Acta Cryst. (1998). D**54**, 1132–1138

New Tools and Resources for Analysing Protein Structures and Their Interactions

Nicholas M. Luscombe,[a] Roman A. Laskowski,[b] David R. Westhead,[c] Duncan Milburn,[a] Susan Jones,[a] Maria Karmirantzou[a] and Janet M. Thornton[a,b,c]*

[a]*Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, England,* [b]*Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, England, and* [c]*EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England. E-mail: thornton@biochem.ucl.ac.uk*

(*Received 27 February 1998; accepted 21 May 1998*)

Abstract

The determination of protein structures has furthered our understanding of how various proteins perform their functions. With the large number of structures currently available in the PDB, it is necessary to be able to easily study these proteins in detail. Here new software tools are presented which aim to facilitate this analysis; these include the PDBsum WWW site which provides a summary description of all PDB entries, the programs *TOPS* and *NUCPLOT* to plot schematic diagrams representing protein topology and DNA-binding interactions, SAS a WWW-based sequence-analysis tool incorporating structural data, and WWW servers for the analysis of protein–protein interfaces and analyses of over 300 haem-binding proteins.

1. Introduction

Here we present a number of new software tools and WWW servers developed for the analysis of protein structures and their interactions with other molecules. These tools were developed in the course of our research, involving computational analysis of many structures in the Protein Data Bank (PDB, Bernstein *et al.*, 1977). Many of the ideas have arisen from studies by crystallographers on individual proteins and their complexes, in which analyses and diagrams are usually performed by hand using *ad hoc* programs. When faced with hundreds of structures to analyse, it becomes necessary to develop more robust software which is then of use for studying any new structure. The tools we have developed are, for the most part, freely available to the academic community *via* the WWW (http://www.biochem.ucl.ac.uk/bsm/biocomp/). Additionally they have been used to establish a number of WWW-based resources to provide information on all entries in the PDB. In the descriptions below we use the structure of a protein–DNA complex as an example: namely, PDB entry 1ber which holds the structure of *E. coli* catabolite gene activator protein (CAP) bound to the DNA molecule 31-2E, determined using X-ray crystallography to a resolution of 2.5 Å (Parkinson *et al.*, 1996).

2. PDBsum

We start with PDBsum (Laskowski *et al.*, 1997) which is a WWW-based database of structural analyses of all entries in the PDB. It makes use of, or provides links to, the majority of the software tools described in this review. Each PDB entry has its own WWW page within PDBsum (http://www.biochem.ucl.ac.uk/bsm/pdbsum) giving an at-a-glance summary of what the entry contains: its protein chains and their secondary structure, any DNA/RNA chains, ligands and water molecules. Attached is a wealth of structural analyses of these molecules as well as extensive links to data in other WWW-based databases. The majority of these analyses are automatically generated soon after a new entry is released by the PDB. The entries can be accessed in a number of ways: by their PDB code, by a simple keyword search, *via* the Het groups they contain, by E.C. number, or from other databases including our own CATH database (Orengo *et al.*, 1997) and PDB's 3DB database (Stampf *et al.*, 1995).

Figs. 1 and 2 show extracts from the PDBsum page for 1ber. Fig. 1 gives the header information relating to the structure as a whole, including schematic diagrams of the molecules in the file, icons for viewing the coordinates in three-dimensional using a VRML browser or *RasMol* (Sayle & Milner-White, 1995), names of authors, resolution, *R* factor, and so on. Various links go to other databases including SWISS-PROT (Bairoch & Boeckmann, 1994) and the Nucleic Acid Database (Berman *et al.*, 1992).

Fig. 2 gives a schematic, or 'wiring diagram', of the secondary structure and motifs in the *A* chain of 1ber. The motifs, computed by the *PROMOTIF* program (Hutchinson & Thornton, 1996), include helices, strands, β-turns, γ-turns, and β-hairpins. Also shown are the domain assignments and which residues are in contact with the DNA. For each domain a link goes to the appropriate structural classification in the CATH database (http://www.biochem.ucl.ac.uk/bsm/cath/).

Where a PDB file contains one or more small-molecule ligands the PDBsum entry includes a *LIGPLOT* (Wallace *et al.*, 1995) of the ligand interactions with the

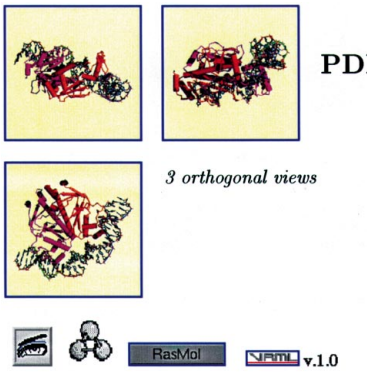
ISSN 0907-4449

protein. For protein–DNA complexes, such as 1ber, a *NUCPLOT* (see below) of the interactions between the protein and DNA is given. Links are provided to *TOPS* topology cartoons and the SAS sequence search and annotation server, both described below.

3. *TOPS*

TOPS is an 'atlas' of protein topology cartoons at http://tops.ebi.ac.uk/tops, described by Westhead *et al.* (1998). The *TOPS* cartoons are schematic diagrams representing the overall topology of a protein chain, or of its constituent domains (Flores *et al.*, 1994). Fig. 3 shows the *TOPS* diagram for the two domains in chain *A* of 1ber. The two cartoons show the protein's helices as circles, its strands as triangles and their connectivity along the chain as lines joining these symbols. This provides a simple representation of the relative directions and positions of the secondary-structural elements within each fold. In the case of 1ber, the first domain is an $\alpha\beta$ domain incorporating several helices and a jelly roll, while the second has a simpler topology consisting of a single β-sheet with three α-helices.

The *TOPS* server holds a topology cartoon to represent every structure in the PDB together with a large amount of information about protein topology in general. Additionally, one can submit a set of coordinates and generate a topology cartoon which can be modified using a Java-based editor.



Complex (gene-regulatory protein/dna)

Title: *Structure of the cap-dna complex at 2.5 Angstroms resolution: a complete picture of the protein-dna interface*
Structure: *Dna (5'-d(Gp Cp Gp Ap Ap Ap Ap Gp Tp Gp Tp Gp Ap C)-3'). Chain: c, e. Engineered: yes. Dna (5'-d(Ap Tp Ap Tp Gp Tp Cp Ap Cp Ap Cp Tp Tp Tp Tp Cp G)- 3'). Chain: d, f. Engineered: yes. Catabolite gene activator. Chain: a, b*
Source: *Synthetic: yes. Synthetic: yes. Escherichia coli. Escherichia coli*

Resolution: 2.50Å. **R-factor:** 0.199. **R-free:** 0.279.

Authors: G.Parkinson, C.Wilson, A.Gunasekera, Y.W.Ebright, R.H.Ebright, H.M.Berman - **Date:** 26-Jan-96

Further information: PDB header (including references), 3DB Browser and coords, NDB entry (ref: PDR023), MMDB entry, CATH and SCOP classification, FSSP structural alignments, PROCHECK summary, PDBREPORT, PROMOTIF analyses.

SWISS-PROT entry: CRP_ECOLI (Chains: **A**, **B**).

Fig. 1. PDBsum header details for the CAP–DNA complex, 1ber. The three thumbnail pictures at the top left show schematic diagrams of the molecules in the complex as seen from three orthogonal viewpoints. Here the two protein chains (*A* in purple and *B* in red) have their helices represented by cylinders and their strands by arrows, while the four DNA chains (*C* to *F*) are shown as stick models with the backbone of each chain traced in a different colour. Interactive versions of these schematic diagrams can be viewed and manipulated using either *RasMol* or a VRML browser, *via* the icons on the line below. The 'Title', 'Structure' and other information are taken from the header records of the PDB file, while below this are a number of links to further analyses such as *PROCHECK* (Laskowski *et al.*, 1993) and *PROMOTIF* as well as to external WWW-based databases including PDB's own 3DB, the Nucleic Acids Database at Rutgers, Entrez's Molecular Modeling Database (Hogue *et al.*, 1996), the structural classifications of proteins in SCOP (Murzin *et al.*, 1995; Hubbard *et al.*, 1997) and CATH, the FSSP database of structurally aligned protein fold families (Holm & Sander, 1994), the PDBREPORT database (Hooft *et al.*, 1996), and SWISS-PROT.

4. *NUCPLOT*

NUCPLOT (Luscombe *et al.*, 1997) is a program written to aid the analysis of protein–nucleic acid complexes. It generates a schematic diagram showing the protein residues that are involved in binding to DNA and RNA and how they interact with the bases and the sugar–phosphate backbone of nucleic acids. The resulting diagrams give a clear and simple representation of the important interactions within the complex.

Fig. 4 shows part of a *NUCPLOT* for 1ber. In this structure, the protein binds as a homodimer to a 30 base-pair site in an approximately symmetrical fashion. Each monomer contains a helix–turn–helix motif which provides both the DNA binding site and the dimer interface. The part of the DNA chain shown in Fig. 4 is a segment of the half site bound to chain *A* of the protein. The protein residues shown on the plot are those that interact with the DNA either *via* hydrogen bonds or through van der Waals contacts. From the diagram it can be seen that amino acids are hydrogen bonded to the DNA backbone between base 3 and 6 on chain *C* and bases 9 and 11 on chain *F*.

The *NUCPLOT* program is available *via* ftp from URL http://www.biochem.ucl.ac.uk/~nick/nucplot.html. *NUCPLOT* diagrams for all protein–DNA complexes in the PDB can be found in the PDBsum database.

5. SAS

In analysing structures and their interactions it is often of value to compare related proteins, especially if the structures of different complexes have been determined (*e.g.* a series of enzyme–inhibitor complexes). Also in analysing sequences (*e.g.* from different species) the structural information may be of importance. To facilitate the use of structural information in sequence

analysis, we have developed a WWW-based tool called SAS – Sequence Annotated by Structure.

This tool annotates the sequences of known structures with structural information at the residue level, derived by programs developed at UCL. The annotations are represented by colouring individual residues in a sequence, according to selected structural properties such as secondary structure, interatomic contacts and active-site information.

The WWW interface (http://www.biochem.ucl.ac.uk/bsm/sas) has several uses. It can be used to annotate a single sequence from the PDB to view its structural features along the length of the sequence. Alternatively, a multiple sequence alignment can be submitted to show, say, the trends and differences in the structural features of a family of related proteins. And finally, and perhaps most usefully, if a sequence of unknown structure is submitted to SAS the sequences in the PDB are scanned and all related sequences are extracted and annotated by their structural features. This can help in identifying distant homologues by showing whether structurally important residues are present in equivalent positions in the query sequence.

Fig. 5 illustrates the SAS output for a target sequence (SWISS-PROT code P51007) which has as its closest match the sequence of 1ber (sequence identity 24.7%). The structural annotation of the 1ber sequence shows its secondary structure and its residues coloured according to the numbers of contacts they make with the DNA. It can be seen that the predicted secondary structure for the target sequence is in good agreement with the actual secondary structure of 1ber, and the region exhibiting the largest numbers of DNA interactions (bottom line of the alignment) also has a high degree of similarity

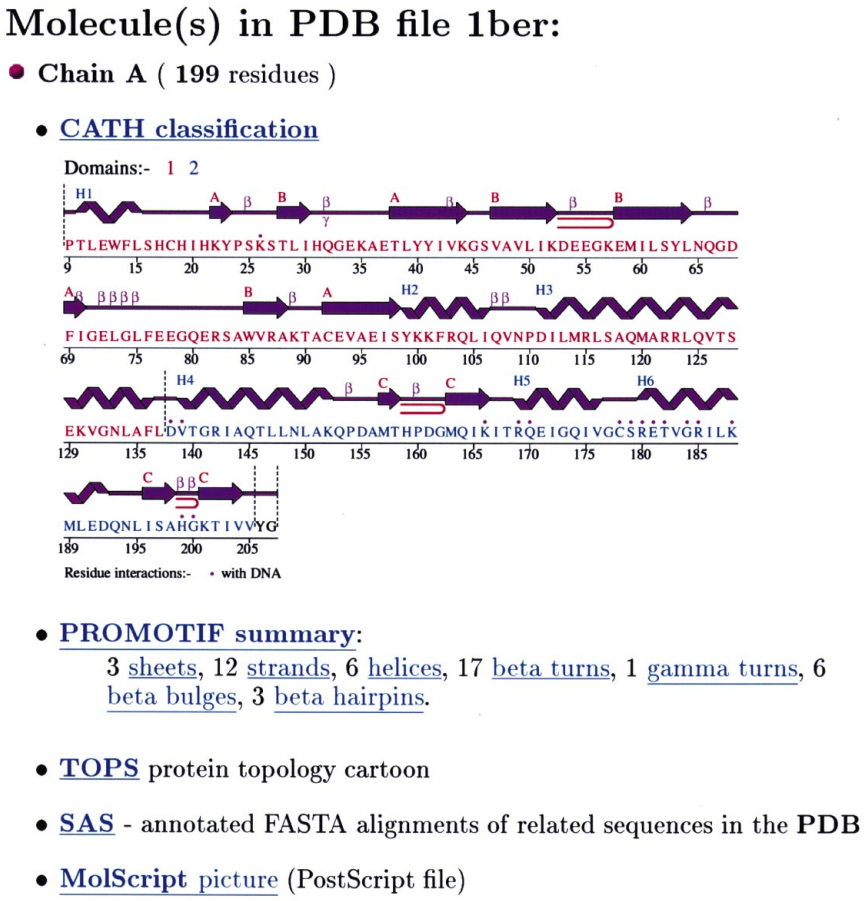


Fig. 2. A schematic plot, or 'wiring diagram', of protein chain *A* in the PDB file 1ber, giving its sequence, secondary structure, domains and motifs. The amino-acid sequence is coloured by domain: red for residues belonging to domain 1, blue for domain 2, and black for fragments belonging to neither. Domains are assigned as described in Jones *et al.* (1998). The CATH classification, based on the domains, can identify which other proteins contain structurally similar domains. The motifs, defined by the *PROMOTIF* analysis, are marked on the diagram; here they include β- and γ-turns, β-hairpins, and helices and strands (labelled *H*1–*H*6 and *A*–*C*, respectively). The small blue dots identify residues interacting with the DNA. Further details on all the annotations are provided by the *PROMOTIF* links below the wiring diagram. The link labelled *TOPS* leads to a protein topology cartoon generated by the *TOPS* program. The SAS link finds similar sequences in the PDB using *FASTA* (Pearson & Lipman, 1988) and annotates the alignment by structural features. Finally, the last of the links generates a *MolScript* (Kraulis, 1991) picture of the chain in question.

between the two sequences, strongly suggesting the target sequence is structurally homologous to 1ber.

6. Protein–protein interactions

Protein–protein interactions are the basis for many biological functions, and a clear description of an interface is an essential starting point to understanding how the complex is formed and perhaps to guide the design of molecules to inhibit complex formation. A new WWW site (http://www.biochem.ucl.ac.uk/bsm/PP/server) has been created for the analysis of protein–protein interaction sites in multimeric structures. The protein interaction server enables the user to submit the three-dimensional coordinates of a protein complex to obtain a set of physical and chemical parameters that characterize the nature of the protein–protein interface.

Fig. 6 shows the information provided by the server for the interface between chains *A* and *B* in the 1ber homodimer. The server calculates data on the size of the protein interface in terms of the lost accessible surface area per chain, the shape (length, breadth and planarity), the intermolecular bonding, polarity, bridging water molecules and packing (Jones & Thornton, 1995). A listing of the residues involved in the protein–

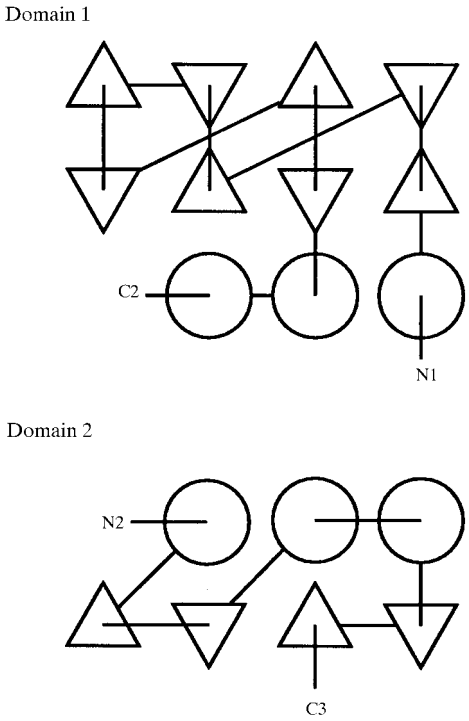


Fig. 3. Topology cartoons of the two domains of chain *A* of 1ber. β-strands are represented by triangular symbols and α-helices by circular ones with the peptide chain following the connecting lines between symbols. In domain 1, the chain runs from N1 to C2, while in domain 2 it runs from N2 to C3. The relative direction of β-strands is shown by the orientation of the triangles. Strands are viewed as having one of two directions: 'up' strands are shown as upward pointing triangles and should be thought of as representing strands directed out of the plane of the diagram; 'down' strands are shown as downward pointing triangles and represent strands directed into the plane of the diagram. The directions of the helices can be deduced by studying how connecting lines are drawn: if the N-terminal connection is drawn to the centre of the symbol and the C terminal one to the edge then the direction is down, otherwise if the N-terminal connection is drawn to the edge and the C-terminal one to the centre and the direction is up.

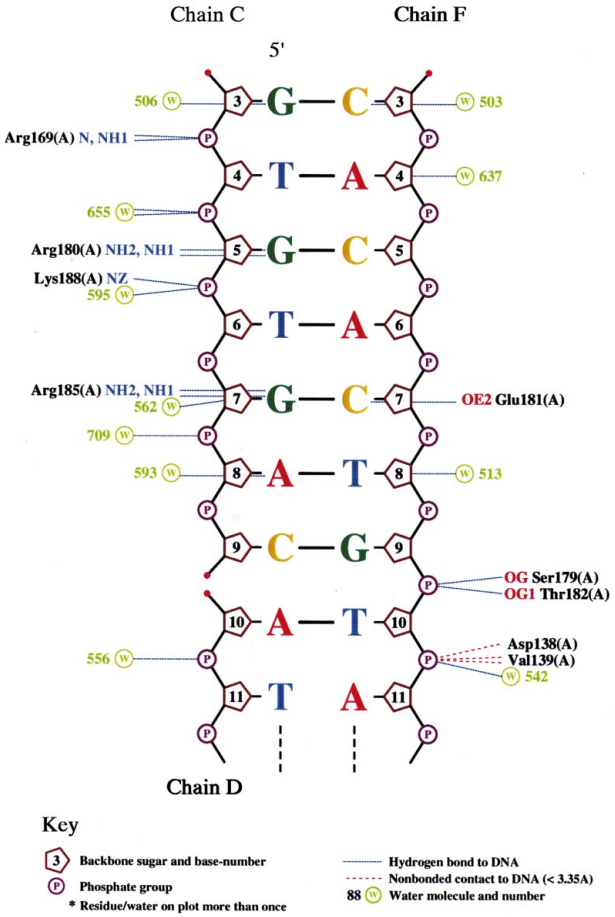


Fig. 4. A *NUCPLOT* of a segment of the DNA in 1ber and the protein residues of chain *A* that interact with it. Bases are represented by one-letter codes and are coloured according to their type. Base pairs are connected by a solid black line between them. The DNA backbones are drawn next to the bases: 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 DNA strands with hydrogen bonds drawn as blue dotted lines and non-bonded contacts as red dotted lines. Interacting protein residues are represented by their atom name, residue name and number and the chain identifier in brackets. The atom names are coloured blue for nitrogen and red for oxygen, although they are omitted from residues interacting only by non-bonded contacts. Water molecules are drawn as light green circles and labelled by their PDB number. A break in the DNA backbone is shown on the left-hand strand between bases 9 and 10. The chain above the break is labelled C and the chain below is D in accordance with the original PDB file.

protein interface (*i.e.* whose accessible surface area decreases by >1 Å^2 on complex formation) is also given, indicating the relative importance of each residue. The parameters for individual structures can be compared to the distributions obtained from data sets of known protein–protein complexes (Jones & Thornton, 1996). Such comparisons allow an estimation of the 'normality' of interfaces in new protein complexes, and may be helpful in distinguishing crystal contacts from those of biological relevance.

7. Protein–haem interactions

In the PDB there are many examples of protein–haem complexes, which have provided the data for a detailed analysis of how proteins recognize and bind this common biomolecule. This analysis has considered many different aspects including the conformation and relative burial of the haem and the nature of the protein interface.

Haem is an aromatic porphyrin molecule acting as a prosthetic group bound to a variety of functionally diverse proteins that have widely differing tertiary structures. In total we analysed 13 non-homologous families, including more than 321 entries in the PDB. The results of this analysis are available on the Internet (http://www.biochem.ucl.ac.uk/bsm/proLig) and as an example Fig. 7 shows the variation in planarity of the haem group in structures representing the 13 different families. The information is useful for

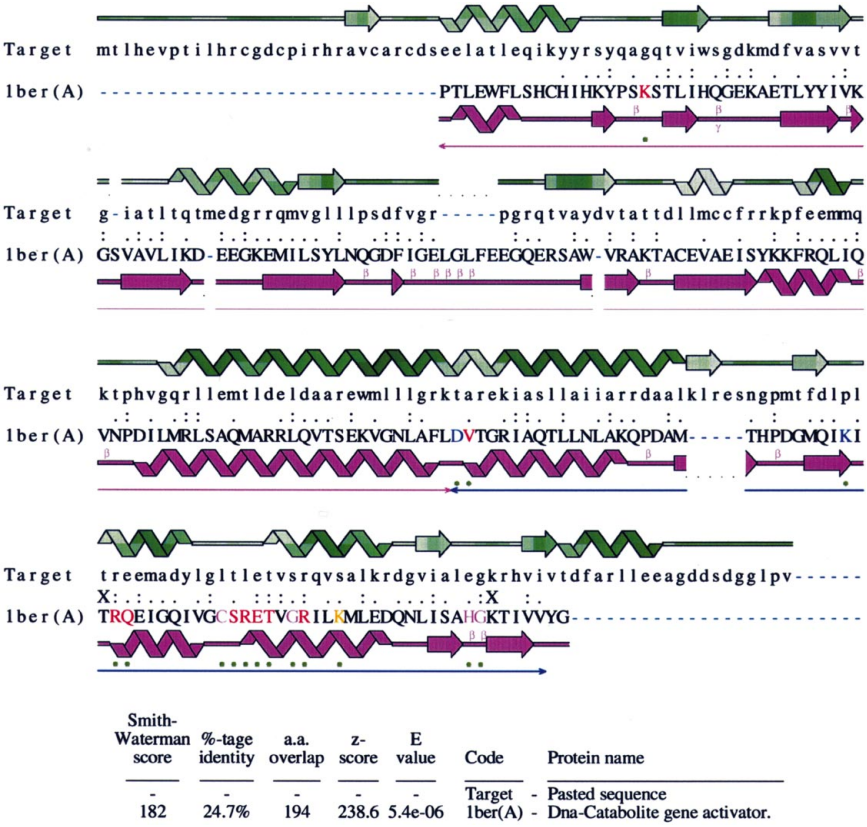


Fig. 5. An annotated SAS alignment of a target sequence (transcriptional activator protein FnrL, SWISS-PROT code P51007) and the closest of its 17 hits from the sequences in the PDB (chain *A* of 1ber). The search and alignment were performed by *FASTA* giving a 24.7% identity over a 194-residue overlap (the other *FASTA* scores for this alignment are summarized in the table at the bottom). The target sequence is shown in lower case while the 1ber sequence is shown in upper case. Residues are coloured according to the numbers of contacts made with the DNA molecule in the complex (black for none, purple for one, blue for two, green for three, orange for four and red for five or more contacts). The colons between the two sequences indicate identical residues, while conservative replacements are denoted by full-stops, with the two Xs signifying the ends of the initial region found by the local similarity search of *FASTA*. The purple wiring diagram shows the secondary structure of 1ber, as calculated by *PROMOTIF*, with green dots again showing residues in contact with the DNA and the thin purple and blue arrows below the wiring diagram denoting the protein's two domains. The green wiring diagram shows the predicted secondary structure for the target sequence as given by the *DSC* program (King & Sternberg, 1996), with the darker green regions corresponding to a higher confidence in the prediction. From the diagram a clear agreement between the target sequence and that of 1ber can be seen in the crucial helix–turn–helix DNA-binding motif in the last row of the alignment, both in terms of the high sequence similarity in the region (particularly of the residues involved in DNA binding) and in the agreement between the predicted and actual secondary structure. This can provide the basis for a better alignment than the simple FASTA alignment shown here and maybe for a three-dimensional model of the target sequence.

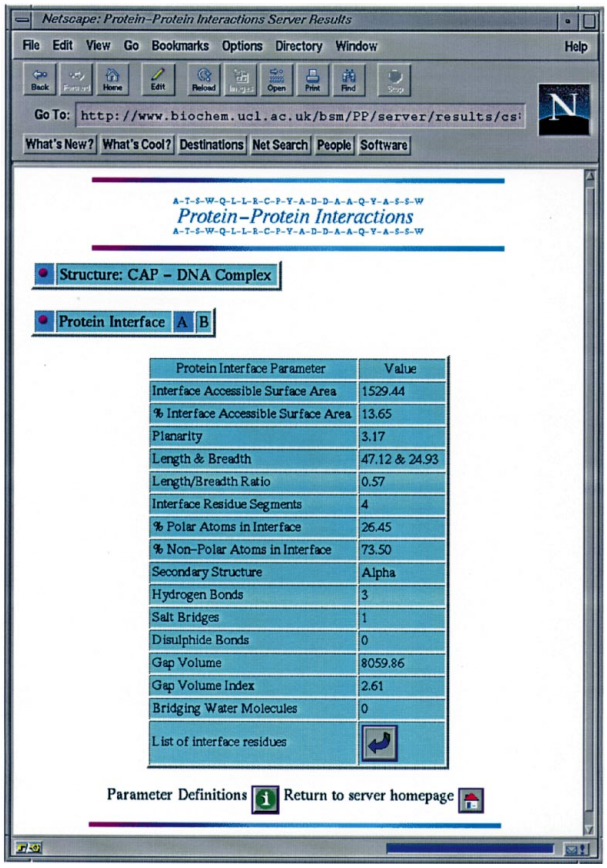


Fig. 6. The interface parameters for chain *A* of 1ber calculated by the Protein–Protein Interaction Server. The table shows that protomer *A* of the CAP dimer buries 1529 Å^2 of accessible surface area in the dimer interface, forms three intermolecular hydrogen bonds with protomer *B*, and that the interface between the two protomers has a gap volume index of 2.6. All these values are close to the means of the interface parameter distributions observed in a non-homologous data set of homodimeric structures.

comparing the structures of haem groups in newly solved structures.

In future we plan to generalize this approach to analyse any protein–ligand complex to enable a comparative analysis of the intermolecular interactions. Such studies can reveal the differences in binding sites for one molecule bound to a variety of different protein families and facilitate prediction of the geometry of such protein–ligand complexes and rules to guide the design of novel binding proteins.

We acknowledge support from the following: NML is supported by a BBSRC special studentship, DM is supported by a BBSRC CASE studentship, sponsored by Roche Discovery Welwyn, MK is supported by a European Union Training and Mobility of Researchers Programme.

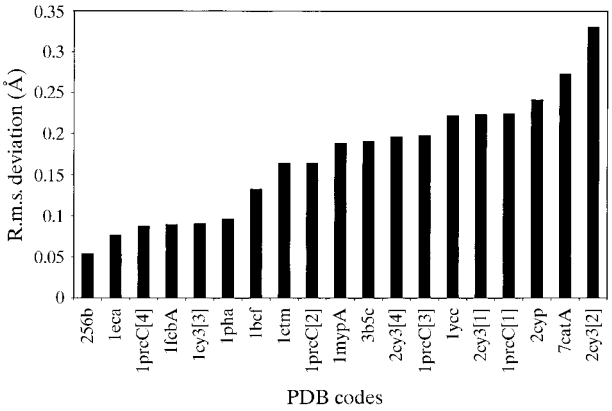


Fig. 7. A plot showing the planarity of bound haem groups in structures representing the 13 different haem-binding protein families. The histogram shows the r.m.s. deviations from a calculated best-fit plane of all atoms in the core of the 19 haem groups present in the 13 representative structures, ordered by increasing r.m.s. Each protein is identified by its four-character PDB code and chain identifier in capitals where relevant. The numbers in brackets after the PDB codes refer to different haem molecules bound to the same chain.

References

Bairoch, A. & Boeckmann, B. (1994). *Nucleic Acids Res.* **22**, 3578–3580.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S-H., Srinivasan, A. R. & Schneider, B. (1992). *Biophys. J.* **63**, 751–759.

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

Flores, T. P., Moss, D. S. & Thornton, J. M. (1994). *Protein Eng.* **7**, 31–37.

Hogue, C. W. V., Ohkawa, H. & Bryant, S. H. (1996). *Trends Biochem. Sci.* **21**, 226–229.

Holm, L. & Sander, C. (1994). *Nucleic Acids Res.* **22**, 3600–3609.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.

Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. & Chothia, C. (1997) *Nucleic Acids Res.* **25**, 236–239.

Hutchinson, E. G. & Thornton, J. M. (1996). *Protein Sci.* **5**, 212–220.

Jones, S., Stewart, M., Michie, A. D., Swindells, M. B., Orengo, C. A. & Thornton, J. M. (1998). *Protein Sci.* **7**, 233–242.

Jones, S. & Thornton, J. M. (1995). *Prog. Biophys. Mol. Biol.* **63**, 31–65.

Jones, S. & Thornton, J. M. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

King, R D. & Sternberg, M. J. E. (1996). *Protein Sci.* **5**, 2298–2310.

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

Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L. & Thornton, J. M. (1997). *Trends Biochem. Sci.* **22**, 488–490.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (1997). *Nucleic Acids Res.* **25**, 4940–4945.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Orengo, C. A., Michie, A. D., Jones, S., Swindells, M. B., Jones, D. T. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.

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

Pearson, W. R. & Lipman, D. J. (1988). *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Sayle, R. A. & Milner-White, E. J. (1995). *Trends Biochem. Sci.* **20**, 374–376.

Stampf, D. R., Felder, C. E. & Sussman, J. L. (1995). *Nature (London)*, **374**, 572–574.

Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1995). *Protein Eng.* **8**, 127–134.

Westhead, D. R., Hatton, D. C. & Thornton, J. M. (1998). *Trends Biochem. Sci.* **23**, 35–36.