed to determine multiple structural conformers of nucleic acids and their populations based on NMR data. With the current version of the program, it takes about 2.5 min of CPU time on an O2 SGI with R5000 processor to refine two copies of the dimer, using the protocol described in Table 3. The protocol entailed 25 cycles of minimization, with about 5000 evaluations of energy per copy in each cycle. The refinement of three copies of the pentamer takes about 4 h of CPU time on the same computer, with 40 min spent on 25 cycles of minimization, and about 3 h spent on 40 000 iterations of Metropolis Monte Carlo.

The algorithm was tested on a variety of simulated data. It can successfully calculate target global A- and B-conformations of DNA and their target populations using NOE data typically observed in NOESY spectra. Despite the apparent triviality of this problem, it could not be solved by other methods (Ulyanov et al., 1998), even when both conformations had significant populations. Surprisingly, the new algorithm could solve this problem even if one of the conformers had a population as low as 10%. We believe two factors contribute to the success of this algorithm. (1) Use of internal coordinates, generalized helical parameters, in the refinement engine significantly reduces the number of degrees of freedom in the system, and it simplifies the surface of *conformational energy*. (2) Using floating rather than fixed populations makes the topology of the *restraint potential* open. This facilitates search of low-energy passage between various local minima.

Using the same type of observable parameters, the algorithm can as well determine local conformations, C2′-*endo* and C3′-*endo*, for individual residues within the framework of B-DNA. However, the presence of a potential third conformer, O4′-*exo*, complicates the situation. This conformer could be determined only if its population was sufficiently high (33%). Furthermore, low populations of the O4′-*exo* conformer introduced errors in the determination of populations of the two main conformers, C3′-*endo* and C2′-*endo*. These errors were especially severe if each of the three conformers had a broad continuous distribution. This result is not surprising. The O4′-*exo* conformation of the sugar moiety is, in many respects, intermediate between C3′-*endo* and C2′-*endo*. Correspondingly, many (although not all) observable parameters also have intermediate values. Commonly observed NOE data are simply not sufficient to define uniquely all three local conformations. The situation may change if additional information, such as scalar coupling data, is available; this requires additional investigation.

The algorithm is relatively insensitive to random errors in the observed parameters. However, lack of a sufficient number of restraints could be detrimental for its performance. Internally inconsistent restraints constitute the sole experimental information used by the algorithm to calculate distinct structural conformers. Presence of such inconsistencies in the experimental data must be assessed before attempting actual refinement of multiple conformers. Some approaches for such an assessment have been recently reviewed by Mujeeb et al. (1999). In some (probably, rare) cases, the presence of multiple conformers can be ascertained from plain inspection of observed NOE cross peaks. For example, it has been found for a 17-nucleotide RNA that an adenine H2 proton from the loop region had cross peaks with protons of five different residues (Yao et al., 1997; Schmitz et al., 1998). Such an observation could only be explained by the presence of distinct conformers, which must exist long enough to give rise to the observed NMR signal. More typically, a conventional single-structure refinement should be attempted first, using as accurate structural restraints as possible. If all experimental restraints are satisfied with a single conformation, there is no justification for attempting a multiple-copy refinement. If, on the other hand, some experimental restraints are systematically violated and/or conformational energy is compromised, this may be an indication of multiple conformers contributing to experimental restraints. Because the performance of the algorithm is so dependent on the type of conformers involved and experimental data available, it is advisable to carry out test calculations with simulated data modeling each experimental system studied.

Due to use of DNAminiCarlo as the refinement engine, the application of our program is limited to nucleic acids only. However, the principles of this algorithm may have wider applications. We are planning to apply this method to a number of experimental nucleic acid systems, where data suggest the presence of distinct conformers.

Acknowledgements

This work was supported by NIH grants GM39247 and RR01081. A.G. acknowledges support through a fellowship in the framework of the ‘Gemeinsames Hochschulsonderprogramm III’ of the Federal Government of Germany, mediated through the German Academic Exchange Service (DAAD).

References

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

Bonvin, A.M. and Brünger, A.T. (1995) *J. Mol. Biol.*, **250**, 80–93.

Bonvin, A.M. and Brünger, A.T. (1996) *J. Biomol. NMR*, **7**, 72–76.

Borgias, B.A. and James, T.L. (1989) *Methods Enzymol.*, **176**, 169–183.

Borgias, B.A. and James, T.L. (1990) *J. Magn. Reson.*, **87**, 475–487.

Chandrasekaran, R., Wang, M., He, R.-G., Puigjaner, L.C., Byler, M.A., Millane, R.P. and Arnott, S. (1989) *J. Biomol. Struct. Dyn.*, **6**, 1189–1202.

Conte, M.R., Bauer, C.J. and Lane, A.N. (1996) *J. Biomol. NMR*, **7**, 190–206.

Dickerson, R.E., Bansal, M., Calladine, C.R., Diekmann, S., Hunter, W.N., Kennard, O., Lavery, R., Nelson, H.J.C., Saenger, W., Shakked, Z., Sklenar, H., Soumpasis, D.M., von Kitzing, E., Wang, A.-H.-J. and Zhurkin, V.B. (1989) *EMBO J.*, **8**, 1–4.

Fennen, J., Torda, A.E. and van Gunsteren, W.F. (1995) *J. Biomol. NMR*, **6**, 163–170.

Fletcher, R. (1981) *Practical Methods of Optimization, Vol. 2: Constrained Optimization*, Wiley & Sons, New York, NY.

González, C., Stec, W., Reynolds, M.A. and James, T.L. (1995) *Biochemistry*, **34**, 4969–4982.

Gorenstein, D.G. (1994) *Chem. Rev.*, **94**, 1315–1338.

Gorin, A.A., Ulyanov, N.B. and Zhurkin, V.B. (1990) *Molek. Biol. (Engl. transl.)*, **24**, 1036–1047.

Görler, A. and Kalbitzer, H.R. (1997) *J. Magn. Reson.*, **124**, 177–188.

Hines, J.V., Varani, G., Landry, S.M. and Tinoco Jr., I. (1993) *J. Am. Chem. Soc.*, **115**, 11002–11003.

Keepers, J.W. and James, T.L. (1984) *J. Magn. Reson.*, **57**, 404–426.

Kemmink, J. and Scheek, R.M. (1995) *J. Biomol. NMR*, **6**, 33–40.

Landy, S.B. and Rao, B.D.N. (1989) *J. Magn. Reson.*, **81**, 371–377.

McAteer, K., Jing, Y., Kao, J., Taylor, J.S. and Kennedy, M.A. (1998) *J. Mol. Biol.*, **282**, 1013–1032.

Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, N.M., Teller, A.H. and Teller, E. (1953) *J. Chem. Phys.*, **21**, 1087–1092.

Metzler, W.J., Wang, C., Kitchen, D.B., Levy, R.M. and Pardi, A. (1990) *J. Mol. Biol.*, **214**, 711–736.

Mujeeb, A., Kerwin, S.M., Egan, W., Kenyon, G.L. and James, T.L. (1992) *Biochemistry*, **31**, 9325–9338.

Mujeeb, A., Ulyanov, N.B., Billeci, T.M., Farr-Jones, S. and James, T.L. (1999) In *Biological Magnetic Resonance, Vol. 17: Structure Computation and Dynamics in Protein NMR* (Eds., Krishna, N.R. and Berliner, L.J.), Kluwer Academic/Plenum Publishers, New York, NY, pp. 201–222.

Rinkel, L.J. and Altona, C. (1987) *J. Biomol. Struct. Dyn.*, **4**, 621–649.

Schmitz, U., González, C., Ulyanov, N.B., Blocker, F.H., Liu, H. and James, T.L. (1996) In *Biological Structure and Dynamics, Vol. 2* (Eds., Sarma, R.H. and Sarma, M.H.), Adenine Press, New York, NY, pp. 165–187.

Schmitz, U., Donati, A., James, T.L., Ulyanov, N.B. and Yao, L. (1998) *Biopolymers*, **46**, 329–342.

Tisne, C., Hartmann, B. and Delepierre, M. (1999) *Biochemistry*, **38**, 3883–3894.

Tonelli, M. and James, T.L. (1998) *Biochemistry*, **37**, 11478–11487.

Tonelli, M., Ragg, E., Bianucci, A.M., Lesiak, K. and James, T.L. (1998) *Biochemistry*, **37**, 11745–11761.

Torda, A.E., Scheek, R.M. and van Gunsteren, W.F. (1990) *J. Mol. Biol.*, **214**, 223–235.

Torda, A.E., Brunne, R.M., Huber, T., Kessler, H. and van Gunsteren, W.F. (1993) *J. Biomol. NMR*, **3**, 55–66.

Ulyanov, N.B., Gorin, A.A. and Zhurkin, V.B. (1989) In *Proc. Int. Conf. Supercomp. ’89: Supercomputer Applications* (Eds., Kartashev, L.P. and Kartashev, S.I.), Int. Supercomputing Inst., Inc., St. Petersburg, FL, pp. 368–370.

Ulyanov, N.B., Gorin, A.A., Zhurkin, V.B., Chen, B.-C., Sarma, M.H. and Sarma, R.H. (1992) *Biochemistry*, **31**, 3918–3930.

Ulyanov, N.B., Schmitz, U. and James, T.L. (1993) *J. Biomol. NMR*, **3**, 547–568.

Ulyanov, N.B. and James, T.L. (1995) *Methods Enzymol.*, **261**, 90–120.

Ulyanov, N.B., Schmitz, U., Kumar, A. and James, T.L. (1995) *Biophys. J.*, **68**, 13–24.

Ulyanov, N.B., Mujeeb, A., Donati, A., Furrer, P., Liu, H., Farr-Jones, S., Konerding, D., Schmitz, U. and James, T.L. (1998) In *ACS Symposium Ser. 682: Molecular Modeling of Nucleic Acids* (Eds., Leontis, N.B. and SantaLucia Jr., J.), American Chemical Society, Washington, DC, pp. 181–194.

Yao, L.J., James, T.L., Kealey, J.T., Santi, D.V. and Schmitz, U. (1997) *J. Biomol. NMR*, **9**, 229–244.

Zhurkin, V.B., Lysov, Yu.P. and Ivanov, V.I. (1978) *Biopolymers*, **17**, 377–412.

Zhurkin, V.B., Poltev, V.I. and Florentiev, V.L. (1981) *Molek. Biol. (Engl. transl.)*, **14**, 882–895.

Zhurkin, V.B., Ulyanov, N.B., Gorin, A.A. and Jernigan, R.L. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 7046–7050.