FEBS Letters 345 (1994) 93–98

FEBS 14060

Hypothesis

A homology-based molecular model of the proline-rich homeodomain protein Prh, from haematopoietic cells

Stephen Neidle[a,*], Graham H. Goodwin[b]

[a]*CRC Biomolecular Structure Unit, The Institute of Cancer Research, Cotswold Road, Sutton, Surrey, SM2 5NG, UK*
[b]*Chester Beatty Laboratories, The Institute of Cancer Research, Fulham Road, London SW3 6JB, UK*

Received 11 March 1994

Abstract

A molecular structural model for the homeodomain of the haematopoietic protein Prh together with its DNA recognition sequence, has been built using the known crystal structure of the MATα2 homeodomain as a starting-point. The modelling procedure used main and side-chain optimisations by means of molecular mechanics/simulated annealing procedures to obtain stereochemically plausible geometries. The resulting structure has a number of specific interactions in both major and minor grooves of the DNA that serve to define the consensus binding sequence for Prh. In particular, the side-chain of glutamine 50 is postulated to be involved in hydrogen bonds to adjacent adenine and cytosine bases within the consensus sequence.

Key words: Prh protein; Homeodomain; Molecular modelling; Protein–DNA recognition

1. Introduction

The homeodomain motif has been found in a large number of gene-regulatory proteins, and is widely distributed in eukaryotic species, ranging from *Drosophila* to *H. sapiens*. Many of these proteins have been identified as transcription factors [1–4]. The 61 amino acid homeodomain motif contains a helix–turn–helix pattern [5], which is largely responsible for the sequence specificity of homeodomains. Crystallographic studies have been reported on two homeodomain–DNA complexes [6,7] from the *engrailed* protein of *Drosophila*, and the yeast MATα2 repressor, together with NMR analyses of the *antennapedia* development regulatory protein from *Drosophila* [8,9]. These have shown that there are a small number of direct major groove protein–DNA interactions involving the 'recognition helix' of each protein that contribute to their sequence specificities. The crystal structures have also indicated that the N-termini of the proteins play a significant role in the overall binding to DNA sequences by interacting in the DNA minor grooves, although due to disorder very few residues have been located in the minor groove. Analysis of homeodomain sequences has suggested that minor groove contacts are universal to homeodomains [1], although there is large variation in N-terminal sequences. There is some evidence that the N-terminal region plays a major role in determining functional specificity in vivo for a homeodomain [9,10].

Homeodomain genes have been found in haematopoietic cells [11,12] where they may play a role in differentiation, cell development [13] and the regulation of haematopoiesis; abnormal homeodomain expression, for example, as a result of chromosomal translocation, can result in a leukaemic state in the cells [14]. A new homeodomain gene has recently been identified in haematopoietic [15], myeloid and liver cells [16] which is highly conserved across vertebrate species, from mouse to humans [17]. The protein encoded by this gene has been termed Prh (proline-rich homeobox), in view of the high proportion of proline residues in the region N-terminal to the homeodomain itself. The Prh homeodomain does not have an arginine at position 5, unlike the *antennapedia* homeodomains [1], although the arginine at position 7 corresponds to that in the MATα2 homeodomain.

We have used the established crystal structure of the MATα2-DNA complex to build a three-dimensional model of the Prh homeodomain and its consensus DNA sequences, using sequence alignment and molecular modelling methods. This model has been used to examine the molecular basis for the DNA sequence preferences shown by Prh [15] for the consensus site 5′-CAATTAAA.

2. Experimental

The amino acid sequence of MATα2 was taken from [1], and directly read from the crystal structure [7], kindly provided by Dr. C. Wolberger. All visualisations and calculations were performed on a Silicon Graphics XS24 Indigo workstation. Molecular mechanics calculations

*Corresponding author.

SSDI 0014-5793(94)00446-3

were carried out with the Hyperchem package [18], using the all-atom AMBER 3.0 force-field [19] and a distance-dependent dielectric constant ($e = 4_{ij}$). Graphics modelling was accomplished with the Midasplus [20] and Hyperchem programs.

Alignment of the MATα2 and Prh sequences was performed manually, taking several of the established invariant residues in the recognition helix as common points [4]. The first residue in the Prh sequence (lysine), has been numbered zero to maintain correspondance with the accepted homeodomain numbering scheme. Generation of the Prh structure from the MATα2 one was also performed manually on the graphics terminal by substitution of side-chains on a residue-by-residue basis. The larger (three) number of loop residues between helices 1 and 2 in MATα2 compared to Prh, were edited out from the structure. The resulting loss of chain connectivity was overcome by systematic manual manipulation of main-chain torsion angles in the residues immediately surrounding the loop, so as to achieve a main-chain bonded distance. A number of the side-chain substitutions along the sequence produced steric clashes; these were relieved by manually making appropriate changes in side-chain orientations. The main and side-chain conformations of the first five residues at the N-terminal end of Prh were generated using the modelling programs, with some main-chain torsion angles initially set in a β sheet conformation and others having to be manually adjusted in order to avoid clashes with the DNA backbone. This procedure was necessitated since the first four residues in the MATα2 crystal structure are disordered, It was necessary to manually adjust side-chain conformations in order to avoid steric clashes – this process was also able to locate plausible attractive non-bonded contacts between these amino acids and the DNA. The DNA sequence, as reported in the MATα2 structure, was modified to be representative of the Prh consensus sequences [15], by means of base substitutions that kept the DNA backbone, sugar and glycosidic angles unchanged from those in the MATα2 structure. The sequence used here was 5′-GCCAATTAAA.

The complete structure was refined in stages by molecular mechanics minimisation. Initial refinement concentrated on the protein alone, then the complete structure was minimised with the terminal base pairs in the DNA kept constrained so as to avoid fraying effects. Refinement was judged to be complete when the RMS gradient was <0.1 kcal·mol^{-1}·Å^{-1}. This procedure was followed by a simulated annealing protocol to optimise side-chain geometries.

3. Results

The alignment of residues between MATα2 and Prh is shown in Table 1. Helix 1 extends approximately from residues 10 to 22, helix 2 from 29 to 40 and helix 3 from 43 to 60. The shortened loop between helices 1 and 2 has two consecutive prolines at the start helix 2, with (ϕ, φ) angles of (−49°, −53°) and (−61°, −40°) for these prolines 28 and 29, respectively. There is excellent conservation of the key hydrophobic residues beyond those in helix 3 that were used in the alignment, with leucine 34, alanine 35 and leucine 38 in helix 2, and valine 45 in helix 3, together with leucine 40 in the turn between these two helices. The characteristic invariant homeodomain residues, viz. tryptophan 48, phenylalanine 49, asparagine 51 and arginine 53, are all within helix 3.

Fig. 1 shows that there are a number of side-chains extending out from the same side of helix 3, that can make contacts with the major groove of the DNA. These contacts are detailed in Table 2 and Fig. 2b. Several of these are to the oxygen atoms of phosphate groups, especially to those in strand 2 of the DNA duplex. The phosphate group of cytosine 24 has contacts with the side chain of three amino acids. The basic terminus of the side-chain of arginine 53 bridges between the phosphate groups of cytosine 23 and cytosine 24. The basic terminus of arginine 31 contacts the phosphate of guanine 22 as well as glutamic acid 42 (between helix 2 and helix 3) where it is in direct hydrogen-bonding contact with one of the two acidic terminal carbonyl oxygen atoms of this residue (2.8 Å separation). The terminal hydroxyl group of tyrosine 25, which is between helix 1 and 2, makes a hydrogen-bond contact with the phosphate group of cytosine 24.

There are several specific contacts with bases in the major groove. The side-chain of asparagine 51 is in hy-

Table 1
Alignment of amino acids in MATα2 and Prh homeodomains. The rectangles indicate the α helices.

```
        1         10        20        30        40        50        60
MATα2 KPYRGHRFTKENVRILESWFAKNIENPYLDTKGLENLMKNTSLSRIQIKNWVSNRRRKEKT
Prh   KRKGGQVRFSEQTIELEKKFETQ---KYLSPERKRLAKLLQLSERQVKTWFQNRRAKWRRLK
                helix 1                 helix 2           helix 3
```

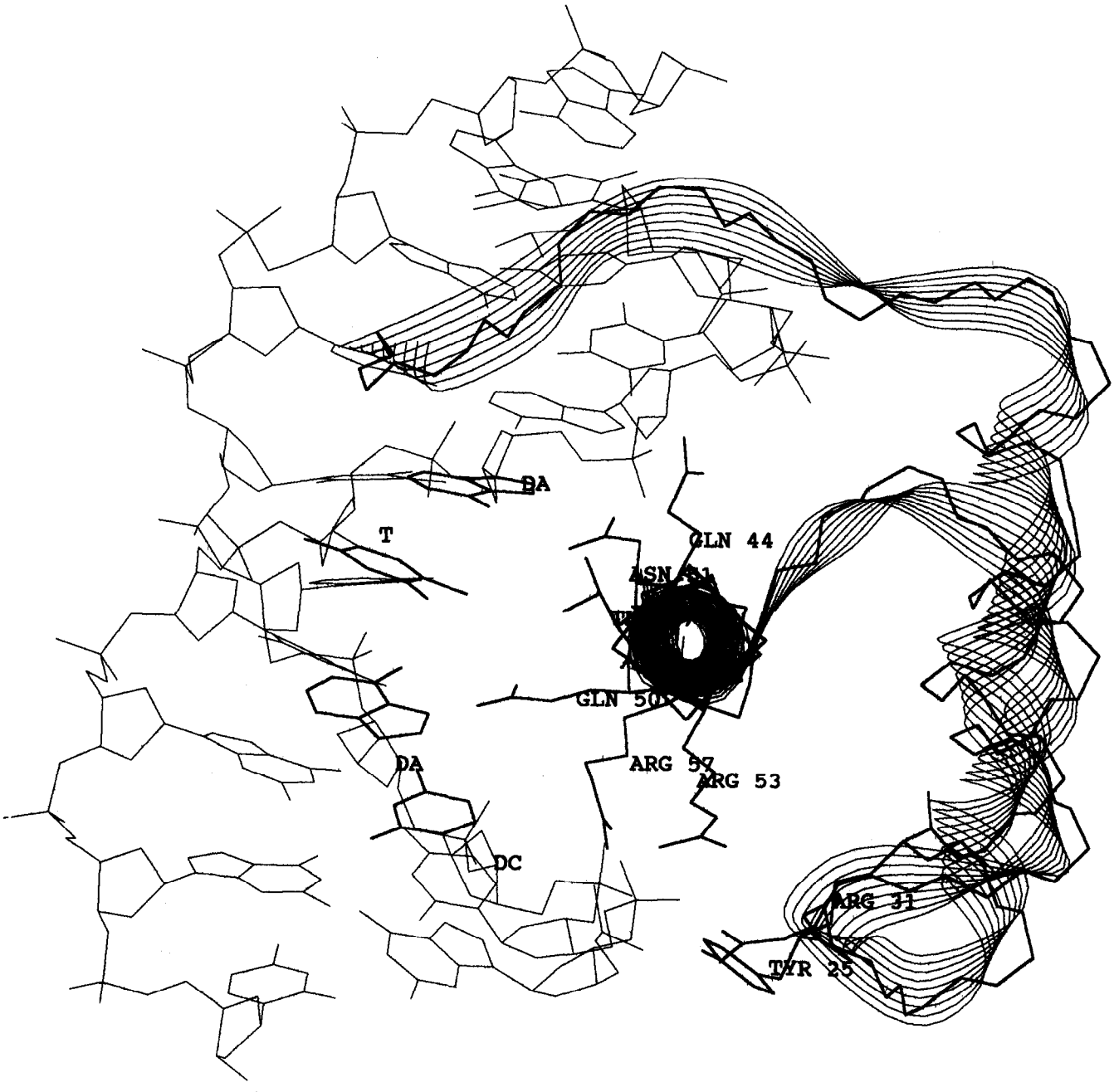

Fig. 1. Overall view of the modelled structure of the Prh–DNA complex, looking down the recognition helix. The α-carbon backbone of the protein is shown in bold, together with the side-chains involved in major groove interactions and the DNA bases to which there is direct hydrogen-bonding.

drogen-bond contact with both atoms N6 and N7 of adenine 16, thereby maintaining a specificity for adenine at this point on the DNA. Glutamine 50 interacts with adenine 25 on strand 1, with the carbonyl oxygen atom at the end of the glutamine side-chain interacting with the N6 hydrogen-bond donor on adenine 25 (Fig. 3). At the same time this carbonyl oxygen atom is in hydrogen-bond contact with the N4 amino group hydrogen-bond donor of cytosine 24 (2.8 Å separation). This pattern of hydrogen bonding thus defines the sequence on strand 1 to be being 5′-cytosine 24, adenine 25. The methyl group of threonine 47 is in close non-bonded van der Waals (3.5 Å) contact with the methyl group of thymine 17, thereby specifying a requirement for thymine at this point in the DNA sequence.

There is an extensive series of contacts between the basic side chains of the N-terminal residues and groups in the DNA minor groove (Table 2 and Fig. 4). For the most part, these are to phosphate groups, although lysine 2 and arginine 7 make direct hydrogen-bond contacts with O2 atoms of the thymine bases 13 and 27 in the groove itself. The main-chain amides of arginine 1

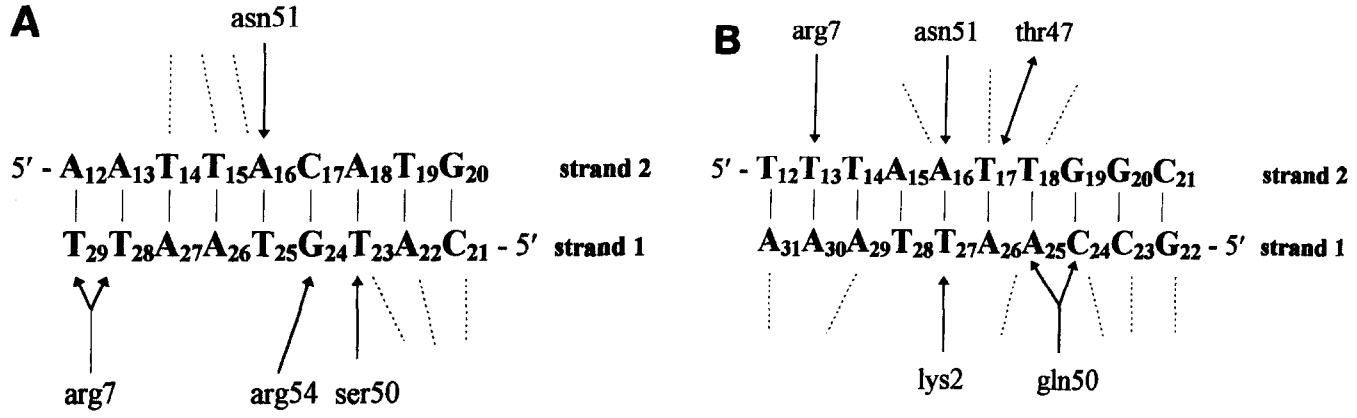

Fig. 2. Schematics of the interactions between (A) MATα2 and its DNA sequence as found in the crystal-structure analysis [7], and (B) the Prh homeodomain and the DNA sequence used in this modelling study. Single-headed arrows indicate hydrogen bonds to the DNA bases, the double-headed arrow shows a hydrophobic contact, and the dashed lines indicate interactions with phosphate groups.

and glycine 4 make contact with a phosphate group and a sugar ring oxygen atom, respectively. The protein backbone has an extended conformation between lysine 2 and valine 6; at each of these residues the backbone is forced to change in direction as a result of the left-handed α-helical conformations that they adopt. In general the backbone is positioned towards the mouth of the minor groove, with only the side-chains of lysine 2 and arginine 7 extending into the groove, towards the bases (Fig. 4b).

4. Discussion

This study has shown that the proline-rich homeobox protein Prh, can be satisfactorily fitted to the experimentally determined tertiary structure of a homeodomain protein with which it shares only relatively low sequence homology (28%). As has been found in the crystal structures of the homeodomain proteins from *engrailed* and MATα2 [6,7], certain key residues provide critical direct base readout contacts in the major groove – in the case of Prh, asparagine 51 hydrogen-bonds to an adenine and threonine 47 is in close hydrophobic contact with the methyl group of a thymine. There are however some differences between Prh and MATα2, as shown in Fig. 2a and b. The latter does not have a hydrophobic residue at position 47, unlike most other homeodomains. The minor groove contacts of arginine 7, are to strand 1 for MATα2, and to strand 2 for Prh – this may be due to a preference for arginine to interact with thymines rather than adenines. In addition, we find for Prh that glutamine 50 plays a key role in DNA sequence recognition, by making a pair of hydrogen bonds to cytosine 24 and adenine 25 (the corresponding amino acid side chains at position 50 in the homeodomain crystal structures are about 1 Å too far from bases for direct contact). The *fushi tarazu* homeodomain similarly has glutamine at

Table 2
Hydrogen bond, van der Waals and close electrostatic interactions between amino acid side chains in the Prh protein, and the DNA sequence used.

	Prh	DNA	Distance (Å)
(i) Major groove interactions			
	Tyr 25	phosphate of C24	2.8
	Arg 31	phosphate of G22	2.7
	Gln 44	phosphate of A16	2.8
	Gln 50	N6 of A25	2.8
		N4 of C24	2.8
	Asn 51	N6 of A16	2.9
		N7 of A16	2.9
	Arg 53	phosphate of C23	2.9
		phosphate of C24	2.7
	Arg 57	phosphate of C24	2.7
(ii) Minor groove interactions			
	Lys 0	phosphate of T18	2.8
	Arg 1 (main-chain amide)	phosphate of T17	2.8
	Arg 1	O3' atom of A29	2.8
	Lys 2	O2 of T27	3.2
	Gly 4	O4' sugar ring atom of A16	3.0
	Gln 5	phosphate of T31	2.9
	Arg 7	O2 of T13	2.9
(iii) van der Waals interactions			
	Thr 47	methyl of T17	3.5

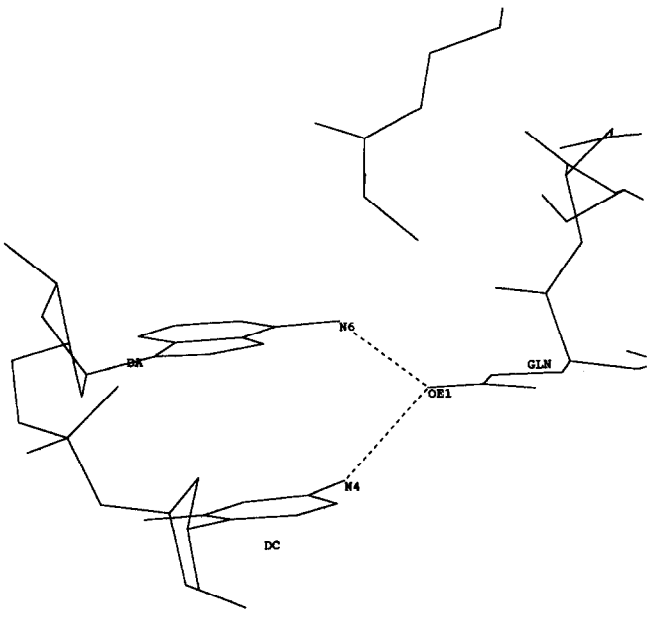

Fig. 3. A detailed view of the three-centre hydrogen-bond interactions between glutamine 50, adenine 25 and cytosine 24.

position 50, which, it has been suggested [21], makes direct contacts with the analogous bases in its consensus sequence. Together, these three amino acids directly specify the sequence 5′-CAAT in the Prh consensus sequence [15]. Recognition by glutamine 50 is not restricted to the CpA dinucleotide sequence, although stereochemically it is clearly the preferred one; others involving equivalent patterns of hydrogen bonding are also possible, and have been found amoung the sequences selected by Prh [15].

The particular conformations found here for the eight N-terminal residues of Prh in the DNA minor groove results in interactions that are directly to two thymines, one of which plays a direct role in maintaining this sequence preference, by providing the preference for a thymine at the 5′ end of the consensus sequence. Although we have not as yet extensively explored all possibilities, there appear to be a number of distinct low-energy conformations available to this region of Prh, of which the model presented here is but one; in the absence of directly comparable crystallographic data the model for the N terminus must therefore remain somewhat conjec-

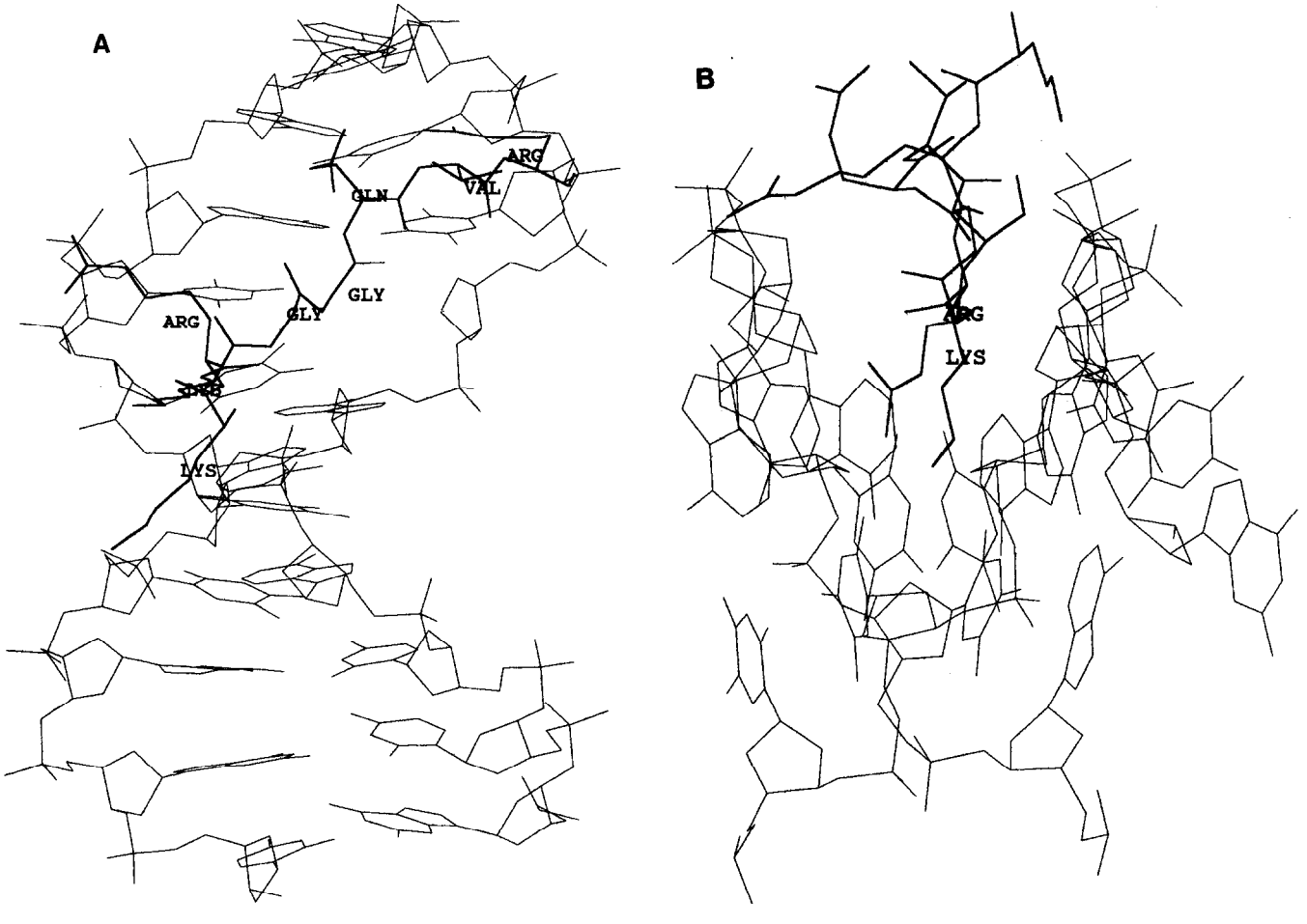

Fig. 4. Views of the eight N-terminal residues of Prh (A) in the minor groove of the DNA sequence, with the DNA helix axis being vertical, and (B) looking down the minor groove, showing the two side-chains (lysine 2 and arginine 7) that directly interact with DNA bases.

tural, although it is clearly consistent with, and rationalises, the consensus sequence data. Overall the fact that the minor groove binding extends over ca. six base pairs in itself provides a preference for an extended narrow-groove AT-rich region since such a sequence would bind the N-terminal basic residues more strongly in both steric and electrostatic terms. The phosphate contacts detailed here for the highly flexible arginine and lysine side chains, are by no means the only ones that can be envisaged; alternative ones would still involve other phosphate groups and bases in this region of the DNA. The extended conformation of part of the N-terminus region, together with the pattern of side-chain contacts to phosphate groups and bases, is analogous to that found in the crystallographic analysis of *Hin* recombinase [22], as well as having some correspondance to the mode of interactions of several minor-groove binding drugs.

The interactions suggested by this detailed molecular model are in large part in accord with our earlier, more tentative predictions of major groove contacts [15]. The present model suggests that the extensive nature of the minor groove interactions (which are only partly observed in the homeodomain crystal structures [6,7]), are a significant component of the overall binding of Prh to its DNA sequence, and perhaps to target selection [5].

Overall, the modelling study suggests that recognition of the TAAT core homeodomain sequence is achieved by a combination of major and minor groove interactions with the recognition helix and the N-terminus of the protein. It has been proposed [23] that the TpA component of this core sequence is recognised by the N-terminus and the ApT by helix 3. In the case of Prh at least, we see that this simple picture is not maintained, with in particular bases of the key base pair A16·T27 being in hydrogen-bond contact with both lysine 2 from the N-terminus and asparagine 51 from helix 3. Although the detailed results of a modelling study must be interpreted with caution, the DNA structure in the Prh complex (Fig. 1) does not appear to be significantly bent, suggesting that the N-terminus of homeodomains may not have a large perturbing effect on overall DNA conformation.

Acknowledgements: This work was funded by grants from the Cancer Research Campaign and the Leukaemia Research Fund.

References

[1] Laughon, A. (1991) Biochemistry 30, 11357–11367.
[2] Hanes, S.D. and Brent, R. (1991) Science 251, 426–430.
[3] Affolter, M., Shier, A. and Gehrung, W.J. (1990) Curr. Opin. Cell Biol. 2, 485–495.
[4] Pabo, C.O. and Sauer, R.T. (1992) Annu. Rev. Biochem. 61, 1053–1095.
[5] Kornberg, T.B. (1993) J. Biol. Chem. 268, 26813–26816.
[6] Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B. and Pabo, C.O. (1990) Cell 63, 579–590.
[7] Wolberger, C., Vershon, A.K., Liu, B., Johnson, A.D. and Pabo C.O. (1991) Cell 67, 517–528.
[8] Otting, G., Qian, Y.Q., Billeter, M., Müller, Affolter, M., Gehring, W.J. and Wüthrich, K. (1990) EMBO J. 9, 3085–3092.
[9] Qian, Y.Q., Billeter, M., Otting, G., Müller, M., Gehring, W.J. and Wüthrich, K. (1989) Cell 59, 573–580.
[10] Lin L. and McGinnis, W. (1992) Genes Dev. 6, 1071–1081.
[11] Magli, M.C., Barba, P., Celetti, A., De Vita, G., Cillo, C. and Boncinelli, E. (1991) Proc. Natl. Acad. Sci. USA 88, 6348–6352.
[12] Kongsuwan, K., Webb, E., Houslaux, P. and Adams, J.M. (1988) EMBO J. 7, 2131–2138.
[13] Allen, J.D., Lints, T., Jenkins, N.A., Copeland, N.G., Strasser, A., Harvey, R.P. and Adams J.M. (1991) Genes Dev. 5, 509–520.
[14] Kennedy, M.A., Gonzalez-Sarmiento, R., Kees, U.R., Lampert, P., Dear N., Boehm, T. and Rabbitts, T.H. (1991) Proc. Natl. Acad. Sci. USA 88, 8900–8904.
[15] Crompton, M.R., Martlett, T.J., MacGregor, A.D., Manfioletti, G., Buratti, E., Giancotti, V. and Goodwin, G.H. (1992) Nucleic Acids Res. 20, 5661–5667.
[16] Hromas, R., Radich, J. And Collins, S. (1993) Biochem. Biophys. Res. Commun. 195, 976–983.
[17] Bedford, F.K., Ashworth, A., Enver, T. and Wiedemann, L.M. (1993) Nucleic Acids Res. 21, 1245–1249.
[18] Hyperchem, a molecular modelling package (Autodesk, Inc.), v. 2.0, 1992.
[19] Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. (1984) J. Am. Chem. Soc. 106, 765–784.
[20] Ferrin, T., Huang, C.C., Jarvis, L.E. and Langridge, R. (1988) J. Mol. Graphics 6, 13–27.
[21] Florence, B., Handrow, R. and Laughon, A. (1991) Mol. Cell. Biol. 11, 3613–3623.
[22] Feng, J.-A., Johnson, R.C. and Dickerson, R.E. (1994) Science 263, 348–355.
[23] Ekker, S.C., von Kessler, D.P. and Beachy, P.A. (1992) EMBO J. 11, 4059–4072.