Journal of Biomolecular NMR, 6 (1995) 294–305
ESCOM

J-Bio NMR 303

# Determination of the NMR solution structure of a specific DNA complex of the Myb DNA-binding domain

Souichi Morikawa[a], Kazuhiro Ogata[b,c], Ai Sekikawa[b], Akinori Sarai[c], Shunsuke Ishii[c], Yoshifumi Nishimura[b] and Haruki Nakamura[a,*]

[a]*Protein Engineering Research Institute, Suita, Osaka 565, Japan*
[b]*Graduate School of Integrated Science, Yokohama City University, Kanazawa-ku, Yokohama 236, Japan*
[c]*Tsukuba Life Science Center, The Institute of Physical and Chemical Research, Tsukuba, Ibaraki 305, Japan*

Received 31 May 1995
Accepted 1 August 1995

*Keywords:* Distance geometry; Simulated annealing; DNA binding; Myb

## Summary

The solution structure of a specific DNA complex of the minimum DNA-binding domain of the mouse c-Myb protein was determined by distance geometry calculations using a set of 1732 nuclear Overhauser enhancement (NOE) distance restraints. In order to determine the complex structure independent of the initial guess, we have developed two different procedures for the docking calculation using simulated annealing in four-dimensional space (4D-SA). One is a multiple-step procedure, where the protein and the DNA were first constructed independently by 4D-SA using only the individual intramolecular NOE distance restraints. Here, the initial structure of the protein was a random coil and that of the DNA was a typical B-form duplex. Then, as the starting structure for the next docking procedure, the converged protein and DNA structures were placed in random molecular orientations, separated by 50 Å. The two molecules were docked by 4D-SA utilizing all the restraints, including the additional 66 intermolecular distance restraints. The second procedure comprised a single step, in which a random-coil protein and a typical B-form DNA duplex were first placed 70 Å from each other. Then, using all the intramolecular and intermolecular NOE distance restraints, the complex structure was constructed by 4D-SA. Both procedures yielded the converged complex structures with similar quality and structural divergence, but the multiple-step procedure has much better convergence power than the single-step procedure. A model study of the two procedures was performed to confirm the structural quality, depending upon the number of intermolecular distance restraints, using the X-ray structure of the *engrailed* homeodomain–DNA complex.

## Introduction

The c-*myb* proto-oncogene product (c-Myb) is a transcriptional activator that binds to specific DNA fragments containing the consensus sequence AACNG. c-Myb is mainly expressed in the immature cells of various hematopoietic lineages and plays a critical role in the proliferation and differentiation of hematopoietic progenitor cells (Luscher and Eisenman, 1990; Graf, 1992). The DNA-binding domain consists of three imperfect tandem repeats of 51–52 amino acids. It has been found that the second (R2) and third (R3) repeats are sufficient for the specific recognition of the consensus DNA sequence, and that the first repeat (R1) contributes slightly to the tight binding of DNA in a nonspecific manner. Therefore, the minimum DNA-binding domain is the R2R3 fragment (Myb-R2R3).

In order to understand the mechanism of the Myb–DNA interaction at the atomic level, we have determined the solution structure of each repeat, i.e., R1, R2, and R3 (Ogata et al., 1992,1995), and of the DNA-complexed structure (Ogata et al., 1994) by NMR/distance geometry

*To whom correspondence should be addressed.

*Abbreviations:* rmsd, root-mean-square deviation; NOE, nuclear Overhauser enhancement; 4D-SA, simulated annealing in four-dimensional space; Myb-R2R3, repeats 2 and 3 of the DNA-binding domain of the c-Myb protein; DNA16, Myb-specific binding DNA duplex with 16 base pairs; 1HDD-C, residues 3 to 59 of the C-chain of the *engrailed* homeodomain–DNA complex; DNA11, DNA duplex with base pairs 9 to 19 of the *engrailed* homeodomain–DNA complex.

calculations. For the latter study, the Myb-binding 16-mer DNA duplex (designated here as DNA16) was used. From this series of studies, we have found that the three repeats have a very similar overall architecture; each contains a helix–turn–helix variation motif and forms a hydrophobic core of tryptophan residues (Ogata et al., 1992,1995). Each repeat is independent in the DNA-free form. However, in the DNA-complexed form, R2 and R3 become closely packed in the major groove, so that the two recognition helices cooperatively bind to the specific base sequence (Ogata et al., 1994).

The details of each repeat structure, the interaction scheme between Myb-R2R3 and DNA16, and their biological importance have been described previously (Ogata et al., 1992,1994,1995). However, we have not yet reported a precise method to construct the DNA-complexed structure by distance geometry calculations. In this paper, we examine the quality of the complex structures by comparing the simple single-step procedure with the more effective multiple-step procedure; the latter was used in our previous paper (Ogata et al., 1994).

Many protein solution structures have been determined from geometrical information obtained by multidimensional NMR experiments, as well as by X-ray crystallographic studies. In order to construct three-dimensional structural models from the distance and torsion angle restraints resulting from NMR experiments, several algorithms have been proposed, and associated programs have been developed (James, 1994). We have developed our own method (the EMBOSS program), which uses a simulated annealing optimization in four-dimensional space (4D-SA). The reliability and quality of the distance geometry structures generated by the 4D-SA protocol have been evaluated, and it could be concluded that the method yields an extremely large radius of convergence with excellent, and almost exhaustive, sampling properties (Nakai et al., 1993).

Recently, several DNA complex structures of DNA-binding domains have been determined by NMR/distance geometry calculations. For example, the structures of the *lac* repressor headpiece–operator complex (Boelens et al., 1988; Chuprina et al., 1993), the DNA-bound homeodomain (Billeter et al., 1993; Qian et al., 1993), the DNA complex of the GATA-1 DNA-binding domain (Omichinski et al., 1993), and the *trp* repressor–operator complex (Zhang et al., 1994) have been published. In all of these structure determinations, except for the specific DNA complex of the DNA-binding domain of GATA-1, the proteins and specific DNA structures were essentially docked either by the rigid-body docking procedure (Boelens et al., 1988), by the molecular dynamics procedure in three-dimensional space (De Vlieg et al., 1989), or by graphic modeling (Zhang et al., 1994), based upon previously constructed models of the individual proteins and DNA molecules. Billeter et al. (1993) systematically changed the initial arrangements of the homeodomain and the DNA, and showed that the final structures did not depend upon the initial guess. In our previous paper (Ogata et al., 1994), we implemented the last procedure in our 4D-SA protocol with more random initial deposition to determine the DNA-complexed structure of Myb-R2R3.

However, it has not been determined whether such a docking procedure can sample all possible complex structures completely independent of the initial estimate. In the structure determination of the DNA complex of the DNA-binding domain of GATA-1 by Omichinski et al. (1993), a procedure was used involving the hybrid distance geometry (embedding) -SA method. It has been noticed that the original embedding algorithm searches a rather limited and biased conformational space if randomized metrization is not incorporated (Havel, 1990; Kuszewski et al., 1992).

In this report, we investigate whether our previous docking procedure (designated here as the multiple-step procedure) can construct the complex structure after searching a conformational space as wide as the simple 4D-SA protocol, without any docking of the two constructed structures (designated as the single-step procedure). These procedures are used to examine the structural quality of the DNA complex of Myb-R2R3 and the homeodomain–DNA complex model.

## Methods

*NMR spectroscopy of the Myb–DNA complex*

Various NMR spectra of the Myb-R2R3–DNA16 complex in salt-free solutions (1.8–2.5 mM at pH 6.8) containing 20 mM DTT-$d_6$ and 1 mM $NaN_3$ were measured using a Bruker AMX-500 spectrometer. $^{13}C$- or $^{15}N$-labeled proteins were prepared for heteronuclear multidimensional NMR experiments, in order to obtain unambiguous assignments of chemical shifts and intramolecular

TABLE 1
NUMBER OF SIMULATED INTRAMOLECULAR NOE RESTRAINTS FOR THE RECONSTRUCTION OF THE HOMEODOMAIN–DNA COMPLEX

| Simulated NOE restraints | Number |
|---|---|
| **1HDD-C** | |
| Intraresidue[a] | 0 |
| Interresidue sequential ($|i-j|=1$) | 217 |
| Interresidue short range ($1<|i-j|\leq 5$) | 256 |
| Interresidue long range ($5<|i-j|$) | 141 |
| Total | 614 |
| **DNA11** | |
| Intraresidue | 364 |
| Sequential intrastrand | 197 |
| Total | 561 |

[a] None of the intraresidual proton pairs of 1HDD-C were selected for the current simulated NOE data sets.

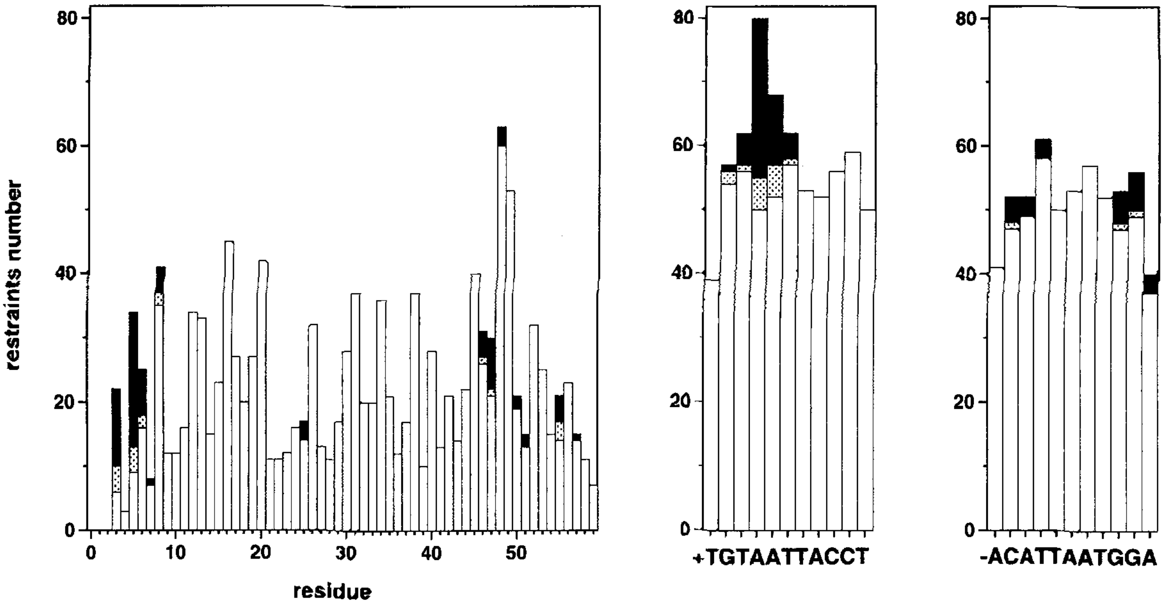

Fig. 1. The distribution of the simulated distance restraints of the complexed structure of 1HDD-C and DNA11. The white bars represent intramolecular distance restraints. Light hatched bars, dark hatched bars and black bars represent intermolecular distance restraints within 3.0, 3.5 and 4.5 Å, respectively.

and intermolecular NOE signals. Details of the NMR experiments have been given elsewhere (Ogata et al., 1994,1995).

*Distance restraints of the Myb–DNA complex*

Cross-peak intensities of the NOEs between the protein backbone protons were classified into four ranges, i.e., 1.9 to 3.0, 1.9 to 4.0, 1.9 to 5.0, and 1.9 to 6.0 Å, corresponding to strong, medium, weak and very weak NOEs, respectively. The intensities between the protons within the protein and between the protein and DNA protons were classified into four ranges, i.e., 1.9 to 4.0, 1.9 to 4.5, 1.9 to 5.5, and 1.9 to 6.5 Å, corresponding to strong, medium, weak, and very weak NOEs, respectively. The

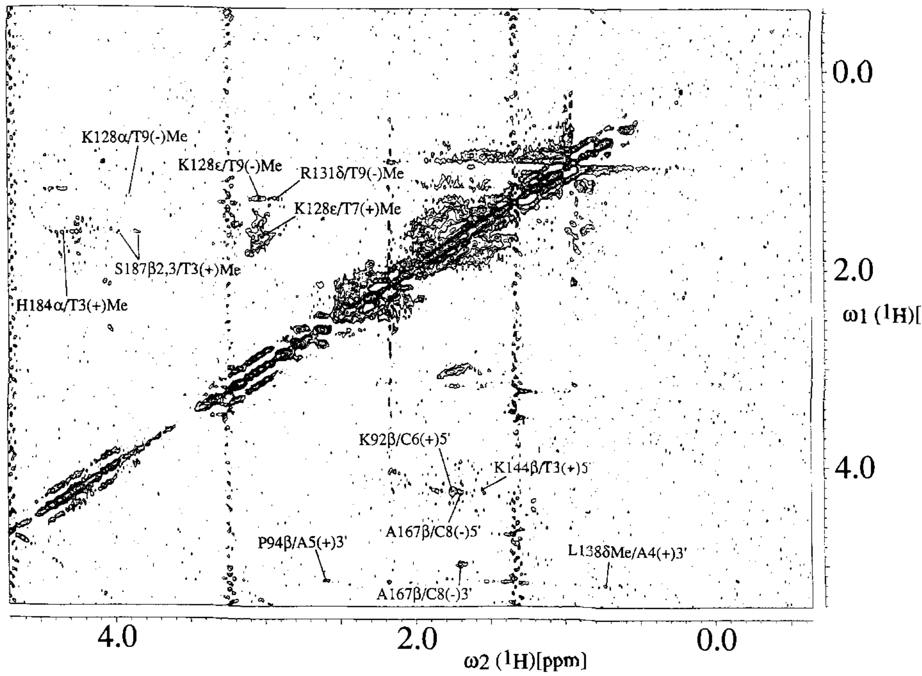

Fig. 2. $^{13}C(\omega_1)$-filtered-$^{13}C(\omega_2)$-selected 2D [$^1H$,$^1H$]-NOESY spectrum of the complex formed by uniformly $^{13}C$-labeled Myb-R2R3 and nonlabeled DNA16, with a mixing time of 100 ms. The NMR sample contained 2.5 mM protein, 20 mM DTT-$d_6$ and 1 mM $NaN_3$ in $D_2O$, pH 6.8. The spectrum was recorded at 500 MHz on a Bruker AMX-500 spectrometer. The temperature during data acquisition was 310 K. Several intermolecular NOE cross peaks are labeled with their assignments. There are also some intramolecular NOE cross peaks and diagonal peaks, due to imperfect isotope filtering.

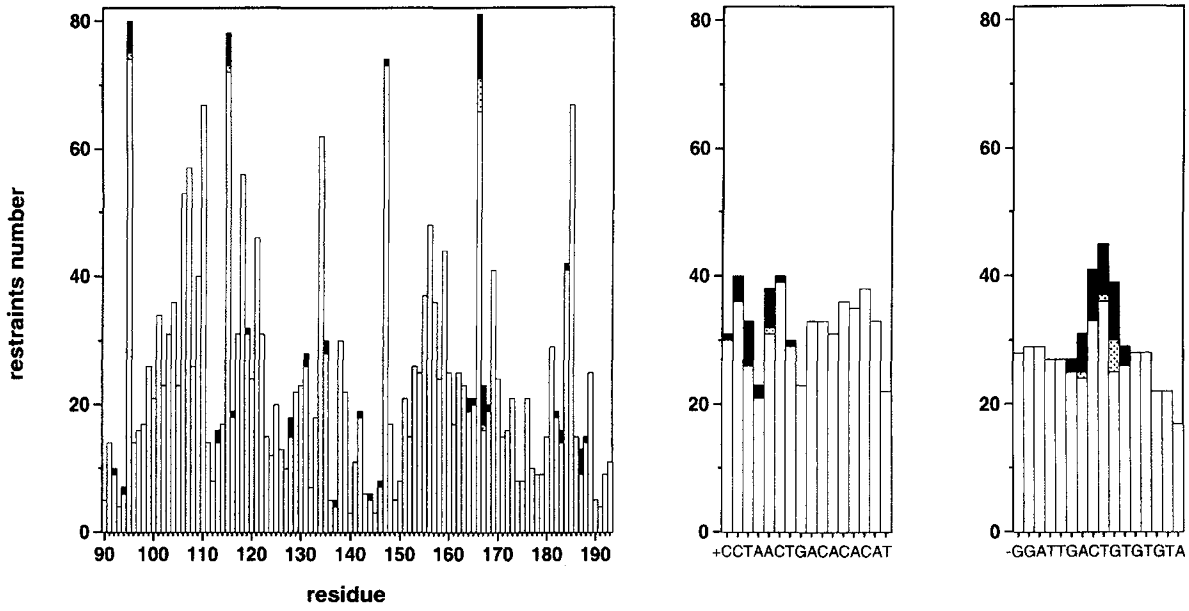

Fig. 3. The distribution of the observed distance restraints in the complexed structure of Myb-R2R3 and DNA16. The base sequences of the + and – strands of DNA16 are also shown. White bars represent intramolecular distance restraints. Light hatched bars represent intermolecular restraints with 4.0 and 4.5 Å upper bounds, dark hatched bars indicate restraints with 5.5 Å upper bounds, and black bars represent restraints with upper bounds larger than 5.5 Å.

intensities between the DNA protons were classified into six ranges, i.e., 1.9 to 2.5, 2.3 to 3.0, 2.3 to 3.5, 2.3 to 4.0, 2.5 to 5.0, and 3.0 to 6.0 Å, corresponding to strong, medium-strong, medium, medium-weak, weak, and very weak NOEs, respectively. A total of 66 intermolecular and 1666 intramolecular NOE restraints were observed. The torsion angles of the DNA backbones were restricted, so that the right-handed phosphate backbones are maintained without the local mirror image conformation (Gronenborn and Clore, 1989). The α, β, γ, ε and ζ torsion angles were restricted to broad ranges of –85° ± 50°, 180° ± 50°, 70° ± 50°, 130° ± 50° and –60° ± 40°, respectively.

### *Structure calculations of the Myb–DNA complex*

From the distance and torsion angle restraints, the three-dimensional complexed structure of Myb-R2R3 and DNA16 was constructed by two different protocols, the single-step and multiple-step procedures, as follows.

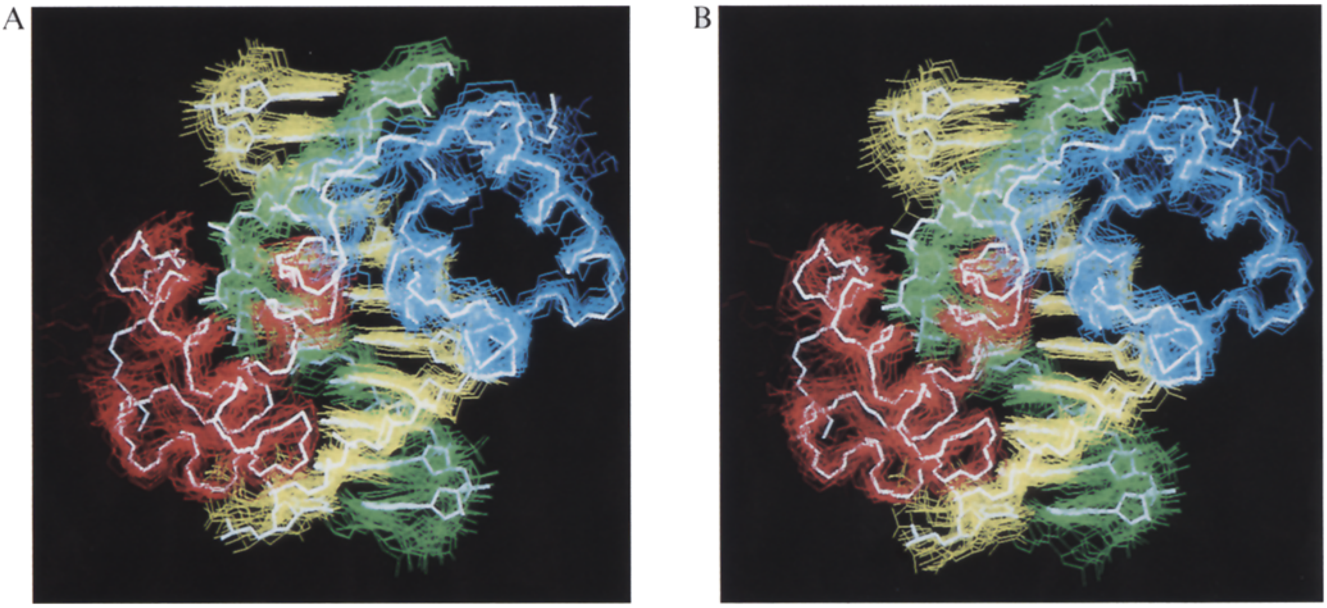

Fig. 4. The superimposed complexed structures of Myb-R2R3 and DNA16 generated by (A) the single-step procedure and (B) the multiple-step procedure. The backbone atoms in the R2 region of Myb-R2R3 are shown in red; those in the R3 region are shown in blue. All heavy atoms in the + strand of DNA16 are shown in green, and those in the – strand are shown in yellow. The refined average structures are shown in white.

TABLE 2

NUMBER OF SIMULATED INTERMOLECULAR NOE RESTRAINTS AND RMSDs OF THE HOMEODOMAIN-DNA COMPLEXES RECONSTRUCTED FROM THE SIMULATED DATA SETS

| | Single-step procedure[a] | | | | Multiple-step procedure[b] | | |
|---|---|---|---|---|---|---|---|
| | S1[c] | S2 | S3 | S4 | M1 | M2 | M3 |
| No. of intermolecular NOEs[d] | 89 | 39 | 17 | 9 | 89 | 39 | 17 |
| $N^{4D\text{-}SA}$ [e] / $N^{opt}$ [f] | 59/37 | 55/37 | 58/44 | 45/31 | 71/45 | 68/57 | 82/59 |
| **Rmsd values vs. X-ray structure**[g] | | | | | | | |
| Protein–DNA complex[h] | | | | | | | |
| backbone atoms[i] | 1.06±0.21 | 1.12±0.23 | 1.71±0.53 | 1.66±0.45 | 1.05±0.19 | 1.13±0.19 | 1.81±0.59 |
| heavy atoms | 1.40±0.25 | 1.48±0.27 | 1.93±0.47 | 1.90±0.43 | 1.41±0.24 | 1.48±0.22 | 2.00±0.52 |
| Protein[j] | | | | | | | |
| backbone atoms | 0.37±0.10 | 0.41±0.13 | 0.41±0.11 | 0.42±0.12 | 0.38±0.10 | 0.41±0.11 | 0.40±0.10 |
| heavy atoms | 1.35±0.25 | 1.42±0.27 | 1.41±0.25 | 1.40±0.27 | 1.38±0.24 | 1.42±0.22 | 1.38±0.21 |
| DNA double helix[k] | | | | | | | |
| phosphate backbone atoms[l] | 1.27±0.26 | 1.23±0.25 | 1.61±0.44 | 1.51±0.39 | 1.25±0.23 | 1.24±0.21 | 1.71±0.55 |
| heavy atoms | 1.27±0.24 | 1.25±0.23 | 1.53±0.36 | 1.46±0.34 | 1.25±0.22 | 1.25±0.20 | 1.61±0.45 |
| Protein–DNA interface[m] | | | | | | | |
| backbone atoms[i] | 0.57±0.13 | 0.74±0.18 | 1.07±0.37 | 1.02±0.35 | 0.57±0.13 | 0.71±0.16 | 1.14±0.40 |
| heavy atoms | 1.00±0.19 | 1.17±0.23 | 1.36±0.32 | 1.31±0.32 | 1.01±0.19 | 1.13±0.19 | 1.40±0.33 |
| **Rmsd values among reconstructed structures**[n] | | | | | | | |
| Protein–DNA complex[h] | | | | | | | |
| backbone atoms[i] | 0.86±0.20 | 1.05±0.27 | 1.40±0.45 | 1.54±0.48 | 0.83±0.21 | 1.01±0.24 | 1.51±0.57 |
| heavy atoms | 1.30±0.13 | 1.43±0.20 | 1.66±0.33 | 1.77±0.38 | 1.31±0.15 | 1.41±0.17 | 1.76±0.44 |
| Protein[j] | | | | | | | |
| backbone atoms | 0.44±0.09 | 0.49±0.12 | 0.50±0.12 | 0.48±0.11 | 0.46±0.10 | 0.48±0.11 | 0.48±0.10 |
| heavy atoms | 1.40±0.12 | 1.46±0.13 | 1.45±0.13 | 1.45±0.13 | 1.44±0.13 | 1.45±0.13 | 1.44±0.12 |
| DNA double helix[k] | | | | | | | |
| phosphate backbone atoms[l] | 0.82±0.24 | 0.91±0.27 | 1.22±0.46 | 1.35±0.45 | 0.78±0.23 | 0.88±0.25 | 1.34±0.63 |
| heavy atoms | 0.91±0.18 | 0.96±0.20 | 1.19±0.31 | 1.28±0.35 | 0.87±0.16 | 0.94±0.18 | 1.28±0.49 |
| Protein–DNA interface[m] | | | | | | | |
| backbone atoms[i] | 0.57±0.16 | 0.77±0.25 | 0.97±0.36 | 1.12±0.40 | 0.55±0.16 | 0.68±0.19 | 1.02±0.41 |
| heavy atoms | 0.98±0.15 | 1.13±0.20 | 1.23±0.27 | 1.33±0.31 | 0.98±0.14 | 1.07±0.16 | 1.28±0.31 |

[a] In the single-step procedure, the initial structures were different random coil 1HDD-C peptides and a typical B-type DNA helix, and each center of mass was separated by 70 Å. The complex structures were constructed simultaneously by the 4D-SA protocol. Here, the final $N^{opt}$ structures selected after energy minimization are compared.

[b] In the multiple-step procedure, the initial structures of 1HDD-C and DNA11 in the first step were different random coils and a typical B-type DNA helix, respectively. Both peptide and DNA structures were individually constructed by the 4D-SA protocol using the intramolecular NOEs. In the second step, the converged structures of 1HDD-C and DNA11 from the first step were separated by 50 Å as the initial complex structure. Adding the intermolecular NOEs, the complex structures were constructed. Here, the final $N^{opt}$ structures after energy minimization are compared.

[c] The simulated data sets have the same intramolecular NOE restraints, but different intermolecular restraints between 1HDD-C and DNA11.

[d] The number of intermolecular NOEs between 1HDD-C and DNA11. For the S1 and M1 data sets, the intermolecular proton pairs with distances shorter than 4.5 Å were selected. For S2 and M2, the intermolecular proton pairs with distances shorter than 3.5 Å were selected. For S3 and M3, the intermolecular proton pairs with distances shorter than 3.0 Å were selected. For the S4 data set, the intermolecular proton pairs with distances shorter than 2.85 Å were selected.

[e] $N^{4D\text{-}SA}$ represents the number of structures, after the 4D-SA computation, with no distance violations greater than 0.5 Å, with deviations of the bond lengths and angles from ideal less than 0.015 Å and 3°, respectively, and where the maximum value of the four-dimensional coordinate is less than 0.001 Å. All calculations started from 100 initial structures.

[f] $N^{opt}$ represents the number of structures, after further energy minimization, with no distance violation larger than 0.4 Å and no dihedral angle violation larger than 4°, and where the total energy is smaller than −680 kcal/mol.

[g] The rmsd comparison of backbone atoms and all heavy atoms of the constructed structures with the regularized X-ray crystal structure of the 1HDD-C–DNA11 complex. The means ± standard deviations of the $N^{opt}$ rmsd values are listed.

[h] The rmsd calculations were carried out with respect to residues Ser$^{9}$–Lys$^{57}$ of 1HDD-C and the 11 base pairs of DNA11.

[i] Only the backbone atoms (N, $C^{\alpha}$, and C') of 1HDD-C and the phosphate backbone atoms (P, O5', C5', C4', C3', O3') of DNA11 were considered.

[j] The rmsd calculations were carried out with respect to residues Ser$^{9}$–Lys$^{57}$ of 1HDD-C.

[k] The rmsd calculations were carried out with respect to the 11 base pairs of DNA11.

[l] Only the phosphate backbone atoms (P, O5', C5', C4', C3', O3') of DNA11 were considered.

[m] The rmsd calculations were carried out with respect to the interface between 1HDD-C (the residues in the third helix) and DNA11, i.e., residues Gln$^{44}$–Lys$^{55}$ of 1HDD-C and base pairs 4–8 of DNA11.

[n] The rmsd values of the backbone atoms and all heavy atoms among the $N^{opt}$ constructed structures. The means ± standard deviations of the $N^{opt} \times (N^{opt} - 1) / 2$ rmsd values are listed.

*Single-step procedure* Using both the intramolecular and intermolecular restraints, 500 structures of the complex of Myb-R2R3 and DNA16 were constructed simultaneously, following the four-dimensional simulated annealing (4D-SA) protocol with the program EMBOSS, as described previously (Nakai et al., 1993). The initial structures of Myb-R2R3 were 500 different random coils and that of DNA16 was a typical B-form double helix (Arnot and Hukins, 1972). The centers of mass of the two molecules were separated by 70 Å. Out of the 500 calculated complex structures, 59 structures were selected that had no individual distance violation larger than 0.5 Å.

The selected structures were further energy minimized, with the experimental restraints, by performing 5000 conjugate gradient steps with the program PRESTO (Morikami et al., 1992), using the AMBER all-atom force field (Weiner et al., 1986). The electrostatic interactions were included with a dielectric constant $2r_{ij}$ for the nonbonded atoms i and j, where $r_{ij}$ is the distance between the atoms i and j. After energy minimization, 22 structures were selected that had no individual distance violation larger than 0.4 Å, no dihedral angle violation larger than 4.0°, and a total energy smaller than −1000 kcal/mol.

*Multiple-step procedure* First, using the individual intramolecular restraints, 20 structures of Myb-R2R3 and 10 of DNA16 were independently constructed following the 4D-SA protocol. The initial structures of Myb-R2R3 were 20 different random coils and that of DNA16 was a typical B-form double helix. Half of the resulting conformations without violations were further energy minimized, in the same manner as mentioned above. The four best individual structures were taken as the initial conformations for the next docking procedure. Second, Myb-R2R3 and DNA16 were docked by the 4D-SA protocol with all the restraints, including additional intermolecular distance restraints, by the program EMBOSS. As the initial structures of the complex, the Myb-R2R3 and DNA16 structures constructed in the preceding step were positioned with random molecular orientations around each center of mass, separated 50 Å from each other. Out of the 100 calculated complex structures, 60 structures were selected with no individual distance violation larger than 0.5 Å.

Finally, energy minimization was carried out for the converged structures, including all the restraints. Out of the 60 structures, 25 were selected that satisfied the same geometrical and energetic criteria as those for the final stage in the single-step procedure, i.e., no individual distance violation larger than 0.4 Å, no dihedral angle violation larger than 4.0°, and a total energy smaller than −1000 kcal/mol.

*Model calculation for the homeodomain–DNA complex*

In order to examine the reliability of the current protocols, we simulated the distance restraints from the X-ray crystal structure of the *engrailed* homeodomain–DNA complex, including residues 3–59 of the C-chain and base pairs 9–19 of entry 1HDD (Kissinger et al., 1990) of the Protein Data Bank. Similar model studies were previously performed by Billeter et al. (1993), evaluating their own protocol.

After attaching hydrogen atoms to the heavy atoms of the crystal structure, the positions of the hydrogen atoms were optimized, in order to fix all the heavy atoms, by 3000 steps of energy minimization with the program PRESTO, using the AMBER all-atom force field. For the

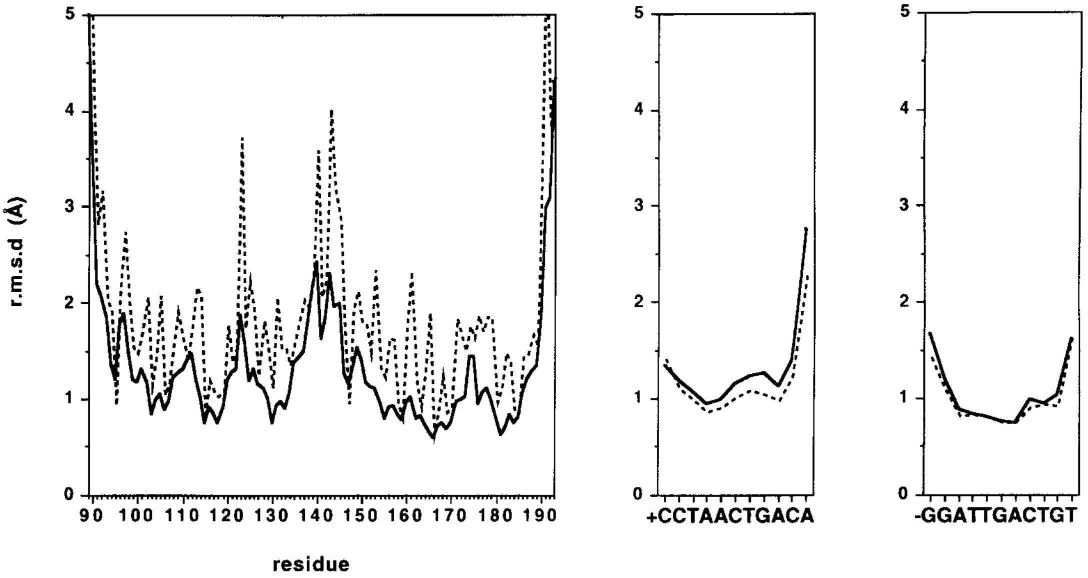

Fig. 5. The rmsd values per residue from the average coordinates of 22 final complexed structures of Myb-R2R3–DNA16, produced by the single-step procedure. The solid lines indicate rmsd values of the backbone atoms in Myb-R2R3 and the phosphate backbone atoms in DNA16. The dashed lines represent the rmsd values of all heavy atoms. In the DNA duplex, the well-defined base conformation decreases the rmsd values, due to the additional hydrogen bond restraints.

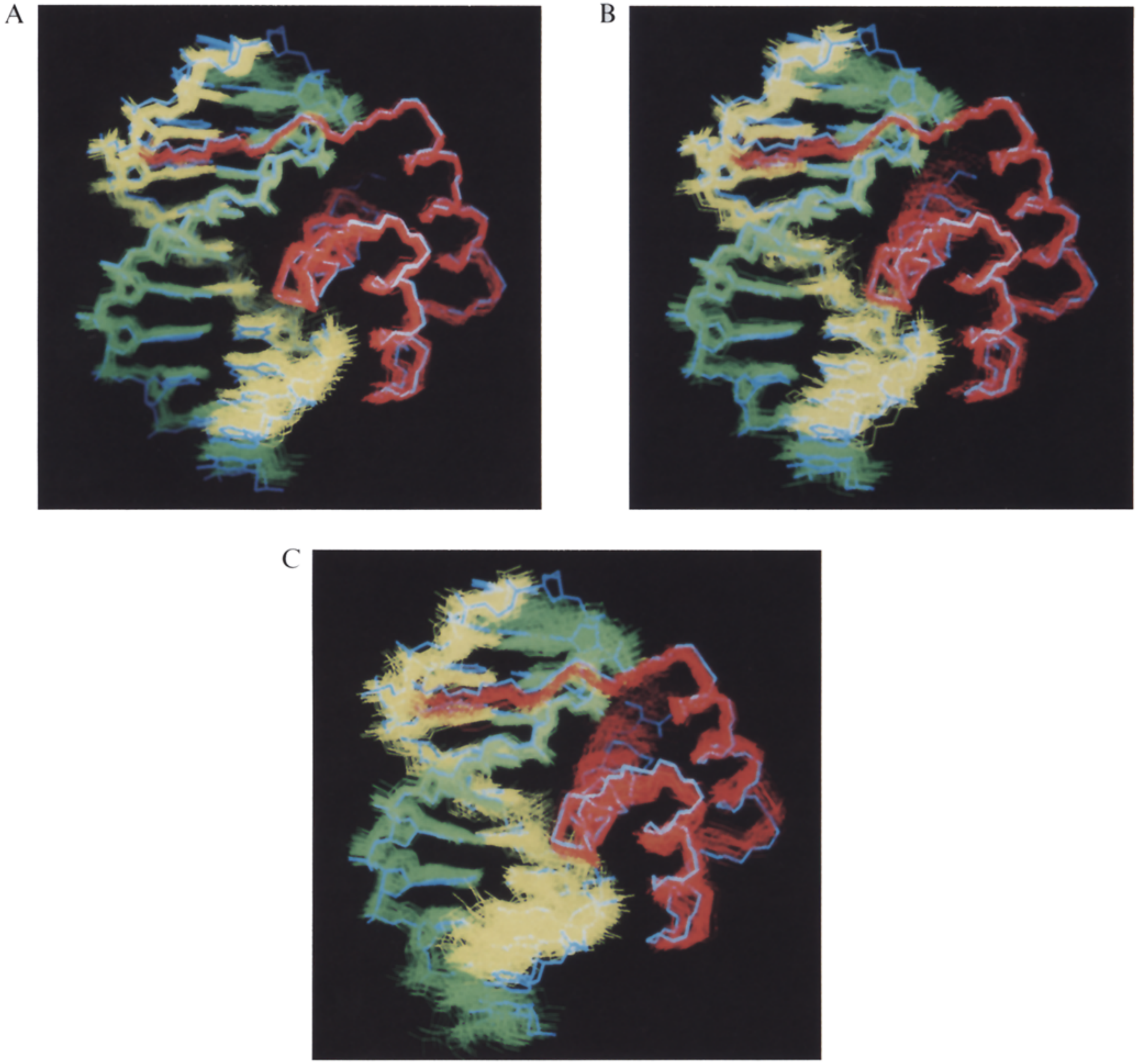

Fig. 6. The superimposed complexed structures of the homeodomain 1HDD-C and DNA11 generated by the single-step procedure, using different data sets of the simulated distance restraints; (A) S1, (B) S2, and (C) S3 (see Table 2). The backbone atoms in 1HDD-C are shown in red. All heavy atoms in the + strand of the DNA are shown in green, and those in the – strand are shown in yellow. The original X-ray structure is shown in blue.

protein and DNA molecules, intramolecular NOE restraints were individually simulated for the proton pairs (i,j) with distances $d_{ij}$ less than 4.5 Å. Here, the aliphatic and aromatic protons, the backbone amide protons, and the $H^{\varepsilon}$ protons in arginine side chains were taken into consideration, whereas other labile protons were neglected. For the $CH_2$ and $CH_3$ protons, and the protons at symmetrical positions in aromatic groups, the averages of $(1/d_{ij}^{6})$ were calculated. For proton pairs containing at least one protein proton, the upper bounds of their distances shorter than 3.0, 4.0, and 4.5 Å were set to 3.0, 4.0, and 5.0 Å, respectively. All lower bounds were set to the sum of the van der Waals radii of the corresponding atoms. For DNA proton pairs, the distance restraints were classified into five ranges, i.e., 1.9 to 2.5, 2.3 to 3.0, 2.3 to 3.5, 2.3 to 4.0, and 2.5 to 5.0 Å, corresponding to distances shorter than 2.5, 3.0, 3.5, 4.0, and 4.5 Å, respectively. The precise numbers of these intramolecular distance restraints are summarized in Table 1. In addition to the distance restraints for the hydrogen bonds between base pairs in the double-stranded DNA, the torsion angles of the DNA backbones were restricted in a similar manner to those restraints of the DNA in the Myb–DNA complex.

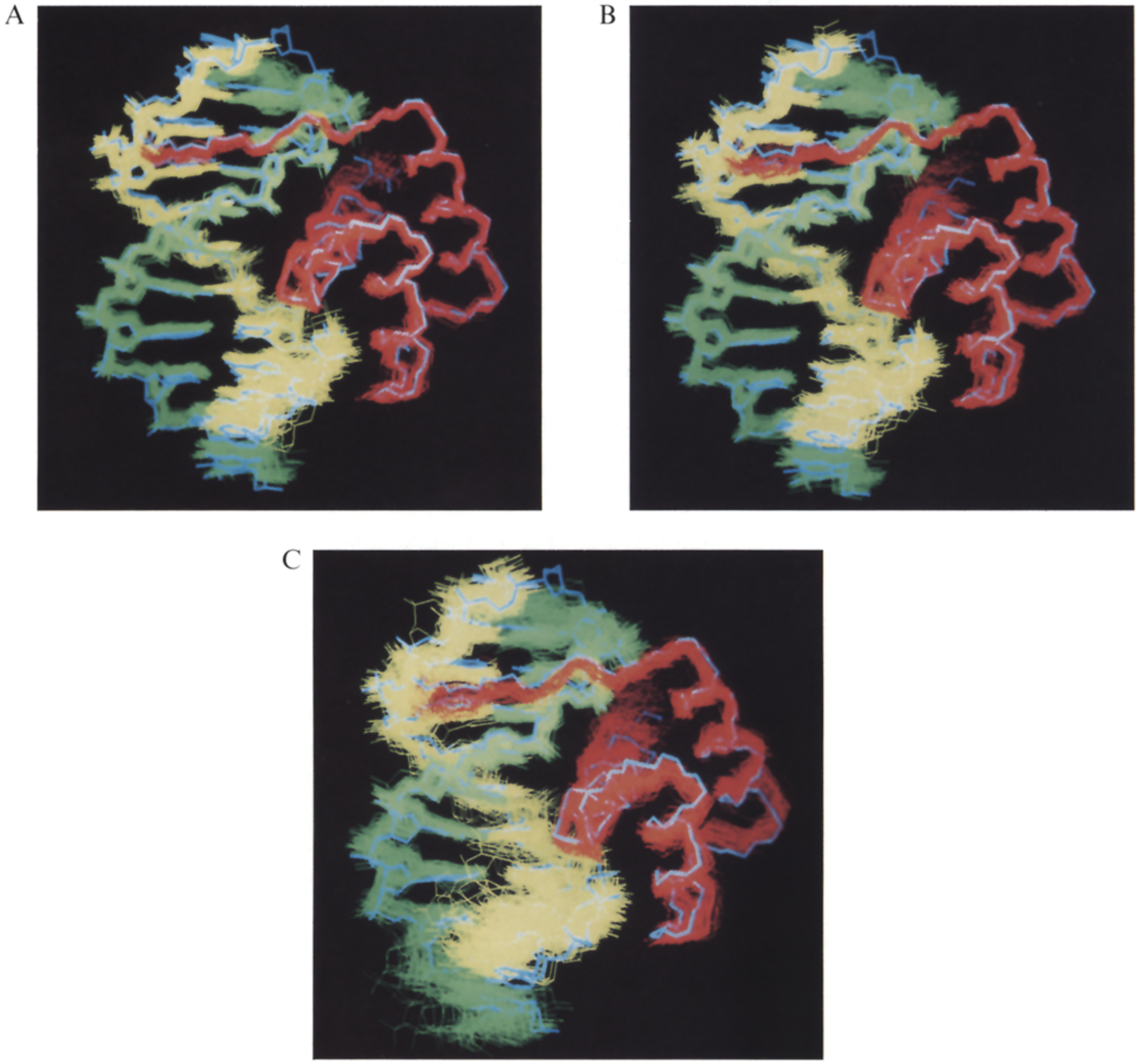

Fig. 7. The superimposed complexed structures of the homeodomain 1HDD-C and DNA11 generated by the multiple-step procedure, using different data sets of the simulated distance restraints; (A) M1, (B) M2, and (C) M3 (see Table 2). The colors are the same as those in Fig. 6.

In order to examine the dependence of the structure quality on the number of intermolecular NOEs between the protein and the DNA, several data sets were prepared. Each data set contained a different number of simulated intermolecular NOE restraints, depending upon the upper bounds of the distances between the protein and DNA protons, as shown in Table 2. The distribution of the simulated intramolecular and intermolecular distance restraints is indicated in Fig. 1.

The complexed structure of the homeodomain (1HDD-C) with an 11-base-pair double-stranded DNA (DNA11) was reconstructed with the above two protocols, the single-step and multiple-step procedures, for the different simulated restraint data sets in Table 2. In the single-step procedure, about 30–40 out of the 100 initial structures were obtained that did not display any individual distance violation larger than 0.4 Å, any dihedral angle violation larger than 4.0°, or a total energy larger than −680 kcal/mol. In the multiple-step procedure, about 50–60 out of the 100 initial structures satisfied the same geometrical and energetic criteria.

All calculations, using the well-vectorized programs EMBOSS and PRESTO (Morikami et al., 1992; Nakai et al., 1993), were carried out on a FACOM VP2600 machine at the Protein Engineering Research Institute (Osaka).

## Results and Discussion

### *Myb–DNA complex structure*

Figure 2 shows a $^{13}$C($\omega_1$)-filtered-$^{13}$C($\omega_2$)-selected 2D [$^1$H,$^1$H]-NOESY spectrum of the complex formed by the uniformly $^{13}$C-labeled Myb-R2R3 and the nonlabeled DNA16. Intermolecular NOE cross peaks are selectively observed and labeled in this figure. The precise numbers of all distance restraints have been summarized in Table 1 of our previous paper (Ogata et al., 1994); distributions of the intramolecular and intermolecular distance restraints are indicated in Fig. 3. All distance restraints have been deposited in the Protein Data Bank.

The complexed structures of Myb-R2R3 and DNA16, constructed by the single-step and multiple-step procedures using the 4D-SA protocol, are shown superimposed in Figs. 4A and B, respectively. The refined average structures, which were obtained by the protocol described in our previous report (Ogata et al., 1994), are also shown in white. The structure of the DNA terminal base pairs 12 through 16, which are not involved in protein binding, was not well defined, and so this region is excluded from the figure and from all the structural analyses below. Several residues of the N- and C-termini of Myb-R2R3 (from Met$^{89}$ to Pro$^{94}$ and from Asn$^{186}$ to Val$^{193}$) were poorly defined as well, so these were also neglected in the following analyses.

When we tried to construct the complexed structure using conventional simulated annealing in three-dimensional space (3D-SA), no structure satisfying all the given distance restraints was attained, even when the multiple-step procedure was applied. In most cases, the 3D-SA protocol failed to dispose two recognition helices of Myb-R2R3 at the correct depth in the major groove of DNA-

TABLE 3

RMSDs OF BACKBONE ATOMS AND ALL HEAVY ATOMS OF THE Myb-R2R3–DNA16 COMPLEX AMONG THE STRUCTURES PRODUCED BY TWO DIFFERENT PROCEDURES

| | Single-step procedure[a] | Multiple-step procedure[b] | Single vs. multiple-step[c] |
|---|---|---|---|
| **Complex of Myb-R2R3 and DNA16[d]** | | | |
| backbone atoms[e] | 1.95±0.29 | 1.92±0.27 | 2.00±0.29 |
| all heavy atoms | 1.92±0.22 | 1.90±0.21 | 1.98±0.24 |
| **Myb-R2R3[f]** | | | |
| backbone atoms | 1.81±0.29 | 1.70±0.29 | 1.78±0.32 |
| all heavy atoms | 2.07±0.24 | 2.00±0.25 | 2.06±0.26 |
| **Myb-R2[g]** | | | |
| backbone atoms | 1.08±0.16 | 0.97±0.16 | 1.03±0.16 |
| all heavy atoms | 1.51±0.15 | 1.43±0.15 | 1.47±0.15 |
| **Myb-R3[h]** | | | |
| backbone atoms | 0.85±0.14 | 0.82±0.12 | 0.84±0.13 |
| all heavy atoms | 1.29±0.13 | 1.28±0.12 | 1.29±0.12 |
| **DNA16[i]** | | | |
| phosphate backbone atoms[j] | 1.88±0.45 | 1.78±0.34 | 1.92±0.39 |
| all heavy atoms | 1.40±0.28 | 1.36±0.23 | 1.43±0.25 |
| **Interface between Myb-R2R3 and DNA16[k]** | | | |
| backbone atoms[e] | 1.54±0.20 | 1.51±0.22 | 1.55±0.21 |
| all heavy atoms | 1.45±0.15 | 1.42±0.17 | 1.45±0.17 |

[a] In the single-step procedure, the initial structures were different random coil Myb-R2R3 peptides and a typical B-type DNA duplex, and each center of mass was separated by 70 Å. The complex structures were constructed simultaneously by the 4D-SA protocol. Here, the final 22 structures selected after energy minimization are compared. The means ± standard deviations of $22\times(22-1)/2$ rmsd values are listed.

[b] In the multiple-step procedure, each initial structure of Myb-R2R3 and DNA16 in the first step was a different random coil and a typical B-type DNA helix, respectively. Both peptide and DNA structures were individually constructed by the 4D-SA protocol using only the intramolecular NOE information. In the second step, the converged structures of Myb-R2R3 and DNA16 from the first step were separated by 50 Å as the initial complex structure with random orientation. Adding the intermolecular NOEs, the complex structures were constructed. Here, the final 25 structures after energy minimization are compared. The means ± standard deviations of $25\times(25-1)/2$ rmsd values are listed.

[c] The means ± standard deviations of $22\times25$ rmsd values are listed between 22 structures constructed by the single-step procedure and 25 structures generated by the multiple-step procedure.

[d] The rmsd calculations were carried out with respect to residues Trp$^{95}$–Trp$^{185}$ of Myb-R2R3 and base pairs 1–11 of DNA16.

[e] Only the backbone atoms (N, C$^{\alpha}$ and C') of Myb-R2R3 and the phosphate backbone atoms (P, O5', C5', C4', C3', O3') of DNA16 were considered.

[f] The rmsd calculations were carried out with respect to residues Trp$^{95}$–Trp$^{185}$ of Myb-R2R3.

[g] The rmsd calculations were carried out with respect to residues Trp$^{95}$–Trp$^{134}$ of Myb-R2R3.

[h] The rmsd calculations were carried out with respect to residues Trp$^{147}$–Trp$^{185}$ of Myb-R2R3.

[i] The rmsd calculations were carried out with respect to base pairs 1–11 of DNA16.

[j] Only the phosphate backbone atoms (P, O5', C5', C4', C3', O3') of DNA16 were considered.

[k] The rmsd calculations were carried out with respect to the interface between Myb-R2R3 (the residues in the third helices of both R2 and R3) and DNA16, i.e., residues Gly$^{127}$–His$^{137}$ and Asp$^{178}$–Thr$^{188}$ of Myb-R2R3, and base pairs 3–10 of DNA16.

16. Only when an initial complexed structure was built by manual docking employing interactive graphics, the 3D-SA protocol could construct a complexed structure without large distance violations. This indicates the advantage of the 4D-SA protocol, having a large radius of convergence, over the conventional 3D-SA protocol, as shown previously in structure determinations of globular proteins (Nakai et al., 1993).

Since an interaction scheme between Myb-R2R3 and DNA16, based upon their complexed structure, has been described previously (Ogata et al., 1994,1995), we focus here on the quality of the complex structures determined by the two protocols, the single-step and multiple-step procedures.

The rmsd values per residue from the average coordinates of 22 final complex structures produced by the single-step procedure were calculated and are shown in Fig. 5. It is evident that the six helices in Myb-R2R3 are well defined, leaving poorly defined N- and C-termini and loops between the helices. Especially the rmsd values of the linker loop between the third helix of R2 and the first helix of R3 are very high, because of the limited number of NOE signals present, as indicated in Fig. 3. As shown in Table 3, each individual R2 and R3 structure was very well converged, but the whole Myb-R2R3 peptide, including the linker, displayed backbone rmsd values that were almost twice as large as those of each separate repeat.

The DNA16 structure was well defined only around the base positions recognized by Myb-R2R3, AACNG, as shown in Fig. 5. In fact, the rmsd values of the interface between Myb-R2R3 and DNA16 are relatively lower than those of the whole complex, Myb-R2R3 or DNA16 (Table 3). Several effective intermolecular NOE restraints, shown in Fig. 3, correctly dispose the interacting helices of R2 and R3 with the AACNG portion of the DNA duplex.

From Figs. 4A and B, the final structures and the degree of structural convergence between the single-step and multiple-step procedures seem quite similar. We calculated the averages and standard deviations of the rmsd values for $N^{opt} \times (N^{opt} - 1)/2$ pairs of structures constructed by the individual procedures. Here, $N^{opt}$ represents the number of final refined structures: 22 for the single-step and 25 for the multiple-step procedure. The rmsd values are indicated in the first and second columns of Table 3, both for the backbone atoms and for all heavy atoms. The atoms P, O5', C5', C4', C3', and O3' are defined as the DNA phosphate backbone atoms. The averages and standard deviations of the rmsd values between the two procedures, which are indicated in the third column of Table 3, were calculated for 22×25 structure pairs. The values in the third column of Table 3 are almost the same as those in the first and second columns, indicating that the structures constructed by the two different procedures are essentially equivalent.

In the single-step procedure, only 59 structures were selected as satisfying the experimental restraints, starting from 500 different random coil peptide structures. In contrast, in the multiple-step procedure, 60 docked structures could be selected by the same criteria from 100 initial structures, which were peptide and DNA molecules previously constructed using individual intramolecular restraints. In our 4D-SA protocol, it is inevitable that mirror image peptide structures are generated during the molecular dynamics calculation in four dimensions. In the current calculation, there is no direct NOE observed between the repeat 2 and repeat 3 cores of Myb-R2R3, and so there are three structural units, R2, R3, and DNA16 (see Fig. 7B of the paper by Ogata et al., 1995). Using the torsion angle restraints for the phosphate backbone, in all cases right-handed double helices were obtained for DNA16. Therefore, only one quarter of the constructed peptide structures should have the probability of assuming the correct overall structural chirality. This is the reason why the single-step procedure has poorer convergence than the multiple-step procedure. Since the final structures produced by the two procedures are equivalent, the multiple-step procedure is considered to be preferable because of its higher convergence. This conclusion is evaluated in the model calculation described below.

At first, the DNA duplex structures were constructed using putative intramolecular NOEs with hydrogen bond and torsion angle restraints for the right-handed double helix, starting from various initial structures: typical A-form, B-form, and Z-form double helices. Since the 4D-SA protocol attains an extremely large radius of convergence, as previously indicated (Nakai et al., 1993), right-handed double helices were always constructed. Therefore, only a typical B-form double helix was used as the initial estimate of DNA16 in the current calculations.

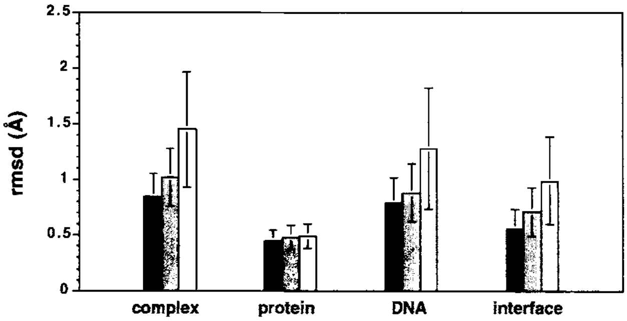

Fig. 8. The average backbone rmsd values of the complexed structure of 1HDD-C and DNA11 between the single-step and multiple-step procedures for the different simulated restraints. Dark hatched bars represent backbone rmsd values between S1 and M1, light hatched bars indicate those between S2 and M2, and white bars represent those between S3 and M3. The rmsd values were calculated for the whole complex, for only the protein region, for only the DNA duplex, and for the interface region between the protein and DNA. The standard deviation is also shown on each bar.

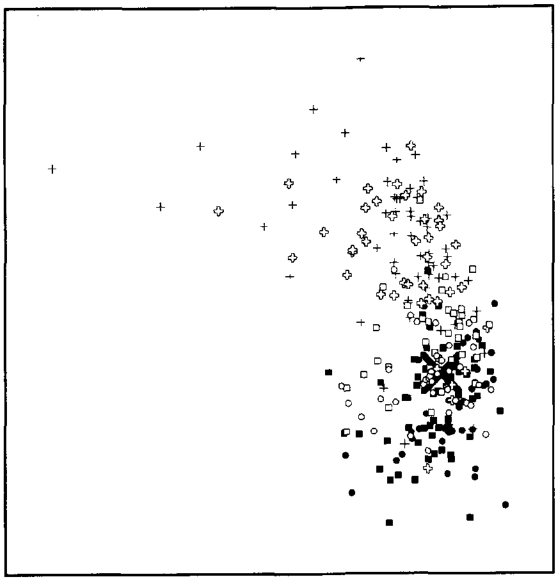

Fig. 9. Two-dimensional representation of the 279 complex structures generated with the two different procedures, using three data sets of the simulated distance restraints; S1 (○), M1 (●), S2 (□), M2 (■), S3 (✧), and M3 (+), plus the original homeodomain X-ray structure, indicated by a large cross (×).

*Model calculation for the homeodomain–DNA complex*

Using several data sets, containing different numbers of simulated intermolecular NOE restraints from the X-ray structure, the complexed structures of the homeodomain 1HDD-C with DNA11 were reconstructed by the single-step (Figs. 6A–C) and multiple-step procedures (Figs. 7A–C). The original X-ray complex structure is indicated in blue in each figure. The average rmsd values of the structures reconstructed from the simulated restraints by the individual procedures are shown in Table 2. With more than 39 intermolecular distance restraints in data sets S1, S2, M1, and M2, the two procedures are able to reconstruct the X-ray structure very well.

The averages and standard deviations of the rmsd values for $N^{opt} \times (N^{opt} - 1)/2$ pairs of reconstructed structures by the individual procedures are also summarized in Table 2. Comparison of the values in columns S1 and M1, S2 and M2, and S3 and M3, shows that the corresponding rmsd values are all similar. Moreover, the average backbone rmsd values between the two procedures shown in Fig. 8 are similar to the corresponding rmsd values obtained by the individual procedures in Table 2. This means that both procedures give almost equivalent structural ensembles.

In the two-dimensional representation of the distribution in Fig. 9, the situation is more explicitly illustrated. The figure was obtained by principal component analysis of a $(37+37+44+45+57+59+1) \times 280$ matrix, the elements of which were the rmsd values between any pair of the 280 structures generated by the two different procedures using three restraint data sets, including the X-ray structure. When the data sets with only 17 intermolecular NOE restraints were used (S3 and M3), the reconstructed structures significantly deviated from the original X-ray structure in both procedures. In contrast, when the data sets with 39 or 89 intermolecular restraints were used (S1, S2, M1, and M2), the two procedures gave overlapping clusters, all of which contained the X-ray structure.

In the single-step procedure, 45–59 structures were selected as satisfying the simulated NOE restraints, starting from 100 different random coil peptide structures. In

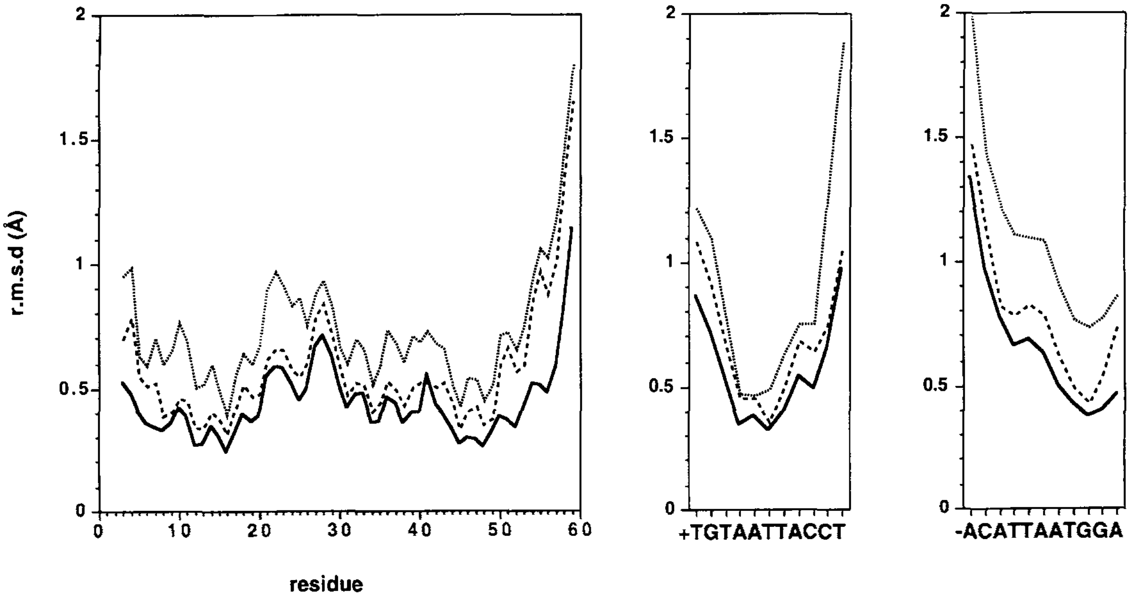

Fig. 10. The backbone rmsd values per residue from the average coordinates of the final complexed structures of the homeodomain 1HDD-C and DNA11 produced by the single-step procedure, constructed from three different data sets of the simulated distance restraints. The solid lines indicate rmsd values of the backbone atoms in 1HDD-C and the phosphate backbone atoms in DNA11, using data set S1. The dashed lines represent those from data set S2, and the dotted lines are from data set S3.

the multiple-step procedure, 68–82 docked structures were selected by the same criteria from 100 initial structures, which were peptide and DNA molecules constructed previously, using individual intramolecular restraints. The differences between the two procedures are less revealing than those found in the determination of the Myb-R2R3–DNA complex. This is because 1HDD-C is composed of only one structural unit, and half of the reconstructed peptide structures should have the probability of assuming the wrong chirality.

The backbone rmsd values per residue from the average coordinates of the final complex structures by the single-step procedure were calculated for three data sets, S1, S2 and S3, as shown in Fig. 10. Due to the absence of intermolecular restraints between residue 25 and the DNA in the S3 data set (see Fig. 1), the rmsd values per residue in the loop between the first and second helices are significantly large.

It is interesting that, in spite of the few intermolecular restraints in data sets S3, M3, or even S4, the backbone rmsd values of the interface between 1HDD-C and DNA-11 are as small as 1 Å (Table 2). In these cases, however, DNA11 and the whole complex structure have large rmsd values. These results suggest that only a few substantial intermolecular distance restraints are required to determine the relative disposition of the rigid structural units. In other words, the distribution of distance restraints in the whole complex is very different from the distribution in a domain of a globular protein of the same molecular size. Rather, the distribution inherently indicates the two different structural domains. This may be the reason why the current multiple-step procedure can provide a correct answer where the structural ensemble overlaps with that determined by the simple single-step procedure. In both procedures, the conformation of DNA11 is not definitively determined by the intramolecular distance restraints alone, but the base pairs at the interface between 1HDD-C and DNA11 are also well defined.

## Conclusions

In this paper, it was shown that the multiple-step procedure using the 4D-SA protocol can construct the specific DNA complex of the minimum DNA-binding domain of the c-Myb protein. This was accomplished with a high convergence rate; moreover, the searched conformational space was as large as that used by the simple single-step procedure. Employing model calculations, using the X-ray structure of the homeodomain–DNA complex, it was also confirmed that the two procedures yield equivalent structural ensembles. The dependence of the structural quality upon the number of intermolecular restraints was investigated, and it is suggested that only a few substantial intermolecular distance restraints are sufficient to determine the relative disposition of a protein and a DNA fragment. The current multiple-step procedure is applicable to the solution structure determination of not only protein–DNA complexes, but also of various complexes of biomolecules.

## Acknowledgements

We are grateful to Taiko Inoue and Hiroko Kanai for help with the preparation of the DNA16 sample, and to Akinori Kidera for helpful discussions. This work was partly supported by Grants in Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, and Culture of Japan.

## References

Arnot, S. and Hukins, D.W.L. (1972) *Biochem. Biophys. Res. Commun.*, **47**, 1504–1510.

Billeter, M., Qian, Y.Q., Otting, G., Muller, M., Gehring, W. and Wüthrich, K. (1993) *J. Mol. Biol.*, **234**, 1084–1097.

Boelens, R., Lamerichs, R.M.J.N., Rullmann, J.A.C., Van Boom, J.H. and Kaptein, R. (1988) *Protein Seq. Data Anal.*, **1**, 487-498.

Chuprina, V.P., Rullmann, J.A.C., Lamerichs, R.M.J.N., Van Boom, J.H., Boelens, R. and Kaptein, R. (1993) *J. Mol. Biol.*, **234**, 446–462.

De Vlieg, J., Berendsen, H.J.C. and Van Gunsteren, W.F. (1989) *Protein Struct. Funct. Genet.*, **6**, 104–127.

Graf, T. (1992) *Curr. Opin. Genet. Dev.*, **2**, 249–255.

Gronenborn, A.M. and Clore, G.M. (1989) *Biochemistry*, **28**, 5978–5984.

Havel, T.F. (1990) *Biopolymers*, **29**, 1565–1585.

James, T.L. (1994) *Curr. Opin. Struct. Biol.*, **4**, 275–284.

Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B. and Pabo, C.O. (1990) *Cell*, **63**, 579–590.

Kuszewski, J., Nilges, M. and Brünger, A.T. (1992) *J. Biomol. NMR*, **2**, 33–56.

Luscher, B. and Eisenman, R.N. (1990) *Genes Dev.*, **4**, 2235–2241.

Morikami, K., Nakai, T., Kidera, A., Saito, M. and Nakamura, H. (1992) *Comput. Chem.*, **16**, 243–248.

Nakai, T., Kidera, A. and Nakamura, H. (1993) *J. Biomol. NMR*, **3**, 19–40.

Ogata, K., Hojo, H., Aimoto, S., Nakai, T., Nakamura, H., Sarai, A., Ishii, S. and Nishimura, Y. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 6428–6432.

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S. and Nishimura, Y. (1994) *Cell*, **79**, 639–648.

Ogata, K., Morikawa, S., Nakamura, H., Hojo, H., Yoshimura, S., Zhang, R., Aimoto, S., Ametani, Y., Hirata, Z., Sarai, A., Ishii, S. and Nishimura, Y. (1995) *Nature Struct. Biol.*, **2**, 309–320.

Omichinski, J.G., Clore, G.M., Schaad, O., Felsenfeld, G., Trainor, C., Appella, E., Stahl, S. and Gronenborn, A.M. (1993) *Science*, **261**, 438–446.

Qian, Y.Q., Otting, G., Billeter, M., Muller, M., Gehring, W. and Wüthrich, K. (1993) *J. Mol. Biol.*, **234**, 1070–1083.

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

Zhang, H., Zhao, D., Revington, M., Lee, W., Jia, X., Arrowsmith, C. and Jardetzky, O. (1994) *J. Mol. Biol.*, **238**, 592–614.