NUCPLOT: a program to generate schematic diagrams of protein–nucleic acid interactions

Nicholas M. Luscombe[1], Roman A. Laskowski[2] and Janet M. Thornton[1,2,*]

[1]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK and [2]Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK

Received September 5, 1997; Revised and Accepted October 28, 1997

ABSTRACT

Proteins that bind to DNA are found in all areas of genetic activity within the cell. To help understand how these proteins perform their various functions, it is useful to analyse which residues are involved in binding to the DNA and how they interact with the bases and sugar–phosphate backbone of nucleic acids. Here we describe a program called NUCPLOT which can automatically identify these interactions from the 3D atomic coordinates of the complex from a PDB file and generate a plot that shows all the interactions in a schematic manner. The program produces a PostScript output file representing hydrogen, van der Waals and covalent bonds between the protein and the DNA. The resulting diagram is both clear and simple and allows immediate identification of important interactions within the structure. It also facilitates comparison of binding found in different structures. NUCPLOT is a completely automatic program, which can be used for any protein–DNA complex and will also work for certain protein–RNA structures.

INTRODUCTION

The interaction of proteins with nucleic acids is an integral part of cellular activity occurring in transcription, translation, replication, repair and rearrangement of nucleic acids. As of June 1997, there were 245 protein–nucleic acid complex structures in the Brookhaven Protein Data Bank (1), and the number is growing continually. Inspection of these structures clearly aids us in understanding these processes and it is becoming increasingly important to conduct this analysis in a systematic manner.

A common property of all DNA binding proteins is their ability to recognise and manipulate DNA structures. This is invariably mediated by the interactions found between the two bodies and thus it is important to know which residues on the protein are responsible for recognition, binding and enzymatic activity.

Owing to the large number of atoms in macromolecules and the complex manner of their interaction in 3D, it is often hard to observe and understand the specifics of the interactions between proteins and DNA without detailed inspection on a graphics terminal. These interactions, once analysed, are often represented in schematic diagrams in order to clarify them. This firstly poses the problem of representing three-dimensional information in two dimensions and secondly, because diagrams are often drawn by hand, they are time consuming to produce.

In this paper we describe a program called NUCPLOT which can automatically generate a schematic 2D plot of protein–DNA interactions directly from the 3D coordinates of the complex as found in PDB files. While there appears to be no standard method of depicting such complexes, the NUCPLOT diagrams have been inspired by the figures used by Houbaviy *et al.* (2).

The resulting plot clearly and intuitively displays the interactions in protein–DNA complexes. It allows quick analysis of interactions within these complexes and facilitates the study of these structures. Particularly helpful uses may be in comparing different proteins or the mode of binding of a particular protein to different DNA sequences.

The program will work for any single or double stranded protein–DNA, DNA–ligand and protein–RNA complexes. Limitations, however, exist in the type of structures which can be displayed at present. Internal base-pairing information such as that found in RNA clover leaf type formations is omitted and more complex structures such as three- or four-stranded DNA cannot be represented.

METHODS

Interaction information

The input to NUCPLOT is a file in PDB format. The program identifies which atoms belong to the protein and other ligands and which to the nucleic acid. Protein residues and water molecules interacting with DNA atoms are then identified from a list of hydrogen bonds, van der Waals contacts and covalent bonds. This list may be supplied by the user, but is more conveniently generated automatically by the HBPLUS program (3) which calculates hydrogen bonds and van der Waals contacts for a given PDB file.

HBPLUS identifies hydrogen bonds as follows. All possible hydrogen atom (H) positions are calculated for donor atoms (D) which satisfy specified geometrical criteria with acceptor atoms (A) in the vicinity. The criteria used are: the H-A distance is <2.7 Å, the D-A distance is <3.35 Å, the D-H-A angle is >90° and the H-A-AA angle is >90°, where AA is the atom attached to the

[*]To whom correspondence should be addressed at: Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK. Tel: +44 171 380 7048; Fax: +44 171 380 7193; Email: thornton@biochem.ucl.ac.uk

acceptor. These criteria can be altered, if required, prior to running HBPLUS. NUCPLOT uses the list of H-bonds generated by HBPLUS to plot all H-bonds between the protein and nucleic acid, between water and nucleic acid, and between protein and nucleic acid via a bridging water molecule.

For van der Waals contacts, all atoms within a certain distance of each other are considered to be interacting. The default distance used by NUCPLOT is 3.9 Å, but this may also be altered by the user. van der Waals interactions between a given protein residue and the nucleic acid are only included if the residue is not already involved in H-bonds to the DNA. van der Waals contacts for water are not included.

Covalent interactions between protein and DNA are either computed from the atomic coordinates using a fixed distance cut-off or are taken directly from the CONECT records in the PDB file.

A complete list of the interactions on the plot is output in an ASCII text file by NUCPLOT. This can be edited as required and used as input for generating another plot.

Base pairing information

To generate the schematic plot, NUCPLOT needs to determine which DNA bases are paired with one another and to work out how to lay out the paired DNA strands on the page to give a sensible plot. This is not always a straightforward matter as there is often little to indicate the base pairing in the PDB file. For example, in some cases, the two DNA strands are represented by two different chains (e.g. chains A and B), with the paired bases having identical residue numbers [e.g. 1run (4)]. In other cases, the two strands are represented by a single chain in the PDB file with, say, residues 1–8 paired with residues 9–16, where base 1 is paired with base 16, base 2 with 15, 3 with 14, and so on [e.g. 2stt (5)]. These are the simple cases.

More complicated examples involve several strands, with some overlap between them, as shown below [e.g. 1ber (6)]:

```
   DDDDDDDDDDDDAAAAAAAAAAAAAAA →
←  CCCCCCCCCCCCCCBBBBBBBBBBBB
```

In cases such as this, the numbering of the bases often gives no clue as to which ones are base-paired and indeed may be wholly eccentric. For example, in the PDB file 1cgp (7), the DNA bases are given in chains C and D, with the base pairing and numbering as indicated in Scheme 1.

To work out what the most likely pairing is, NUCPLOT uses the hydrogen-bonding data read in from the HBPLUS program to determine which bases are paired with which and how the various strands making up the DNA helix should be laid out on the page to give a representation of the helix from one end to the other. Where the DNA is distorted, the base-pairing interactions may be lost, which complicates the process. NUCPLOT chases down each chain, assigning its own internal numbers to the bases, giving the same number to both bases of a pair. Whenever a chain has already had some of its bases numbered, NUCPLOT assesses the direction of the numbering and assigns numbers to all unnumbered bases, giving non-integral numbers to any unpaired bases that appear as insertions between sets of paired bases. Once all bases have been numbered in this way, the plotting commences at the end with the lowest internally numbered bases and proceeds towards the other end.

Drawing

Figure 1 shows an example of a NUCPLOT for the Zif268–DNA complex, 1zaa (8), in which the two DNA strands run vertically down the page. The left-hand strand runs from the 5′ end at the top to the 3′ end at the bottom of the page, and the right-hand strand runs in the opposite direction. The chain names, as given in the PDB file, are placed above each strand. The bases are represented by their one-letter code, and the solid lines between the two strands indicate the base-pairing. The DNA backbone is depicted with its sugars drawn as brown pentagons and its phosphates as purple circles. The base numbers from the PBD file are written within the sugar groups. In cases where a sugar is only connected to an oxygen atom, as is usually the case for strand ends, the phosphate symbol is replaced by a red circle.

Residues from the protein, and water molecules, which interact directly with the DNA are placed as close as possible to their site of interaction in a way which avoids overlap. Also shown are residues that interact with the DNA via a bridging water. The text size of interacting groups is adjusted according to the availability of space around them. Smaller fonts are used when there are many interactions in one place and larger fonts when less crowded. Bond lines are then drawn between the interacting entities.

Each interaction shows the atom name, residue name, number, and the chain identifier. They are coloured brown for carbon, blue for nitrogen and red for oxygen. Water molecules are drawn as blue circles and are labelled by their PDB number.

Bond lines are drawn to connect the interacting entities: blue dotted lines for hydrogen bonds, red dotted lines for van der Waals contacts and solid black lines for covalent bonds. The appearance of the plot and which interactions appear on it can be altered by editing a simple parameter file.

EXAMPLES

We present two examples of NUCPLOTs generated from their respective PDB files: a zinc-finger DNA-binding protein and a Cro protein bound to DNA. The examples serve to show how NUCPLOT depicts protein–DNA interactions and we briefly describe the interactions that were identified as important by the authors who solved the structures.

```
D    D    D    D         D    D    D    C    C    C    C    C    C    C         C    C    C
15   14   13   12   ...  5    4    3    33   32   31   30   29   28   27   ...  18   17   16
|    |    |    |         |    |    |    |    |    |    |    |    |    |         |    |    |
16   17   18   19   20   ...  27   28   29   30   31   32   33   3    4    5    ...  14   15
D    D    D    D    D         D    D    D    D    D    D    D    C    C    C         C    C
```

Scheme 1.

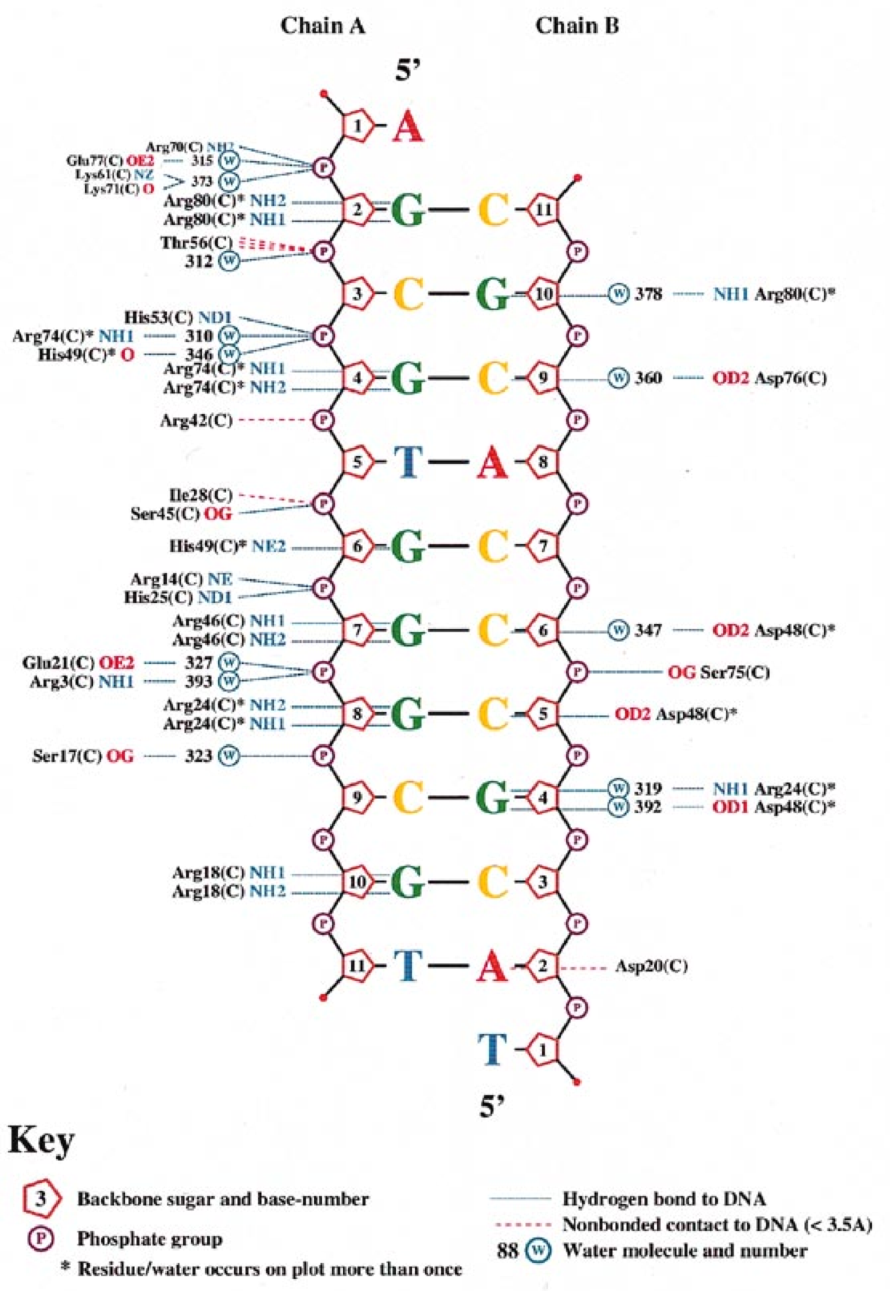

Figure 1. A NUCPLOT diagram of the Zif268–DNA complex (1zaa). Bases are represented by one-letter codes and are coloured according to their type. Base pairs are connected by a solid black line between them. The DNA backbones are drawn next to the bases: the sugars as brown pentagons and phosphates as purple circles. The base numbers, as given in the PDB file, are written inside the sugars. Interactions are plotted on either side of the strands; interacting protein residues are represented by their atom name, residue name, number and the chain identifier in brackets with hydrogen bonds drawn as blue dotted lines and non-bonded contacts as red dotted lines. Atom names are coloured blue for nitrogen and red for oxygen; here, atom names are omitted from residues interacting only by non-bonded contacts. Water molecules are drawn as blue circles and labelled by their PDB number.

Zif268–DNA complex

Structure and function

The first example is shown in Figure 1 and represents a zinc-finger DNA-binding protein, Zif268 (1zaa), solved by Pavletich and Pabo (8) to a resolution of 2.1 Å. The zinc-finger proteins constitute one of the largest structural families of eukaryotic DNA-binding proteins. Unlike the helix–turn–helix motif containing proteins, a lot of structural variation is observed for zinc-finger motifs and at least six subfamilies have been identified. This particular protein has a homologous structure and binding pattern to the fingers found in *Xenopus* transcription factor IIIA.

Figure 2 shows a MOLSCRIPT (9) diagram of the protein–DNA complex. The protein is a single chain consisting of three distinct zinc-fingers that fit into the major groove of B-DNA in a

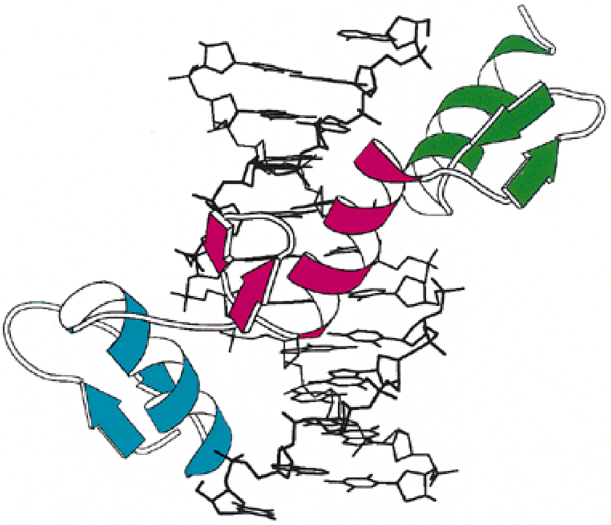

Figure 2. A MOLSCRIPT diagram of the Zif268–DNA complex (1zaa). The protein is a single chain containing three independent zinc fingers coloured cyan, purple and green in the N→C direction. Each finger, consisting of an α-helix and two β-strands coordinating a zinc ion, is bound to a 3 bp subsite in the major groove of the DNA.

semicircular fashion. The zinc-fingers are numbered 1–3 in the order in which they appear on the chain using the conventional N→C direction and are coloured cyan, purple and green, respectively. Equivalent residues in each finger are found 28 positions apart along the peptide. Each finger consists of two antiparallel β-strands and an α-helix. The zinc-fingers get their name from the coordination of a zinc ion by two conserved cysteines in the β strands and two conserved histidines in the α-helix thus forming a stable domain.

Overview of interactions

The protein makes a large number of interactions. Around 25 residues in all are involved in hydrogen bonds to the bases and sugar–phosphate backbone of the DNA, some directly and some via bridging waters. Even when viewing the structure on the 3D graphics terminal it is difficult to see which residues from the protein are interacting with which nucleotides of the DNA. The schematic NUCPLOT in Figure 1 makes these interactions quickly discernible. In particular one can see which residues interact with only the sugar–phosphate backbone and which with the bases and hence are likely to be important for specific recognition of the DNA sequence.

From the plot in Figure 1 it can be seen that most bonds are made to the guanine rich chain A of the DNA. As first suggested by Pavletich and Pabo (8), it is likely that both recognition and stabilisation of binding is mediated through interactions with this strand.

The diagram also shows the periodic binding pattern of the protein to each subsite. Fingers 1–3 are bound to 3 bp subsites as follows: subsite 1 for bases 8–10 (GCG), subsite 2 for 5–7 (TGG) and subsite 3 for bases 2–4 (GCG) on chain A of the DNA. Subsites 1 and 3 have identical sequences.

Backbone interactions

The interactions to the DNA backbone, described by Pavletich and Pabo (8), include two histidine residues (His25 and His53) hydrogen-bonding to the 5′ phosphates of bases 7 and 4 on chain A (Fig. 1). These residues are found in position 7 of the α-helix in both fingers 1 and 2. His81 is also conserved in finger 3 but the plot shows that this residue is not bound to the equivalent phosphate in subsite 3.

Arg42 and Arg70 of fingers 2 and 3 bind the 5′ phosphates of bases 5 and 2, respectively; the corresponding Arg14 from finger 1 binds nucleotide 7 shifted towards the 3′ end of the subsite by 1 bp.

Base interactions

Of the residues that bind directly to the bases, the plot shows a clear binding pattern to the three bases of each subsite (8). Fingers 1 and 3 show common interactions at the 5′ base in their subsite, while only finger 2 makes contact at the central position and all fingers bind to the 3′ base of each subsite.

Looking at the 5′ base interactions, we can see that the two arginine residues, Arg24 and Arg80, each make two hydrogen bonds to the guanine bases (G8 and G2). Inspection of the structure has shown these residues to reside in position 6 of the α-helix (8). A threonine residue is in the equivalent position in finger 2, but the plot shows no interaction by this amino acid.

It is clear from the diagram that only His49 in finger 2 interacts with the central G6 base in the subsite. This residue occupies position 3 of the α-helix.

The 3′ guanine bases in all three subsites (G10, G7 and G4) make conserved interactions to Arg28, Arg46 and Arg74 (8). The bidentate nature of these interactions is identical to that found in the bonds with the 5′ bases. These amino acids precede the α-helix by one residue.

As can be seen from the plot, all base contacts are made between nitrogen group containing amino acids and guanine bases. Comparison of interactions with the three fingers shows that a combination of the correct residue with the right base is needed to make the required bonds.

Phage 434 Cro/OR1 complex

Structure and function

The second example is that of the binding of Cro protein to DNA in PDB file 3cro solved by Mondragon and Harrison (10) to a resolution of 2.5 Å. The Cro protein is part of a regulatory switch in phage 434. The DNA site in the complex, termed OR1, is one of six related binding sites with similar sequences. The underlying factors which control the binding affinities to each site inevitably are the protein–DNA interactions and the site specificity achieved by the protein. It is reported that the central six residues of each binding site modulate the affinities for the protein and therefore the interactions with these particular bases are of great interest.

Figure 3 is a MOLSCRIPT diagram of the Cro–DNA complex. The protein interacts with the binding site as a dimer and each monomer binds to a half site. The protein chains are coloured purple for the L-chain and cyan for the R-chain. The monomers consist of a bundle of five helices with helices 2 and 3 forming a helix–turn–helix motif. Helix 2 of the motif interacts with the DNA backbone on one side of the major groove while the turn and loop

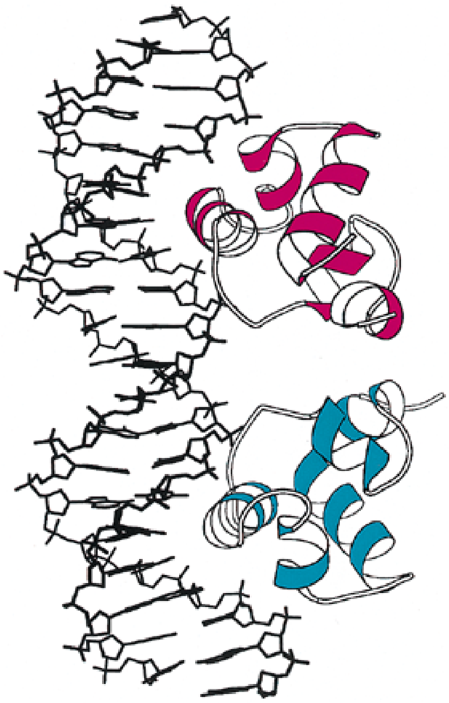

Figure 3. A MOLSCRIPT diagram of the phage 434 Cro/OR1 complex (3cro). The Cro protein is a homodimer, with each monomer consisting of a five helix bundle. Helices 2 and 3 in each chain form the helix–turn–helix DNA binding motif. The L-chain is shown in purple and the R-chain in cyan.

connecting helices 3 and 4 interact with the backbone on the other side of the groove. Helices 4 and 5 act as the dimer interface.

Overview of interactions

A plot of the interactions between chain L of the protein and the DNA is shown in Figure 4. Only half of the OR1 site is shown which is symmetrical about base 12 on chain A. The plot shows two regions of interactions: between bases 13 and 16 on the DNA chain A and between bases 2 and 5 on chain B. These regions constitute the same region on the DNA helix on either side of the major groove.

Backbone interactions

The combination of hydrogen and van der Waals bonds to the sugar–phosphate backbone of the DNA, observed by Mondragon and Harrison (10), can be seen in Figure 4. We can see that on chain A, five sequential residues interact with the 5′ phosphates of T13 and C14. Lys40 and Pro42 make van der Waals contacts while Lys40, Arg43 and Phe44 make hydrogen bonds. The plot shows two other proximal residues, Ser30 and Lys27 are also hydrogen bonded to neighbouring 5′ phosphates of T15 and 16. In fact, these residues all belong to the turn between helices 2 and 3.

The figure highlights that fewer backbone interactions occur on chain B. Gln32 which also interacts with the bases, is bonded to the 5′ phosphate and sugar of A4. Thr16 is also multiply bound to the 3′ phosphate and sugar of A2. The plot shows that this phosphate is also involved in a hydrogen bond with Gln17.

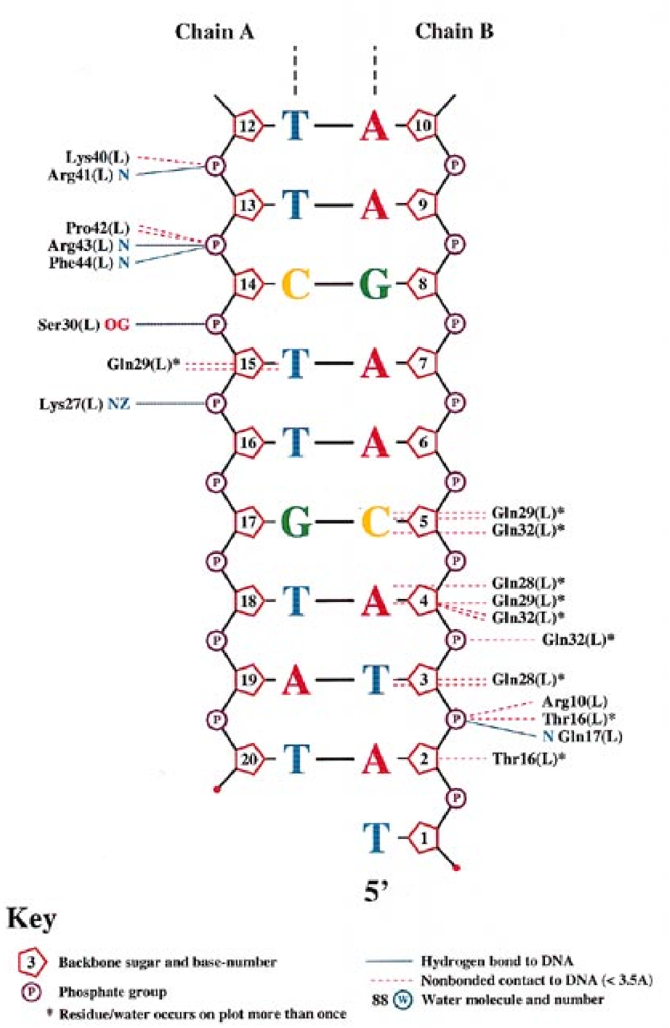

Figure 4. A NUCPLOT diagram of the phage 434 Cro/OR1 complex (3cro) showing the interactions made by the protein's L-chain. The symbols and colouring scheme are the same as those in Figure 1.

Base interactions

The only direct contacts to the DNA bases are through van der Waals interactions from glutamines 28, 29 and 32 to T15 on chain A and T3, A4 and C5 on chain B. Looking at the original structure, one can observe that these interactions are made primarily through helix 3 of the protein and it is unlikely that they are specific for particular DNA sequences. More importantly, no base interactions are found at all in the central bases of the OR1 site (bases 12–14 on chain A), the region thought to confer specificity. The diagram drawn by NUCPLOT depicts the observation made by Mondragon and Harrison (10) that specificity for the OR1 binding site must come from other aspects of binding such as distortion of the DNA.

CONCLUSIONS

The examples have shown some of the potential uses of NUCPLOT in the investigation of protein–nucleic acid complexes. The program automatically produces a schematic diagram of all relevant interactions within such complexes in a clear manner.

The program can be used by both the crystallographer or NMR spectroscopist for analysing a newly solved structure and by other investigators studying protein–nucleic acid structures in general. No standard method of depicting these interactions has existed until now, often making it difficult to compare interactions in different structures. The plots allow the user to instantly identify the interesting interactions and because they are all drawn in the same manner, it is much easier to compare structures. As seen with the examples, combined use of NUCPLOT with a closer look at specific contacts in 3D provides a very powerful method of investigating interactions in protein–DNA complexes.

IMPLEMENTATION AND AVAILABILITY

NUCPLOT is written in C and the source code is available by anonymous ftp on ftp.biochem.ucl.ac.uk after completion of a license agreement. The program is supplied with the source code for HBPLUS, script files for compiling and running under UNIX and full documentation. Enquiries can be made to: nick@biochem.ucl.ac.uk.

The 3D coordinates of the structure to be plotted must be in PDB format. The appearance of the plot may be altered by changing the parameter file using any text editor or word processor. The output is produced in PostScript format (11) in either colour or black and white and may be viewed on a graphics terminal using appropriate software or printed on a Postscript printer.

The program is quick and easy to use. The user only needs to name the 3D coordinates file and plots are produced automatically.

ACKNOWLEDGEMENTS

We would like to thank Helen Berman, Anke Gelbin and Ketan Patel from the Department of Chemistry, Rutgers University, Lawrence Pearl from the Department of Biochemistry, UCL and Lena Nekludova from the Department of Biology, MIT for their help and suggestions. N.M.L. is supported by a BBSRC special studentship.

REFERENCES

1 Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rogers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.

2 Houbaviy,H.B., Usheva,A., Shenk,T. and Burley,S.K. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 13577–13582.

3 McDonald,I.K. and Thornton,J.M. (1994) *J. Mol. Biol.*, **238**, 777–793.

4 Parkinson,G., Gunasekera,A., Vojtechovsky,J., Zhang,X., Kunkel,T.A., Berman,H. and Ebright,R.H. (1996) *Nature Struct. Biol.*, **3**, 837–841.

5 Werner,M.H., Clore,G.M., Fisher,C.L. Fisher,R.J., Trinh,L., Shiloach,J. and Gronenborn,A.M. unpublished results.

6 Parkinson,G., Wilson,C., Gunasekera,A., Ebright,Y.W., Ebright,R.H. and Berman,H.M. (1996) *J. Mol. Biol.*, **260**, 395–408.

7 Schultz,S.C., Shields,G.C. and Steitz,T.A. (1991) *Science*, **253**, 1001–1007.

8 Pavletich,N.P. and Pabo,C.O. (1991) *Science*, **252**, 809–817.

9 Kraulis,P.J. (1991) *J. Appl. Crystallogr.*, **24**, 946–950.

10 Mondragon,A. and Harrison,S.C. (1991) *J. Mol. Biol.*, **219**, 321–334.

11 Adobe Systems Inc. (1985) *PostScript Language Reference Manual.* Wesley Press, Reading, MA.