Journal of Biomolecular NMR, 2 (1992) 467–476
ESCOM

J-Bio NMR 076

Simulation of NOESY spectra of DNA segments: A new scaling procedure for iterative comparison of calculated and experimental NOE intensities

R. Nibedita, R. Ajay Kumar, A. Majumdar and R.V. Hosur*

Chemical Physics Group, Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400 005, India

Received 6 January 1992
Accepted 9 June 1992

Keywords: SIMNOE; NOESY simulation; DNA structure; Intensity scaling

SUMMARY

A new algorithm for simulation of two-dimensional NOESY spectra of DNA segments has been developed. For any given structure, NOE intensities are calculated using the relaxation matrix approach and a new realistic procedure is suggested for 1:1 comparison of calculated and experimental intensities. The procedure involves a novel method for scaling of calculated NOE intensities to represent volumes of digitised cross peaks in NOESY spectra. A data base of fine structures of all the relevant cross peaks with Lorentzian line shapes and in-phase components, is generated in a digitised manner by two-dimensional Fourier transformation of simulated time domain data, assuming a total intensity of 1.0 for each of the cross peaks. With this procedure, it is shown that the integrated volumes of these digitised cross peaks above any given threshold scale exactly as the total intensity of the respective peaks. This procedure eliminates the repetitive generation of digitised cross peaks by two-dimensional Fourier transformation during the iterative process of structure alteration and NOE intensity calculation and thus enhances the speed of DNA structure optimization. Illustrative fits of experimental and calculated spectra obtained using the new procedure are shown.

INTRODUCTION

Two-dimensional (2D) nuclear Overhauser enhancement spectroscopy (NOESY) plays a major role in determination of structures of biological macromolecules in aqueous solutions (Anil Kumar et al., 1980, 1981; Wüthrich, 1986). The cross-peak intensities in NOESY spectra carry the crucial intramolecular interproton distance information and substantial effort has been focussed in recent years on extracting this information by proper interpretation of the cross-peak intensities (Banks et al., 1989; Landy and Rao, 1989; Baleja et al., 1990a; Borgias et al., 1990; Majumdar

* To whom correspondence should be addressed.

and Hosur, 1990). The interproton distance information constitutes the basic input for structure determination algorithms based on distance geometry and/or molecular dynamics calculations. Thus, the NOESY spectrum has often been referred to as a 'finger print' of the 3D structure and structures of several proteins and nucleic acid segments have been determined to date, to different levels of accuracies, by this method (see reviews: Hosur et al., 1988; van de Ven and Hilbers, 1988; Clore and Gronenborn, 1989, 1991; Wagner, 1990; Wüthrich, 1989a,b). It is now believed that simulating the NOESY spectra with some starting structural models, matching them with experimental spectra and iteratively refining the structural model until a best fit of calculated and experimental intensity patterns is obtained, is the most reliable approach to structure determination by NMR (Keepers and James, 1984; Baleja et al., 1990a; Borgias and James, 1990; Borgias et al., 1990; Gochin et al., 1990; Majumdar and Hosur, 1990, 1991; Mertz et al., 1991). Simulations employ relaxation matrix calculations for obtaining NOE intensities between various protons in the molecule and thus take care of spin-diffusion effects which significantly influence the cross-peak intensities in NOESY spectra of macromolecules (Macura and Ernst, 1980). The quality of a structure derived in this way depends on the number of NOE cross peaks, the accuracy of intensity measurements of the cross peaks, the number of variable torsional angles in the molecule, the extent of segmental motions within the molecule and the efficacy and the user friendliness of the simulation algorithm. Thus, several algorithms such as CORMA, MARDIGRAS, COMATOSE (Borgias and James, 1990; Borgias et al., 1990), IRMA (Boelens et al., 1989), BKCALC (Banks et al., 1989), DINOSAUR (Bonvin et al., 1991) are available today for NOESY simulation purposes. These are used in conjunction with either distance geometry or molecular dynamics or molecular mechanics calculations to refine the molecular structures (Banks et al., 1989; Nerdal et al., 1989; Baleja et al., 1990a,b; Gochin and James, 1990).

A crucial step in the iterative structure determination process is the matching of experimental and calculated NOE intensities. Both intensities are normalised in some way and the difference between them is minimised for obtaining the best fit and the best structure. Some authors used a scaling procedure where the observed intensities are multiplied by a factor equal to the ratio of sum of the intensities of all the NOEs calculated to the sum of all the corresponding observed intensities (Baleja et al., 1990a,b; Gochin et al., 1990). Zhou et al. (1987) used a different procedure for normalization of the peaks in experimental and calculated spectra. Each peak was scaled by dividing its intensity (the intensity of a peak is estimated by summing up all the points spanned by the peak) by the total of all the points in the columns that contain this peak. Computed intensities are also scaled similarly by considering all the peaks originating from a given proton and dividing each NOE by the sum of all the NOEs and the auto peak intensity of the particular proton. However, both these approaches do not address the basic problem which is as follows. While the relaxation matrix calculations yield 'analog' values for the intensities, the experimental peaks are 'digital' in nature, have fine structures with overlapping components, contain noise and have finite line widths along the two dimensions of the 2D spectrum. Further, the intensities and shapes of the experimental peaks are influenced by the data processing parameters such as window functions, digital resolutions, phase corrections etc. and the noise levels in different parts of the 2D spectrum are not necessarily identical. Therefore, the calculated 'analog' and experimental 'digital' intensities cannot be directly compared, and we believe that for a proper 1:1 comparison of the experimental and calculated intensities it is first absolutely essential to convert the 'analog' intensities into 'digital' intensities by simulating the 2D spectra with identical digital resolution and

line-width conditions as in experimental spectra. The straightforward method to achieve this is to generate the simulated time-domain data from the knowledge of chemical shifts, coupling constants and the calculated NOE intensities; Fourier-transform the data using identical parameters as used in processing experimental spectra and finally integrate these digital peaks in the same fashion as is done for the experimental peaks. Such an exercise has to be repeated every time the structure is altered in the iterative structure optimisation process. In this paper we describe a novel intensity scaling procedure for comparison of experimental and calculated NOE intensities, which eliminates the repetitive digitisation exercise, and thus significantly enhances the speed of structure optimization. A simulation algorithm, SIMNOE, has been developed incorporating this new procedure.

THE SIMULATION ALGORITHM

Figure 1 shows a general outline of the iterative structure determination process using the SIMNOE algorithm. The figure is self-explanatory and the new feature of the algorithm, namely, scal-

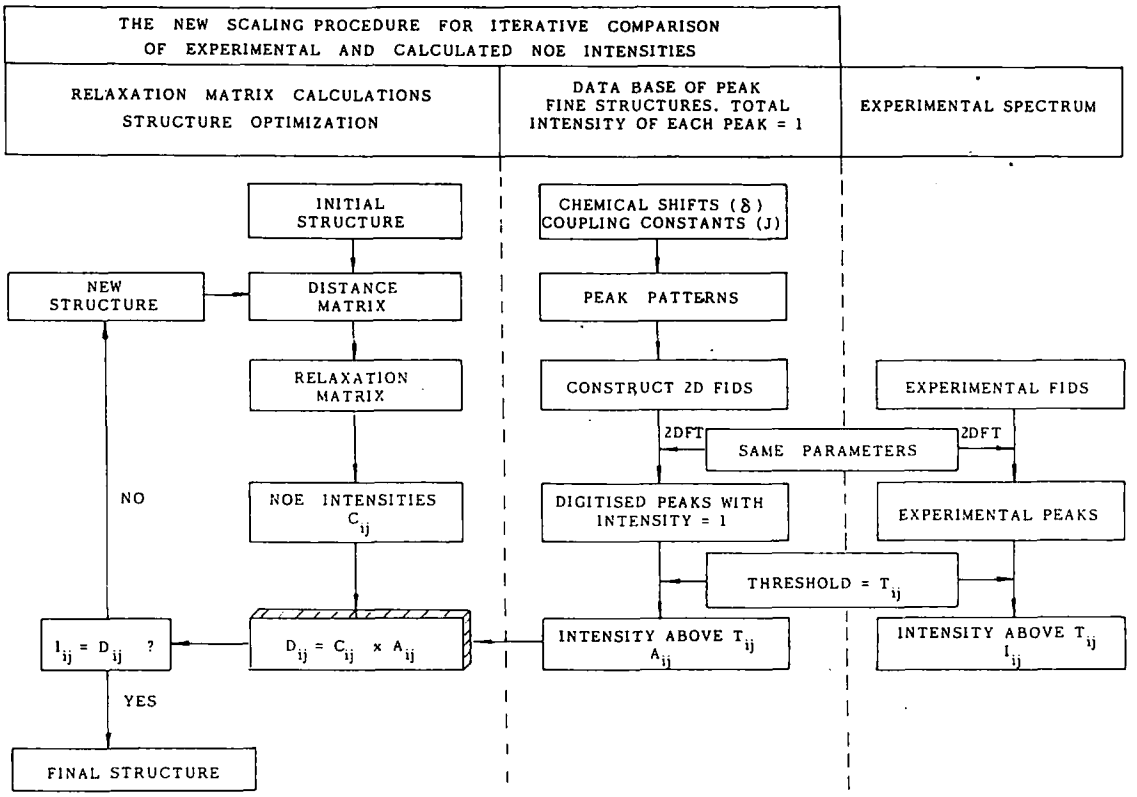


Fig. 1. Flow chart of operation of the SIMNOE program. The entire procedure can be divided into three blocks and the connections between these are indicated by horizontal arrows. Chemical shifts (δ) and coupling constants (J) are used as inputs for the calculation of position and fine structures of the cross peaks. The new scaling procedure for calculated NOE intensities is explicitly indicated. The thick box shows the arithmetic operation of converting computed 'analog' intensities into 'digitised' intensities. The simulation procedure involves an auxiliary DNA modelling program, MODEST (Ajay Kumar, R., to be published), developed on an IRIS 4D/70G workstation.

ing of calculated NOE intensities to represent digitised peaks is explicitly indicated in the top box and also the arithmetic scaling operation is enclosed in the thick box inside the figure. The salient features of the procedure are highlighted in the following paragraphs.

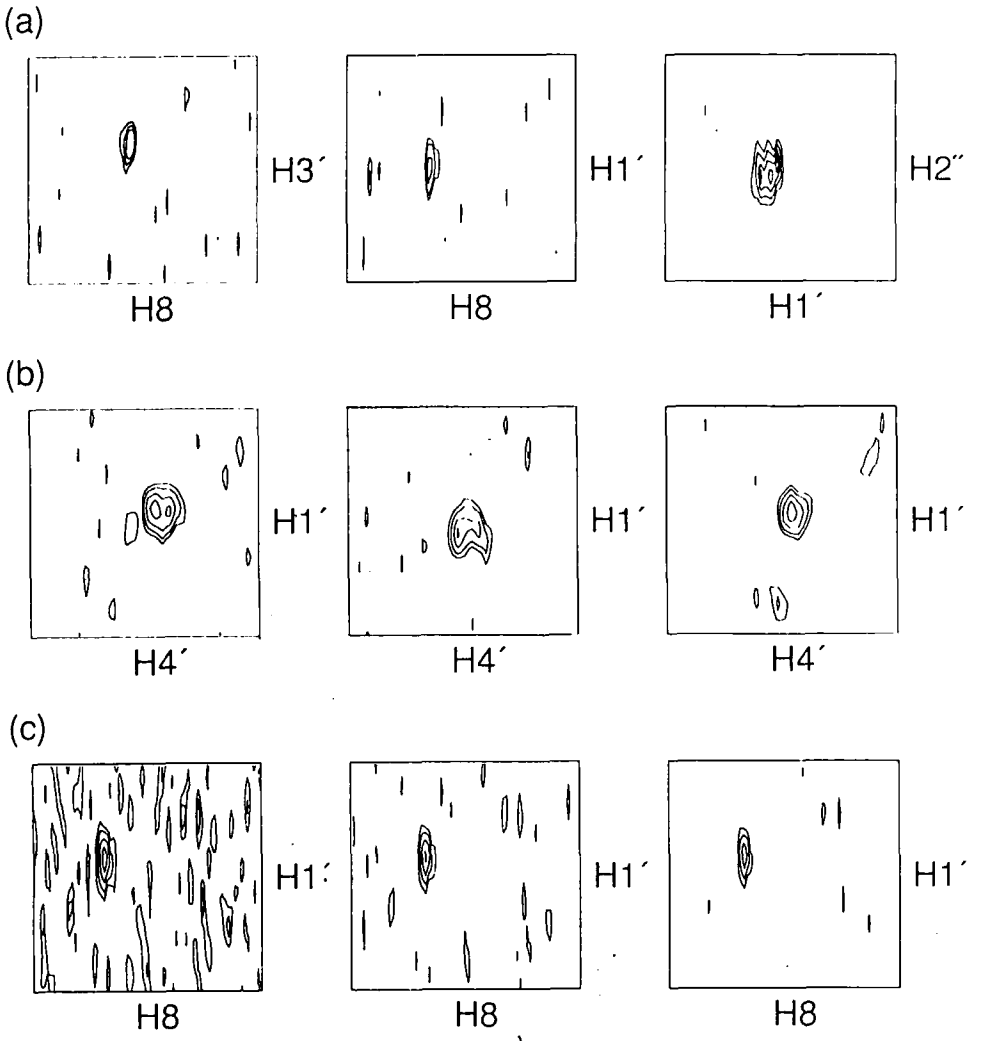
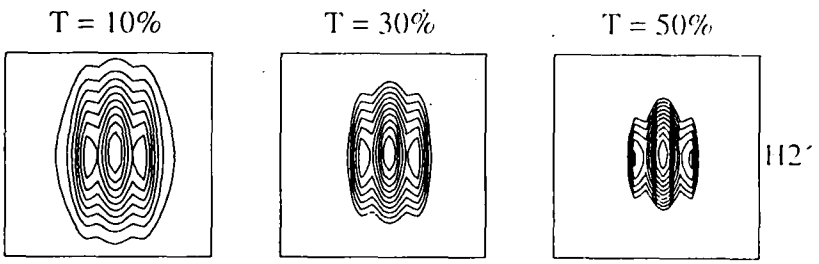


Fig. 2. The strategies of selecting thresholds for peak integration in experimental NOESY spectra: (a) A common threshold is chosen for the entire spectrum by defining a suitable minimum intensity so that the spectral display contains bare minimum noise sufficient to visually discriminate between true peaks and noise. The appearances of three different peaks from three different regions of an experimental spectrum are shown. What appears to be a satisfactory threshold for one peak is not so for another peak in the spectrum. (b) A single threshold has been chosen for a region of the NOESY spectrum. It does not seem to be satisfactory for all the peaks in the given region of the spectrum. (c) A separate threshold is chosen for each peak in the spectrum. As an example of this type of threshold selection, it is shown here how a weak NOESY peak stands out of the noise when we choose higher and higher thresholds. The data are taken from an experimental NOESY spectrum of an oligonucleotide d(GGATTGGCCAATCC) recorded on an AM 500 FTNMR spectrometer; mixing time = 150 ms, temperature = 27°C, 256 t_1 points were collected and each FID had 2048 data points. The data were processed with zero-filling to 1024 points along the t_1 dimension, prior to Fourier transformation. Phase-shifted sine bell ($\pi/8$) window multiplication was employed along both dimensions.

(a) Selection of threshold for peak integration in experimental spectra

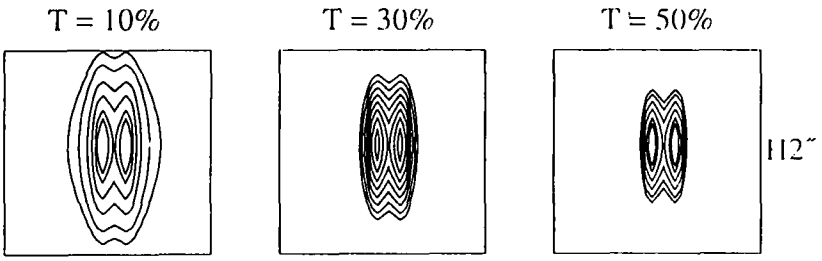
Since the experimental peaks contain noise, it becomes necessary to choose a threshold which is just above the noise level and the volume of the peak above this level will have to be taken to represent the NOE intensity. Figure 2 shows three different strategies for selection of thresholds. In the first case, a minimum intensity is chosen to serve as threshold for the entire spectrum. Figure 2a shows as an illustration three different cross peaks above such a threshold from three different regions of an experimental spectrum. In the second case, the spectrum is divided into different regions to suit the base planes and noise levels and separate minimum intensities are selected for each of the regions to serve as thresholds. Figure 2b shows three different cross peaks above a common threshold in a particular region of the spectrum. In the third case, a separate threshold is selected for each peak in the spectrum. Figure 2c shows the appearance of a particular cross peak above three different thresholds chosen at 10%, 20%, 30% of its peak height.

All the three strategies have been used in the literature and the choice has been dictated by the quality of the spectra. The illustrations in Fig. 2 depict the difficulties that could be encountered in the first two strategies. What appears to be a good threshold – just above noise level – for one peak is not so for another peak and vice versa. In general a number of factors need to be considered for selection of proper thresholds in an experimental spectrum: (i) the noise level varies from one region to another in the spectrum, (ii) the base plane is different in different regions of the spectrum, (iii) the experimental peaks have fine structures and – depending on the multiplicity, line widths, phase characteristics of the components and digital resolutions in the spectra – the overall areas, heights and volumes of the peaks above the noise level are expected to be different, even though they may all have the same total intensity. This is illustrated by simulations of two cross peaks with different multiplicities in Figs. 3 and 4 and also by the extensive data in Table 1. The threshold chosen for Figs. 3 and 4 are 10%, 30% and 50% of respective peak heights and the



A	333.0	168.7	68.1
B	3330.0	1686.9	681.2
C	33300.3	16869.2	6811.6

Fig. 3. Illustrative simulations of an intrasugar H1′-H2′ cross peak above three different threshold levels; the fine structure in the peak corresponds to C2′-endo sugar geometry. See text and the middle column in Fig. 1 for details of the simulation procedure. In all cases, the line widths along both dimensions have been assumed to be 5.0 Hz, and the spectral parameters are: digital resolutions, 1.5 Hz/pt and 3.0 Hz/pt along the H1′ and H2′ axes, respectively. For window multiplication prior to Fourier transformation a cosine function was used. The digital intensities (in arbitrary units) of the peaks for three different analog intensity values, 0.1, 1.0 and 10.0 are shown in A, B and C, respectively, for each of the three thresholds.



A	323.2	170.9	80.2
B	3232.0	1708.6	801.5
C	32319.9	17085.9	8014.8

Fig. 4. Similar data as in Fig. 3, for an H1′-H2″ cross peak from a C3′-endo sugar geometry. It is to be noted by comparison with Fig. 3 that for the same analog intensity, the digitised intensities in the two cross peaks are different. The difference is more pronounced at higher threshold levels.

calculation is repeated for three different total intensity values: 0.1, 1.0 and 10.0. For any single threshold chosen in similar fashion in Figs. 3 and 4, the integrated volumes of the digitised peaks are different and the difference is more pronounced at higher threshold levels. For these reasons, the third strategy of threshold selection (Fig. 2c) seems inevitable and can take care of all the experimental artefacts. The SIMNOE algorithm uses this last strategy for integration of peaks in the experimental spectrum. An array T of peak thresholds is generated once and using these thresholds, the experimental peaks are integrated to generate an array I. The intensities in this array are normalised with respect to a preselected strongest peak in the spectrum.

(b) Data base of fine structures and 'digitised' intensities of cross peaks

The NOE intensity scaling procedure introduced in this paper and described in the next section, requires, to start with, creation of a data base of fine structures of all the usable cross peaks in the NOESY spectrum. This is done by assuming typical multiplicities of the various protons in the molecule and a total intensity of 1.0 for each of the cross peaks. The peaks are created in a digitised manner by first generating the time-domain data and then Fourier-transforming it using the same processing parameters that are used for the experimental spectrum. Then, using the thresholds T_{ij} determined from the experimental spectra, integrated 'digitised' peak volumes (A_{ij}) are obtained for all the peaks. For any given spectrum such a data base has to be created only once, and will then be used repeatedly for scaling of computed intensities (see Fig. 1). It is of course clear that if different spectra with different processing parameters are to be simulated, separate data bases will be needed for each of the spectra.

(c) Conversion of computed 'analog' NOE intensities to 'digitised' intensities

The basic idea is shown in Figs. 3 and 4 and Table 1. For a given peak with in-phase components and Lorentzian line shapes, the digitised intensity is seen to scale identically as the analog intensity, irrespective of the threshold above which the peak is integrated or the multiplet structure of the peak. For example, if the total (or analog) intensity is changed from 1 to 10, the digit-

ised intensity of the peak also gets scaled by the same factor of 10, at each of the three threshold levels chosen (10%, 30% and 50%). Figures 3 and 4 illustrate this point for two specific multiplet patterns. The scaling of digitised intensities for all the commonly used peaks in the NOESY spectrum of a DNA segment is presented in Table 1. This observation is the key to the new analog-to-digital intensity conversion procedure introduced here. Multiplication of the digitised intensity of a peak in the data base (which corresponds to a total intensity of 1.0) by the computed NOE intensity from the relaxation matrix formalism (in array C) should yield the digitised intensity corresponding to that particular total computed intensity (in array D). This scaling operation is highlighted by the thick box in Fig. 1.

TABLE 1
RELATION BETWEEN TOTAL NOE INTENSITY AND INTEGRATED VOLUME OF DIGITISED CROSS PEAKS FOR DIFFERENT THRESHOLDS AND FINE STRUCTURES[a]

Peak	Threshold (%)	Total intensity	Integrated volume ($\times 10^5$)
H1'-H2"	10	1.0	3410.8
		0.1	341.1
	30	1.0	1875.3
		0.1	187.5
H1'-H4'	10	1.0	3200.5
		0.1	320.1
	30	1.0	1690.0
		0.1	169.0
H8-H2'	10	1.0	2787.3
		0.1	278.7
	30	1.0	1322.3
		0.1	132.2
H8-H2"	10	1.0	2789.5
		0.1	278.9
	30	1.0	1351.2
		0.1	135.1
H8-H3'	10	1.0	2587.9
		0.1	258.8
	30	1.0	1199.1
		0.1	119.9
H8-H1'	10	1.0	2669.2
		0.1	266.9
	30	1.0	1250.5
		0.1	125.1

[a] Couplings used (in Hz): J(H1'-H2')=9.0, J(H1'-H2")=6.0; J(H2'-H2")=−14.0, J(H2'-H3')=6.0, J(H2"-H3')=1.0; J(H3'-H4')=1.0, J(H4'-H5')=3.0, J(H4'-H5")=8.0; and J(H5'-H5")=−14.0, J(H3'-P)=5.0.

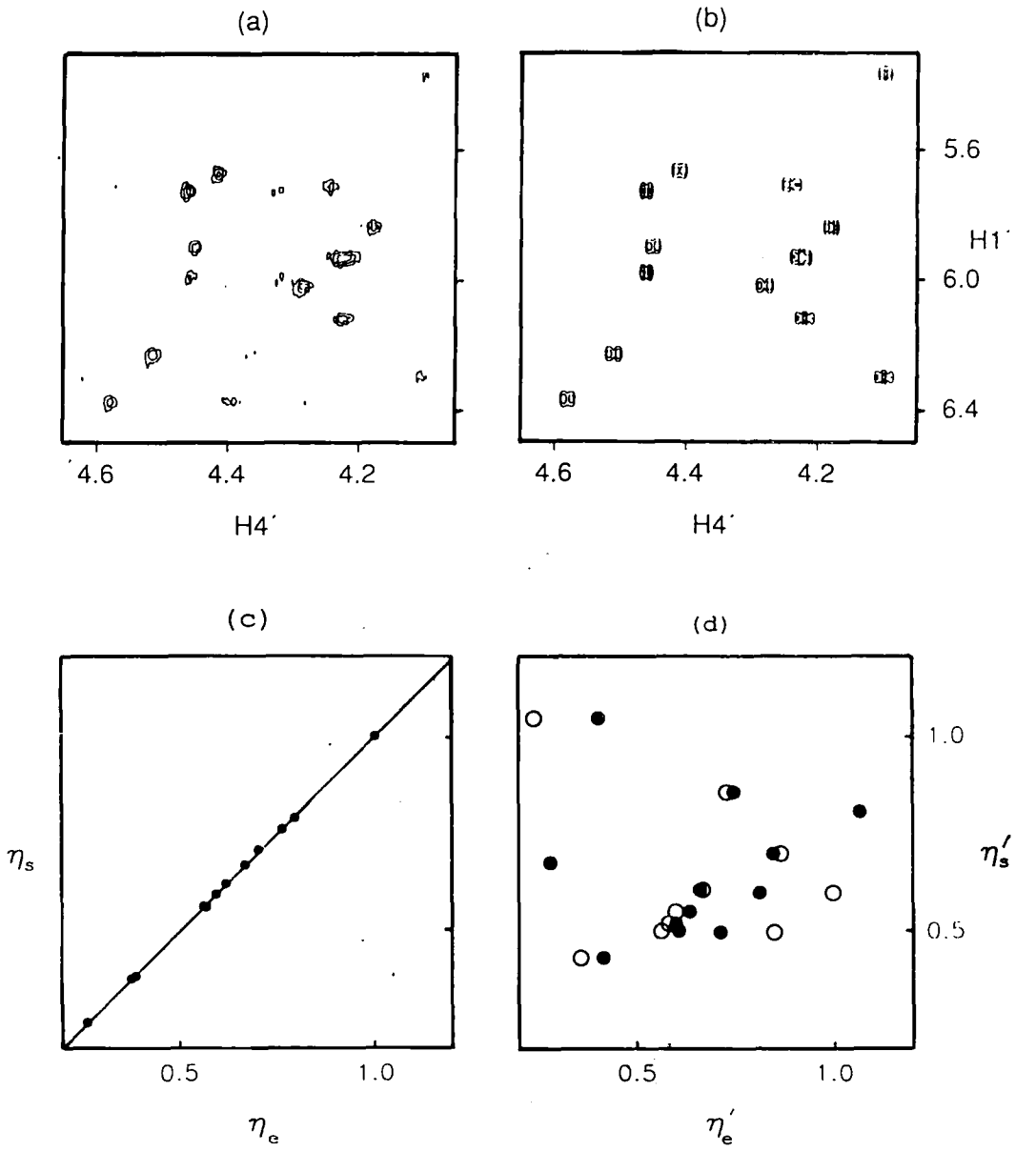


Fig. 5. Comparison of (a) experimental and (b) best fit SIMNOE simulations of the H1'-H4' cross peaks in the NOESY spectrum of the same oligonucleotide as used in Fig. 2. Quantitative comparison of the intensities as per the methodology described in the text is shown in (c). The threshold is chosen at 20% of peak height for each peak. Here η_e and η_s represent experimental and simulated intensities, respectively. The fit is seen to be extremely good. In (d) the fit between calculated and experimental intensities as per the protocol used by others (Baleja et al., 1990a; Gochin et al., 1990) is shown for comparison. Two sets of data are shown by open and filled circles which correspond to (i) a single uniform threshold at 20% peak height of a particular peak in the region (○) and (ii) a separate threshold is chosen at 20% of peak height for each individual peak (●). The digitised experimental intensities above these thresholds were then multiplied by a factor, q=(sum of calculated intensities)/(sum of experimental intensities). These are represented by η'_e in the figure. η'_s are the calculated analog intensities. The quality of fit is seen to be poor in both cases although the structure of the molecule is the same.

Mathematically,

$$D_{ij} = C_{ij} * A_{ij}$$

These elements are normalised using the digitised simulated intensity of the particular peak selected for normalising experimental intensities. The elements of the D array can now be compared with the elements of the I array to check the correctness of the structure. It is clear that such a procedure results in substantial saving of computational time and enhances the speed of structure optimisation.

For the present version of SIMNOE, structure alteration is performed interactively with coupling to a DNA modelling program MODEST (Ajay Kumar, R., to be published). Efforts are being made to automatise the entire procedure. Since relaxation matrix calculations take a fair amount of computational time, a completely automatic search of the conformational space using any of the search algorithms is still computationally prohibitive.

AN EXPERIMENTAL DEMONSTRATION

Figure 5 shows a comparison of an experimental (a) and the corresponding SIMNOE-simulated (b) spectral region of a 150-ms NOESY spectrum of an oligonucleotide d(GGATTGGCCAATCC) at 27°C; the quantitative comparison of the intensities in the two spectra is shown in (c). The simulations have been performed assuming a single correlation time of 2 ns and a single leakage rate of 0.5 sec^{-1} for all the protons in the relaxation matrix calculations. The fit between experimental and simulated intensities is seen to be extremely good. The H1'-H4' cross peaks shown in the spectra determine the sugar geometries of the various nucleotide units in the DNA segment (see reviews Wüthrich, 1986; Hosur et al., 1988; van de Ven and Hilbers, 1988). While the details of the structure derived by the above simulations will be published separately, it suffices to note here that the sugar geometries are constrained to the S domain in the oligonucleotide. In Fig. 5d the fit of digitised simulated and experimental intensities for the same structure of the molecule but as per the scaling procedure used by others (Baleja et al., 1990a; Gochin et al., 1990) is shown for comparison. The fit of simulated and experimental intensities is poor in this case.

ACKNOWLEDGEMENTS

The facilities provided by the 500-MHz FT-NMR National Facility, supported by the Department of Science and Technology, Government of India are gratefully acknowledged.

REFERENCES

Anil Kumar, Ernst, R.R. and Wüthrich, K. (1980) *Biochem. Biophys. Res. Commun.*, **95**, 1–6.
Anil Kumar, Wagner, G., Ernst, R.R. and Wüthrich, K. (1981) *J. Am Chem. Soc.*, **103**, 3654–3656.
Baleja, J.D., Pon, R.T. and Sykes, B.D. (1990a) *Biochemistry*, **29**, 4828–4839.
Baleja, J.D., Germann, M.W., van de Sande, J.H. and Sykes, B.D. (1990b) *J. Mol. Biol.*, **215**, 411–428.
Banks, K.M., Hare, D.R. and Reid, B.R. (1989) *Biochemistry*, **26**, 6996–7002.
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.
Bonvin, A.M.J.J., Boelens, R. and Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 305–309.

Borgias, B.A. and James, T.L. (1990) *J. Magn. Reson.*, **87**, 475–487.
Borgias, B.A., Gochin, M., Kerwood, D.J. and James, T.L. (1990) *Prog. NMR Spectrosc.*, **22**, 83–100
Clore, G.M. and Gronenborn, A.M. (1989) *Crit. Rev. Biochem.*, **24**, 479–564.
Clore, G.M. and Gronenborn, A.M. (1991) *Prog. NMR Spectrosc.*, **23**, 43–92.
Gochin, M. and James, T.L. (1990) *Biochemistry*, **29**, 11172–11180.
Gochin, M., Zon, G. and James, T.L. (1990) *Biochemistry*, **29**, 11161–11171.
Hosur, R.V., Govil, G. and Miles, H.T. (1988) *Magn. Reson. Chem.*, **26**, 927–944.
Hosur, R.V., Chary, K.V.R., Saran, A., Govil, G. and Miles, H.T. (1990) *Biopolymers*, **29**, 953–959.
Keepers, J.W. and James, T.L. (1984) *J. Magn. Reson.*, **57**, 404–426.
Landy, S.B. and Rao, B.D.N. (1989) *J. Magn. Reson.*, **83**, 29–43.
Macura, S. and Ernst, R.R. (1980) *Mol. Phys.*, **41**, 95–117.
Majumdar, A. and Hosur, R.V. (1990) *J. Magn. Reson.*, **88**, 284–304.
Majumdar, A. and Hosur, R.V. (1991) *Prog. NMR Spectrosc.*, **24**, 109–158.
Mertz, J.D., Guntert, P., Wüthrich, K. and Braun, W. (1991) *J. Biomol NMR*, **1**, 257–269.
Nerdal, W., Hare, D.R. and Reid, B.R. (1989) *Biochemistry*, **28**, 10008–10021.
Van de Ven, F. and Hilbers, C.W. (1988) *Eur. J. Biochem.* **178**, 1–38.
Wagnér, G. (1990) *Prog. NMR Spectrosc.*, **22**, 101–139.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.
Wüthrich, K. (1989a) *Science*, **243**, 45–50.
Wüthrich, K. (1989b) *Acc. Chem. Res.*, **22**, 36–44.
Zhou, N., Bianucci, A.M., Pattabhiraman, N. and James, T.L. (1987) *Biochemistry*, **26**, 7905–7913.