

Solving DNA Structures by *MERLOT*

BY RAJAGOPAL CHATTOPADHYAYA\* AND PINAKPANI CHAKRABARTI

Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90024, USA

(Received 13 November 1987; accepted 22 June 1988)

**Abstract**

*MERLOT*, a molecular replacement computer package, has been successfully used for the determination of initial phases for five different oligonucleotide structures. Two of these, CCGG and CGCGCGTTTTCGCGCG, with 316 and 323 non-H atoms respectively in the asymmetric unit, have been solved *ab initio* using the *MERLOT* package. The rest are re-examinations of previously solved DNA oligomer structures, ioCCGG, CGCGAATTCGCG and CGCGAATTbrCGCG, with 160, 486 and 488 non-H atoms in the asymmetric unit respectively. Problems involved in applying the molecular-replacement technique to DNA structures, and their possible solutions, are discussed.

**Introduction**

The crystal structures of about 35 different DNA oligomers solved in recent years (Dickerson, 1987) have provided detailed knowledge about sequence-dependent three-dimensional structures of DNA. In addition to detailed pictures of A- and B-type DNA helices, several left-handed or Z-DNA structures have been discovered by these studies.

Among these 35 different structures, many crystallized in a fashion isomorphous with another structure previously studied, and hence the phase problem was easily solved in those cases. For example, the A-DNA sequences GGGGCCCC (McCall, Brown & Kennard, 1985), GGGCTCC (Brown, Kennard, Kneale & Rabinovich, 1985), GGGTGCCC (Rabinovich, Haran, Eisenstein & Shakked, 1988) all crystallized in isomorphous fashion in space group  $P6_1$  with GGTATACC, for which the crystal structure had been solved earlier (Shakked *et al.*, 1981). Similarly, variants of the B-DNA sequence CGCGAATTCGCG (Drew *et al.*, 1981; Drew, Samson & Dickerson, 1982; Fratini, Kopka, Drew & Dickerson, 1982; Hunter, Brown, Anand & Kennard, 1986; Hunter, Brown, Kneale, Anand, Rabinovich & Kennard, 1987; Brown, Hunter, Kneale, Anand, Rabinovich & Kennard, 1986) all crystallized in space group  $P2_12_12_1$  in isomorphous fashion, and variants of the Z-DNA sequence CGC-

GCG (Wang *et al.*, 1979; Wang, Hakoshima, van der Marel, van Boom & Rich, 1984; Wang, Gessner, van der Marel, van Boom & Rich, 1985; Ho *et al.*, 1985; Chevrier *et al.*, 1986) also crystallized isomorphously with the native sequence. Actually, about 75% of the 35 solved DNA structures belong to one of these isomorphous families. When a new DNA crystal is not isomorphous with any of the previously solved DNA structures, phase information can be obtained using single- or multiple-derivative data. However, it is extremely hard, and sometimes impossible, to obtain a good derivative data set for most DNA crystals. In such a case the molecular-replacement method might be successful in providing a phasing model.

The molecular-replacement method can only work if one has a reasonably good model for the structure in question. Based on the fibre structures, and the recently determined crystal structures for various sequences, one can often construct a reasonable model for the unknown structure in various helical forms. Given this existing body of knowledge about possible structures, the molecular-replacement method becomes the method of choice, provided, of course, the unknown resembles one of the known helical forms of DNA.

A separate computer program, *ULTIMA* (Rabinovich & Shakked, 1984) was used to solve the structure of the octamer GGTATACC. This program is a multidimensional search approach incorporating packing criteria with *R*-factor calculations for low-resolution X-ray data. Initially, we tried to use *ULTIMA* to solve the structure of CGCGCGTTTTCGCGCG (Chattopadhyaya, Ikuta, Grzeskowiak & Dickerson, 1988) thought to exist as a DNA hairpin molecule. Both B and Z models for 75% of the molecule were tried by *ULTIMA*, but the correct combination of rotations and translations was not obtained, probably because we did not have the full model for the structure. This led us to examine *MERLOT* (Fitzgerald, 1988), which is an integrated package of computer programs for the determination of phases using the molecular-replacement technique. Once *MERLOT* had predicted the correct solution for this structure, we were tempted to study other DNA structures using the same package. The computer programs in *MERLOT* are extremely well documented and user friendly, and a detailed description is provided by the author.

\* Present address: Department of Biochemistry, University of California, Berkeley, CA 94720, USA.

The molecular-replacement technique usually involves replacing the unknown structure with a properly oriented and positioned molecule of identical or related structure. The orientation is described by three Euler angles determined in the rotation-function calculation. In *MERLOT*, one can either use *CROSUM*, the fast rotation function of Crowther (1972), or *LATSUM*, the rotation function of Lattman & Love (1970) for this purpose. For determining the translation vector, three alternative ways are provided in *MERLOT* — *TRNSUM*, the translation function (Crowther & Blow, 1967); *RVAMAP*, *R*-factor search (Fitzgerald, 1988); and *PAKFUN*, a packing function (Fitzgerald, 1988). In addition, the package includes *RMINIM*, a rigid-body refinement procedure (Ward, Wishner, Lattman & Love, 1975), and several other useful programs for set-up, output and analysis purposes.

The strategy recommended as a result of the present studies can predict the correct solution quite efficiently, and very few combinations of possible orientations and translations, if not one, need to be further examined by subsequent refinement. Here we will discuss the five different DNA oligomer structures which have been examined by the *MERLOT* package, one by one.

### ioCCGG

The structure of this iodinated sequence was originally solved using anomalous phasing (Conner, Takano, Tanaka, Itakura & Dickerson, 1982). The original 2.1 Å crystal data and the refined coordinates are deposited with the Brookhaven Protein Data Bank. Five different models were generated for this test case in our calculation: (a) A-DNA fibre model for d(CCGG)<sub>2</sub>, (b) B-DNA fibre model for d(CCGG)<sub>2</sub>, (c) A-DNA fibre model for d(ioCCGG)<sub>2</sub>, (d) B-DNA fibre model for d(ioCCGG)<sub>2</sub>, and (e) the refined d(ioCCGG)<sub>2</sub> coordinates without solvent molecules. In each case, the model was placed in a *P1* cell measuring 40 Å on all three sides with all cell angles 90°.

From the *P4<sub>3</sub>2<sub>1</sub>2* cell constants ( $a = b = 41.1$ ,  $c = 26.7$  Å) of the unknown crystal, it can be predicted that there is only one tetramer duplex molecule per asymmetric unit. Program *CROSUM* was run with each of these five different input models for the tetramer duplex, and parameters such as the resolution range of reflections and the Patterson cut-off radius were varied. In some runs, where 8.0–2.1 Å data were used, the Patterson cut off was either 10 or 11.5 Å; in some other runs, where 8.0–3.0 Å data were used, the Patterson cut off was always 11.5 Å. In *CROSUM*, the Patterson cut off has to be less than or equal to six times the high-resolution limit. For the unknown structure in space group *P4<sub>3</sub>2<sub>1</sub>2*, the following rotation-function space was searched:  $\alpha$  from 0 to 88.75° in steps of 1.25°,  $\beta$  from 0 to 90° in steps of 5.0°, and  $\gamma$  from 0 to 355° in steps of 5.0°. This is 1/8th the complete range

Table 1. *Rotation-function peak heights for crystal data from ioCCGG using various input models*

Peak heights are expressed relative to that obtained from the refined model. All calculations used 8.0–3.0 Å data and a Patterson cut off of 11.5 Å.

Description of input model	Maximum peak	Second-highest peak
Refined ioCCGG structure	1.00	—
ioCCGG fibre model, A-DNA	0.88	0.63
ioCCGG fibre model, B-DNA	0.74	0.73
CCGG fibre model, A-DNA	0.62	0.59
CCGG fibre model, B-DNA	0.62	0.61

of Euler angles (Rao, Jih & Hartsuck, 1980). The user has control over the range and the grid size in  $\beta$  only, and the parameters for the other two angles are fixed internally in the program based on the space-group information provided by the user.

For each model, it appeared to matter little whether 3 or 2.1 Å data were used, and the two different Patterson cut offs also had negligible effect on the peaks. As expected, the refined model gave the strongest rotation-function peak, followed by the A-DNA fibre model with I atoms, followed by the B-DNA model with I atoms, followed by both fibre models without I atoms. Without the I atoms, the maximum intensity of the rotation-function peak was about the same with both fibre models. Although the peak was correctly predicted even without the use of the I atoms, the runs with the A-DNA fibre model with I atoms showed fewer spurious peaks compared with the other fibre models. The B model with I atoms showed less discrimination between peaks compared with the A model. Quantitative results for these searches are summarized in Table 1. For the A fibre model with I atoms, there were two major peaks, but they turned out to be equivalent as the model itself possessed pseudo-symmetry. Thus, based on the *intensities* of these peaks, it could be predicted that the model resembles the A-DNA fibre model.

The Euler angles were further refined by using searches with narrower ranges and smaller grid sizes about the peak positions in the program *LATSUM*. A global search with *LATSUM* is impractical as it takes too much computer time. The advantage of *LATSUM* is that the user can specify the grid size of all three angles.

After suitably rotating the various models, *TRNSUM* and *RVAMAP* were run with 8.0–3.0 Å data. From the various translation-function maps, the fractional translations FA and FB were obtained unambiguously for the A model. However, the fractional translation FC could not be obtained from the relevant *TRNSUM* runs. The program *RVAMAP*, however, determined all three fractional translations without ambiguity for the A model, and this agreed with the crystal structure. The values of FA and FB found from *TRNSUM* were also confirmed by *RVAMAP*. The program *PAKFUN* was also run for the A fibre

model with I atoms, and this showed that the FA, FB and FC values obtained earlier are correct.

The program *RMINIM* was run for each model for a given set of angles and translations. The refined crystal structure, of course, had the lowest *R* factor (0.30), and the A fibre model with I atoms had the second lowest *R* factor (0.56), while all other models had *R* values of 0.65 or higher, when 8.0–3.0 Å data were used with a constant temperature factor of 15 Å<sup>2</sup>.

### CCGG

This sequence was in fact crystallized before the iodinated derivative was studied, but the structure could not be solved as all covalent heavy-atom derivatives crystallized in different space groups, and diffusion of heavy-metal ions did not succeed in giving a suitable derivative. Calculation from the *P*<sub>2</sub><sub>1</sub> cell dimensions (*a* = 22.17, *b* = 26.62, *c* = 30.38 Å and  $\beta = 105.65^\circ$ ) and a measured density of 1.52 g cm<sup>-3</sup> showed that if the asymmetric unit contained one d(CCGG)<sub>2</sub> molecule, the crystal would be 30.5% DNA by weight; if it had two such d(CCGG)<sub>2</sub> molecules in the asymmetric unit, then the crystal would be 61% DNA by weight. On average, DNA crystals contain about 50% DNA by weight, neglecting counter cations. Therefore, it is more likely that there are two tetramer duplexes per asymmetric unit in this crystal.

Two different input models were used for the *ab initio* structure solution – one was the A-DNA model generated from fibre coordinates (Arnott & Hukins, 1972) for d(CCGG)<sub>2</sub>, and the other was the crystal structure model obtained from d(ioCCGG)<sub>2</sub> (Conner *et al.*, 1982) in which the two I atoms were deleted. The B model was not pursued as CCGG sequences have always been found to exist in the A form in the crystals studied so far.

The models were placed in a *P*<sub>1</sub> cell measuring 40 Å on all three sides. The following rotation-function space was searched for both models:  $\alpha$  from 0 to 177.5° in steps of 2.5°,  $\beta$  from 0 to 180.0° in steps of 5.0°, and  $\gamma$  from 0 to 355.0° in steps of 5.0°. The program *CROSUM* was run for both models using 8.0–3.0 Å data and 8.0–2.5 Å data, with a Patterson cut-off radius of 10 Å. Both resolution ranges gave similar rotation-function peaks. With the crystal-structure input model, there were four peaks with intensities above 65% of the maximum. Two of these were the major peaks and were nearly equal (100, 98%) in intensity. The two other peaks were at 85 and 76% levels and were judged to be spurious. When the fibre A-DNA model was used in *CROSUM*, the intensity of the highest peak was only 75% compared with the highest peak from the crystal-structure model and, in addition, there was a greater number of peaks above the 65% threshold level in that search. Therefore, the discrimination between peaks appeared to be worse

when the fibre was chosen. This indicated that the model obtained from the iodinated crystal resembled both independent tetramer duplexes in the native CCGG crystal, and probably the two major peaks of nearly equal intensity corresponded to the orientation of the two independent tetramer duplexes. An additional *CROSUM* run for the crystal-structure model using 8.0–4.0 Å data with a Patterson radius of 12 Å showed less discrimination between peaks compared with the 3.0 and 2.5 Å runs.

All four orientations predicted by the rotation-function calculation for the crystal-structure input model were drawn using the program *PLUTO* outside *MERLOT*. They all had their helix axes along the *b* direction of the crystal. In fact this can be predicted simply by the existence of some strong reflections at 3.0–3.2 Å resolution around (0,8,0). In another *CROSUM* run with the crystal-structure input model, 13 of the strongest reflections around (0,8,0) were removed from the input list of reflections. Although the number of peaks above the 65% threshold increased as a result, the first two major peaks from the earlier search were still present, while the third and fourth peaks were missing in this *deletion* search. In yet another *CROSUM* run, the complete list of reflections was used, but all base atoms were deleted from the input list of atoms, leaving just the sugar-phosphate backbone as the model. Here also, the two major peaks were still present, while the other two peaks were absent. In both of these latter runs, the intensities of the peaks were lower than the search where the complete model and the complete list of reflections were used, the number of peaks above 65% of the maximum was larger and the major peaks were slightly shifted from their original locations. The purpose of these runs was to separate the real peaks from the spurious ones.

After both of the independent orientations were further refined using *LATSUM*, the rotated models were separately used for running *TRNSUM*. In *MERLOT*, there is currently no provision for running *RVAMAP* or *PAKFUN* when there is more than one molecule in the asymmetric unit. Therefore, the only positioning tool in this case is *TRNSUM*. In the space group *P*<sub>2</sub><sub>1</sub>, this involves looking at the Harker section  $y = \frac{1}{2}$  for the vector between two tetramer molecules related by a 2<sub>1</sub> axis. Two separate runs of *TRNSUM* for each independent orientation yielded values for FA1, FC1, FA2 and FC2. These values were all close to zero, which means the 2<sub>1</sub>-related neighbour for each independent molecule is directly above itself in the *y* direction. If the first independent molecule is set close to (0,0,0), then it follows from packing considerations that the second independent molecule has to be close to (0.5, 0, 0.5). Therefore, 0.5 was added to the values of FA2 and FC2 obtained from *TRNSUM*. FB1 for anyone of the independent molecules can be arbitrarily set to zero, but FB2 for the other independent molecule

has to be found from another *TRNSUM* run where the vector between the two independent molecules is considered. This indicated the relative distance between both tetramer duplexes along the  $y$  direction, *i.e.* FB2 – FB1 and the values of FA2 – FA1 and FC2 – FC1 for this peak were consistent with the FA1, FA2, FC1 and FC2 values obtained previously from the other two *TRNSUM* runs.

Each independent molecule was positioned according to the result from *TRNSUM* and they were separately refined with *RMINIM* to improve their individual rotational and translational parameters. Although *RMINIM* allows simultaneous rigid-body refinement of two identical but independent molecules, we decided to use *CORELS* (Sussman, Holbrook, Church & Kim, 1977) as it takes care of van der Waals contacts between independent molecules whereas *RMINIM* does not. With these two molecules, the  $R$  factor was 0.54 with 8.0–4.5 Å data in the beginning, and dropped to 0.46 using 8.0–2.9 Å after simultaneous rigid-body refinement of the two domains in *CORELS*. The detailed description of the refinement of this structure will be reported elsewhere (Quintana, Chattopadhyaya, Conner & Dickerson, 1988).

### CGCGCGTTTTCGCGCG

The structure of this 16-base DNA oligomer was recently solved using the *MERLOT* package. It was predicted (Ikuta, Chattopadhyaya, Ito, Dickerson & Kearns, 1986) that the molecule forms a so-called single-stranded *hairpin* structure. As this represents the first crystal structure of a DNA hairpin, a good starting model was not available for the  $T_4$  portion of the structure. However, the remaining portion, or the stem region of the structure, could exist either in the B or the Z form. Accordingly, two models were selected for the *MERLOT* calculations: one was a fibre B-DNA model for d(CGCGCG)<sub>2</sub> (Arnott & Hukins, 1972), the other was simply the crystal-structure model for the same duplex in the Z form (Wang *et al.*, 1979). In either case, the input model had 240 non-H atoms, as opposed to 323 in the entire hexadecamer molecule. The C2 cell dimensions ( $a = 57.18$ ,  $b = 21.63$ ,  $c = 36.40$  Å and  $\beta = 95.22^\circ$ ) suggested that there was only one hexadecamer per asymmetric unit.

For the *CROSUM* calculation, either model was placed in a P1 cell measuring 50 Å on all three sides. For space group C2, the following rotation-function space was searched:  $\alpha$  from 0 to 180° in 2.5° intervals,  $\beta$  from 0 to 180° in 5° intervals, and  $\gamma$  from 0 to 355° in 5° intervals. The resolution ranges used were 8.0–3.0 and 8.0–2.5 Å, and a Patterson cut-off radius of 15 Å. The B model could be eliminated solely on the basis of the *CROSUM* result, as the maximum intensity of the rotation-function peak was only 40% of the

maximum value of the peak obtained from the Z model under identical conditions. The main difficulty was deciding which of the ten rotation-function peaks obtained with the Z model was the real peak. Eventually it turned out that owing to pseudosymmetry in the input model only five peaks were independent. However, the correct rotation-function peak could be sorted by using the *deletion* searches of the type mentioned for the previous example. In numerous *CROSUM* runs performed with 8.0–2.5 Å data, various numbers (between 2 and 28 from a total of about 1700 reflections in the range) of the strongest reflections were deleted from the input list of reflections. In all runs, the two major peaks (found to be equivalent later due to pseudosymmetry of the model) were consistently present. This enabled us to decide upon the correct peak from the *CROSUM* runs.

The two orientations were refined by *LATSUM*. *TRNSUM* and *RVAMAP* were run using the refined angles. For C2, one needs only to find FA and FC, and FB is arbitrarily set to zero. Therefore, *TRNSUM*, *RVAMAP* and *PAKFUN* are all essentially two-dimensional searches at  $y = 0$ . Unlike other cases, *TRNSUM* was not as discriminative with this model, although the highest peak agreed with that from *RVAMAP*. The second-highest peak was 90% in intensity compared with the highest peak in *TRNSUM*. FA and FC values obtained did not show any bad contacts in *PAKFUN*. *RVAMAP* was run using 8.0–3.0 and 8.0–2.4 Å data; the correct solution topped the list in both reflection ranges. This solution corresponded to an  $R$  factor of 0.53 using 8.0–3.0 Å data and a temperature factor of 15 Å<sup>2</sup>. After *RMINIM*, this solution gave an  $R$  factor of 0.49, while all other combinations had  $R$  values of 0.54 or higher after *RMINIM*. Most of these combinations were also unreasonable in terms of packing, and therefore not pursued.

The remaining portions of the molecule came out from successive difference maps after refinement, details of which will be published elsewhere.

### CGCGAATTCGCG and CGCGAATTbrCGCG

These structures were solved some years ago (Drew *et al.*, 1981; Fratini *et al.*, 1982) and will be considered together as they are related. The original crystal data and the coordinates are available from the Brookhaven Protein Data Bank for both oligomers. Although the crystals for these two sequences are nearly isomorphous, and the sequence is identical except for the Br atom at the ninth cytosine on each strand of the duplex, significant differences were found in their crystal structures. Both structures also have significant differences from the ideal fibre model of B-DNA. Our interest in these cases was to see whether *MERLOT*

could predict the correct rotation and translation starting from fibre models.

The model chosen for both structures was the fibre B-DNA structure (Arnott & Hukins, 1972) generated for this dodecamer sequence. The  $P2_12_1$  cell constants suggested one duplex molecule per asymmetric unit. For the runs with the brominated sequence, two additional Br atoms, covalently attached to the C5 atom of the ninth cytosine residue in each strand, were used. A sample run was also performed with a fibre A-DNA model for the native dodecamer data set, but this was not pursued further as the peak intensity of the maximum was only 47% of that obtained for a similar run with the fibre B model.

The models were oriented with their helix axes along the  $x$  direction (if the helix direction was along  $z$ , streaky peaks in  $\beta = 0$  and  $180^\circ$  sections resulted) and placed in a  $P1$  cell measuring  $80 \text{ \AA}$  along  $x$  and  $40 \text{ \AA}$  along the other two directions. Since all our calculations were done using actual coordinates, the three sides of the  $P1$  cell need not be equal to each other. However, if one wishes to use the transform option in *MERLOT*, these have to be equal. For the space group  $P2_12_1$ , the following rotation-function space was searched:  $\alpha$  from  $0$  to  $177.5^\circ$  in  $2.5^\circ$  steps,  $\beta$  from  $0$  to  $90^\circ$  in  $5^\circ$  steps, and  $\gamma$  from  $0$  to  $355^\circ$  in  $5^\circ$  steps. Various resolution ranges and Patterson cut offs were used for both the native dodecamer and the brominated sequence:  $8.0\text{--}2.6 \text{ \AA}$  data, radius  $12.5 \text{ \AA}$ ;  $8.0\text{--}3.0 \text{ \AA}$  data, radius  $17.5 \text{ \AA}$ ;  $8.0\text{--}4.0 \text{ \AA}$  data, radius  $20 \text{ \AA}$ ; and  $8.0\text{--}5.0 \text{ \AA}$  data, radius  $20 \text{ \AA}$ . The radius was not increased above  $20 \text{ \AA}$  to prevent the intermolecular vectors from being included in the *CROSUM* calculation.

For these oligomers, the high-resolution searches using  $8.0\text{--}2.6$  or  $8.0\text{--}3.0 \text{ \AA}$  data were not useful, as the discrimination between peaks was poor with those resolution ranges. For instance, for the native dodecamer, there were 32 peaks above the 65% threshold level in the  $3 \text{ \AA}$  search, all between 94 and 100% of the maximum peak intensity. Really, this corresponds to 16 independent peaks, as the model itself possesses pseudosymmetry, but that number is still too high. This prompted us to use other resolution ranges, using  $4.0$  and  $5.0 \text{ \AA}$  data. As the model used did not represent either structure at high resolution, such behaviour can be expected. However, the correct rotation peak showed more discrimination when  $8.0\text{--}4.0 \text{ \AA}$  data were used for the brominated structure; all other peaks had intensities that were 89% or lower. This indicated the correct rotation for the brominated structure. Incidentally, the same peak was at the top of the list for the brominated dodecamer when the  $3.0 \text{ \AA}$  search was performed, but in the  $2.6 \text{ \AA}$  search, it was the 11th peak with 98.8% of the maximum in that search. However, when the  $3.0 \text{ \AA}$  search was repeated with ten of the strongest reflections around  $(0,0,18)$

deleted, the same peak that topped the list in the  $4.0 \text{ \AA}$  search was the maximum here, and the next peak was only 80% of its intensity. These tests confirmed the orientation of the molecule in the crystal, and it was further confirmed by *PLUTO* drawings and comparison with the published structure. When the base atoms were deleted from the input list of atoms and the complete list of reflections used, however, the results were not too conclusive. For the native dodecamer, the  $8.0\text{--}5.0 \text{ \AA}$  search, rather than the  $8.0\text{--}4.0 \text{ \AA}$  search, was the most discriminative. The combination of Euler angles was similar (all three angles within  $5^\circ$  for both models) to those obtained for the brominated model, but not identical. This is consistent with the observations made earlier (Fratini *et al.*, 1982), where the structure of the native was used as an initial model for the structure solution of the brominated sequence. In the  $5.0 \text{ \AA}$  search for the native, the correct peak had the maximum intensity, and the next-highest peak was at the 94% level. However, when the search was repeated using  $8.0\text{--}3.0 \text{ \AA}$  data, deleting some 12 strongest reflections close to  $(0,0,20)$ , the correct peak was 15% higher in intensity compared with the next-highest peak.

Compared with the native dodecamer, the correct peak showed more discrimination for the brominated dodecamer because of two reasons: firstly, the brominated dodecamer is closer to the fibre structure in the sense that its helix axis is almost straight like the fibre model, whereas the helix axis of the native was found to possess a  $19^\circ$  bend; secondly, the inclusion of the heavy atom in the calculation helps in sifting the correct peak from incorrect ones. In the fibre model, the two Br atoms in the duplex are about  $16.7 \text{ \AA}$  apart, therefore a Patterson cut off of  $17.5$  or  $20 \text{ \AA}$  would have included the Br-Br vector.

*LATSUM* was not performed for these oligomers; instead, it was decided to refine the angles later by *RMINIM*. The angles from *CROSUM* were accurate enough for low-resolution *TRNSUM* runs, but not for a global *RVAMAP* run. Indeed, when a global *RVAMAP* was run, the lowest  $R$  factor was  $0.72$  using  $8.0\text{--}4.0 \text{ \AA}$  data. In order to obtain a reasonable  $R$  factor, the Euler angles have to be quite accurate. *TRNSUM* was run using only  $25\text{--}8 \text{ \AA}$  data using the oriented model. From the three *TRNSUM* runs, FA and FB could be determined unambiguously, but there were two possible solutions for FC. A one-dimensional *RVAMAP* using  $25\text{--}8 \text{ \AA}$  data and assuming the values of FA and FB from *TRNSUM* confirmed one of the values of FC obtained from the earlier *TRNSUM* run. Rigid refinement was continued using *RMINIM*, gradually increasing the resolution range from  $25\text{--}8 \text{ \AA}$  to  $8\text{--}3 \text{ \AA}$ . The  $R$  values obtained for both dodecamers using this strategy are given in Table 2. Again, one notices that for the brominated dodecamer the  $R$  value does not go up as fast with high-resolution data for it resembles the fibre model better than the native dodecamer.

Table 2. *Dependence of R factors obtained using a B-DNA fibre model for the native and brominated dodecamers on the amount of data considered*

Sequence	Resolution range					
	25–8 Å	25–6 Å	10–5 Å	8–4.3 Å	8–4.0 Å	8–3 Å
CGCGAATTCGCG	0.33	0.43	0.50	0.62	0.63	—
CGCGAATTCrCGCG	0.37	0.44	0.44	0.50	0.52	0.59

### Discussion

These examples illustrate that the *MERLOT* package may be conveniently used for the initial phasing of DNA oligomer structures. In general, DNA molecules possess a high degree of internal symmetry, which gives rise to multiple peaks, apparently representing ambiguity in the rotation function (Rabinovich & Shakked, 1984). It appears that, even though there may be multiple peaks from *CROSUM*, the correct model and orientation can be guessed by a variety of searches in which (i) various resolution ranges of reflections are used, (ii) various input models are tried, (iii) some of the strongest reflections corresponding to the parallel stacking of bases are omitted from the search, (iv) the base atoms are omitted from the input model, leaving only the sugar–phosphate backbone. From these experiments, one may be able to guess the correct model and its orientation, solely on the basis of the intensities of the rotation-function peaks, and the discrimination shown between various peaks at various resolution ranges. If the model is accurate, including high-resolution data (e.g. 2.5 or 3.0 Å) helps in the discrimination between various rotation peaks; on the other hand, if the model is coarse, lower-resolution searches (e.g. 4 or 5 Å) will be more discriminative. However, in low-resolution searches (e.g. 8.0–5.0 Å data), even the direction of the helix axis is incorrect for many rotation peaks. The rough direction of the helix axis can be easily predicted in most crystals. Usually an 8.0–3.0 Å search will establish the possible orientations, as all of these solutions agree with the true one regarding the rough direction of the helix axis in the crystal. The more difficult problem is to determine the rotation about the helix axis, and the strategies mentioned above may help in this respect. The true orientation, then, will be determined by the resolution range that shows the most discrimination (e.g. where the second peak is 90% in intensity or lower), and the deletion searches using high-resolution data. The true solution will always appear in all searches, whereas spurious ones may not. Based on these experiments, it is our contention that the rotation problem can be routinely solved for any unknown DNA structure.

Once the rotation problem has been solved, the translation problem can also be solved using a combination of the three approaches provided in the

*MERLOT* package. It is to be borne in mind that the success of the translation problem depends very much on the accuracy of the Euler angles, usually refined with the help of *LATSUM*. The rotation function will usually give a clue as to the accuracy of the model, and the resolution range for the translation function should be chosen accordingly. For example, with the first three examples, the translation problem could be solved using 8.0–3.0 Å data, but as the model was rough in the case of the dodecamers, using low-resolution data for the translation problem made sense. Usually, angles obtained from *CROSUM* are accurate within four degrees, so those angles are always accurate enough for, say, 25–8 Å translation searches.

The success of the molecular-replacement method will, to a great extent, depend on guessing a good model for the unknown, but the fibre and crystal models studied so far will help in this regard. Usually the rotation function will give a clue as to which model is better than the rest. With the advent of faster computers, these calculations are being performed within very short CPU times. The slower step seems to be the human decision-making process in sorting the correct solution from incorrect ones, usually at the translation stage. We believe that in *MERLOT* we have an integrated computer package that enables the user to solve DNA oligomer structures in an extremely efficient manner.

We thank Dr R. E. Dickerson and co-workers for providing coordinates and original crystal data for our experiments. RC also thanks Dr Dickerson for encouragement and support. This research was supported by NSF Research Grant DMB85-01682 and USPHS GM31299.

### References

- ARNOTT, S. & HUKINS, D. W. L. (1972). *Biochem. Biophys. Res. Commun.* **47**, 1504–1509.
- BROWN, T., HUNTER, W. N., KNEALE, G., ANAND, N. N., RABINOVICH, D. & KENNARD, O. (1986). *Proc. Natl Acad. Sci. USA*, **83**, 2402–2406.
- BROWN, T., KENNARD, O., KNEALE, G. & RABINOVICH, D. (1985). *Nature (London)*, **315**, 604–606.
- CHATTOPADHYAYA, R., IKUTA, S., GRZESKOWIAK, K. & DICKERSON, R. E. (1988). *Nature (London)*, **334**, 175–179.
- CHEVRIER, B., DOCK, A. C., HARTMAN, B., LENG, M., MORAS, D., THUONG, M. T. & WESTHOF, E. (1986). *J. Mol. Biol.* **188**, 707–719.
- CONNER, B. N., TAKANO, T., TANAKA, S., ITAKURA, K. & DICKERSON, R. E. (1982). *Nature (London)*, **295**, 294–299.
- CROWTHER, R. A. (1972). In *The Molecular Replacement Method*, edited by M. G. ROSSMANN, pp. 173–185. New York: Gordon & Breach.

- CROWTHER, R. A. & BLOW, D. M. (1967). *Acta Cryst.* **23**, 544–548.
- DICKERSON, R. E. (1987). In *Unusual DNA Structures*, edited by R. D. WELLS & S. C. HARVEY. New York: Springer-Verlag. In the press.
- DREW, H. R., SAMSON, S. & DICKERSON, R. E. (1982). *Proc. Natl Acad. Sci. USA*, **79**, 4040–4044.
- DREW, H. R., WING, R. M., TAKANO, T., BROKA, C., TANAKA, S., ITAKURA, K. & DICKERSON, R. E. (1981). *Proc. Natl Acad. Sci. USA*, **78**, 2179–2183.
- FITZGERALD, P. M. (1988). *J. Appl. Cryst.* **21**, 273–278.
- FRATINI, A. V., KOPKA, M. L., DREW, H. R. & DICKERSON, R. E. (1982). *J. Biol. Chem.* **257**, 14686–14707.
- HO, P. S., FREDERICK, C. A., QUIGLEY, G. J., VAN DER MAREL, G. A., VAN BOOM, J. H., WANG, A. H. J. & RICH, A. (1985). *EMBO J.* **4**, 3617–3623.
- HUNTER, W. N., BROWN, T., ANAND, N. N. & KENNARD, O. (1986). *Nature (London)*, **320**, 552–555.
- HUNTER, W. N., BROWN, T., KNEALE, G., ANAND, N. N., RABINOVICH, D. & KENNARD, O. (1987). *J. Biol. Chem.* Submitted.
- IKUTA, S., CHATTOPADHYAYA, R., ITO, H., DICKERSON, R. E. & KEARNS, D. (1986). *Biochemistry*, **25**, 4840–4849.
- LATTMAN, E. E. & LOVE, W. E. (1970). *Acta Cryst.* **B26**, 1854–1857.
- MCCALL, M., BROWN, T. & KENNARD, O. (1985). *J. Mol. Biol.* **183**, 385–396.
- QUINTANA, J., CHATTOPADHYAYA, R., CONNER, B. & DICKERSON, R. E. (1988). In preparation.
- RABINOVICH, D., HARAN, T., EISENSTEIN, M. & SHAKKED, Z. (1988). *J. Mol. Biol.* **200**, 151–161.
- RABINOVICH, D. & SHAKKED, Z. (1984). *Acta Cryst.* **A40**, 195–200.
- RAO, S. N., JIH, J. & HARTSUCK, J. A. (1980). *Acta Cryst.* **A36**, 878–884.
- SHAKKED, Z., RABINOVICH, D., CRUSE, W. B. T., EGERT, E., KENNARD, O., SALA, G., SALISBURY, S. A. & VISWAMITRA, M. A. (1981). *Proc. Soc. London Ser. B*, **213**, 479–487.
- SUSSMAN, J. L., HOLBROOK, S. R., CHURCH, G. M. & KIM, S.-H. (1977). *Acta Cryst.* **A33**, 800–804.
- WANG, A. H. J., GESSNER, R. V., VAN DER MAREL, G. A., VAN BOOM, J. H. & RICH, A. (1985). *Proc. Natl Acad. Sci. USA*, **82**, 3611–3615.
- WANG, A. H. J., HAKOSHIMA, T., VAN DER MAREL, G., VAN BOOM, J. H. & RICH, A. (1984). *Cell*, **37**, 321–331.
- WANG, A. H. J., QUIGLEY, G. J., KOLPAK, F. L., CRAWFORD, J. L., VAN BOOM, J. H., VAN DER MAREL, G. & RICH, A. (1979). *Nature (London)*, **282**, 680–686.
- WARD, K. B., WISHNER, B. C., LATTMAN, E. E. & LOVE, W. E. (1975). *J. Mol. Biol.* **98**, 161–177.

*Acta Cryst.* (1988). **B44**, 657–663

## Local Structures of Orientationally Disordered Crystals. I. Compatibility Matrices

BY DANIEL ANDRÉ AND HENRI SZWARC

*Laboratoire de Chimie Physique des Matériaux Amorphes (Unité Associée au CNRS, n° 1104), Bâtiment 490, Université Paris-Sud, 91405 Orsay, France*

(Received 12 February 1988; accepted 4 July 1988)

### Abstract

In this paper, steric criteria have been used to study the orientational relationships of neighbouring molecules in orientationally disordered crystals whose structures have been described in terms of the site model. Compatibility matrices have been defined: they are square matrices of order  $N \times N$ , where  $N$  is the number of distinguishable molecular orientations per crystalline site; these elements, which are either 0 or 1, represent the compatibility between pairs of molecular orientations located at two given neighbouring sites. These elements are determined by studying the contacts of the van der Waals envelopes of the molecules involved, and different characteristics of these envelopes (shape, radius, centre location) have been considered. For the five compounds (seven phases) that were examined, the data resulting from recent compilations [Nyburg & Faerman (1985). *Acta Cryst.* **B41**, 274–279; Nyburg, Faerman & Prasad (1987). *Acta Cryst.* **B43**, 106–110] yield the highest compatibility ratios. The variation of the compatibility ratio  $\tau$  vs temperature or vs  $\varepsilon$  (the parameter which defines the way van der Waals

envelopes overlap) gives information on the influence of thermal contraction on orientational freedom. Furthermore, the examination of the compatibility matrices, together with an analysis of the shortest intermolecular distances, indicates the nature of the contacts which entail steric hindrance.

### Introduction

A number of molecular crystals can exhibit a plastic phase (Timmermans, 1938) which is characterized by a great orientational freedom of the molecules. These plastic phases can best be described in terms of the site model (often loosely called 'Frenkel model') according to which the molecules reorient between a limited number of potential wells at the bottoms of which they perform damped librations (Brot & Lassier-Govers, 1976). By cooling such a crystal, the reorientational motion may be frozen to yield a glassy crystal (Adachi, Suga & Seki, 1968; Fuchs, Virlet, André & Szwarc, 1985) within which the orientational disorder is retained in the same crystalline structure (André, Ceccaldi & Szwarc, 1984).