Vol. 5 No. 1, 2
Winter & Spring 1994

J. Sci. I. R. Iran

A MODEL FOR THE BASIC HELIX-LOOP-HELIX MOTIF AND ITS SEQUENCE SPECIFIC RECOGNITION OF DNA[1]

P. Rashidi-Ranjbar

Department of Chemistry, Tehran University and Chemistry and Chemical Engineering Research Center of Iran, Seied Jamaladin Asadabady Ave., Tehran, Islamic Republic of Iran

Abstract

A three dimensional model of the basic Helix-Loop-Helix motif and its sequence specific recognition of DNA is described. The basic-helix I is modeled as a continuous α-helix because no α-helix breaking residue is found between the basic region and the first helix. When the basic region of the two peptide monomers are aligned in the successive major groove of the cognate DNA, the hydrophobic side chains of the helix I-helix I come in van der Waals proximity. In this way, the end residues of the helix I-helix I are placed far from each other so that a "loop" is needed to bring the helix II-helix II close together for hydrophobic interactions and therefore dimerization. The proposed sequence specific recognition is by hydrogen bonding of the conserved Asn (or Thr) and Glu side chains to the consensus guanine and adenine respectively. The methyl group of Thr and the hydrophobic residue next to it also produce a hydrophobic pocket for recognition of the methyl group of the consensus thymine. The highly conserved Arg's interact with backbone phosphates and a direct recognition of base pairs by Arg's is not likely.

Introduction

Protein-DNA interaction is of central importance in biology and plays a role in gene expression and regulation, DNA recombination and replication, strand scission and other biological processes. An essential part of gene expression and regulation is the binding of a regulatory protein (transcriptional factor) to specific DNA sequences. Analysis of these proteins has indicated that they are modular in nature containing separate motives for DNA binding, dimerization and transcriptional activation [1]. These motives are relatively small and they have been grouped in several kinds of structural classes namely: helix-turn-helix, zinc finger, leucine zipper and basic helix-loop-helix.

The X-ray structures of protein-DNA complexes of helix-turn-helix [2] and zinc finger [3] proteins which have been studied have demonstrated many important structural details of the protein-DNA recognition. Also, the secondary structure of a leucine zipper protein has been studied by 2D-NMR [4], but no structural data has appeared on the basic helix-loop-helix proteins.

Keywords: Helix-Loop-Helix; Protein-DNA recognition; DNA binding; Molecular modeling

[1] This paper was presented at the ESOR III conference in Gothenburg, Sweden, July 15-22, 1991 and the abstract is published in the book of abstracts of the conference.

The basic helix-loop-helix (bHLH) motif is about 60 amino acid residue common to a number of proteins (Figure 1) involved in cell type determination or transcriptional regulation [5, 6]. It contains a basic region amino terminal to another subdomain, tentatively assigned a helix-loop-helix (HLH) secondary structure [5].

By mutational studies, it has been shown that the basic region is the DNA binding and recognition element whereas the HLH domain mediates dimerization and forms homo and heterodimers [5-8]. Some of the bHLH proteins bind to DNA only as heterodimers [9, 10]. Recently, proteins have been identified which contain the HLH domain but lack the basic region and it has been suggested that they may negatively regulate other HLH proteins through the formation of non-functional heterodimers [6, 11, 12].

It has also been shown that replacing the helix I, helix II or the loop of MyoD with the analogous sequence of Drosophila AS-C/T4, both of which contain the bHLH motif, has no substantial effect on DNA binding *in vitro* although replacing the basic region allows the DNA binding but fails to activate the muscle program [7]. Deletion of either helix I, loop or helix II as well as mutation of any residue to Pro in helix I and helix II disrupt the dimerization and therefore DNA binding [7]. The loop region is the most flexible part of the protein and most of the mutations and insertions do not affect the DNA binding or dimerization [7, 13]. Only if α-helix forming residues are inserted while deleting the α-helix breaking ones are no dimers formed [7].

In MyoD, the basic region has three distinct groups of basic residues B1-B3, (Figure 1). Only mutation in the B1 region and insertion between B1-B2 is permitted [7], suggesting that the DNA binding and recognition codes are distributed from the B2 to the B3 region. Substitution of Pro in the basic region disrupts DNA binding while keeping the dimerization [7], suggesting that the basic region has an α-helix structure when bound to DNA. The methylation interference studies indicate that the basic region of the bHLH proteins interacts with the major groove of DNA [14].

Although the bHLH families of proteins have individual sequence specificity for binding which are not totally dyad-symmetric [15, 17], a general cognate, NNCANNTGNN, is common among them. Experimental observations indicate that interactions between heterologous bHLH proteins generate complexes that bind specifically to a commom DNA sequence [8].

Here we describe a model for the bHLH motif and its sequence specific recognition of DNA.

Experimental Section

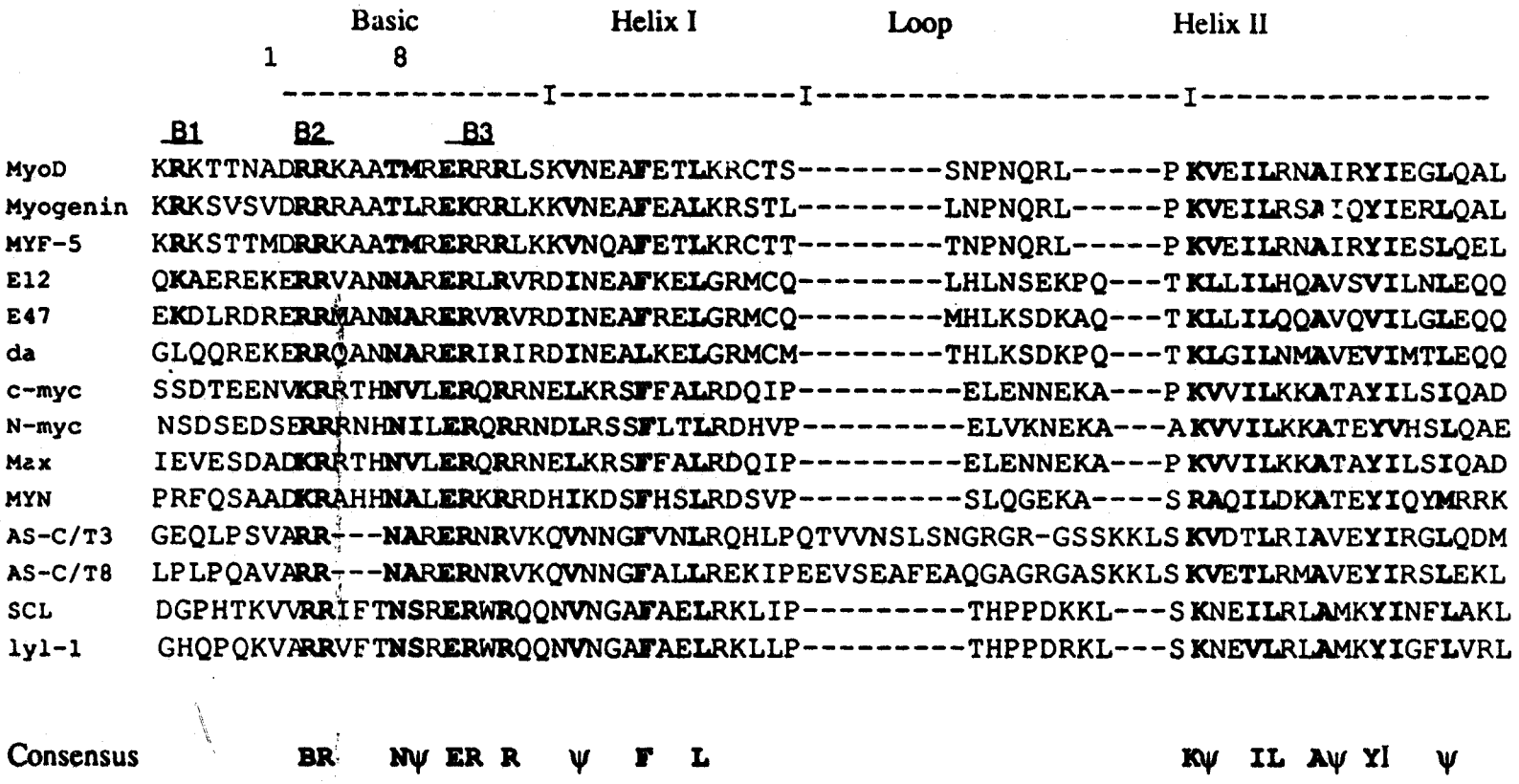
All graphic work was done on a personal *IRIS* [18] using the *Insight II* [19] and *BIOGRAF* [20] programs. The preferred MyoD binding site [15] with dyad-symmetric sequence of GACAGCTGTC (the bold bases refer to the consensus ones) were built in as a right-handed B-DNA [21]. From the other studies it is known that DNA interacts with proteins in the right-handed B conformation [2].

The sequence of the bHLH part of MyoD (Figure 1) was used for model building. The basic region was modeled as an α-helix based on the mutational studies [7] and the similarity between the basic regions of bHLH and leucine zipper motives. The latter has a loose helix structure in solution [4] and forms a defined helix structure when bound to DNA [22]. We predict that the basic region of the bHLH motif behaves in a similar way. It is also known that proteins interact with DNA in the major groove by the topology of an α-helix [2, 3]. An α-helix can readily fit to the major groove of DNA and the amino acid side chains penetrate toward the base pairs and backbone phosphates, (for a proposed β-sheet recognition topology see [23]).

The basic and helix I regions of MyoD were built as a continuous α-helix because no helix breaking residue between the basic region and helix I of bHLH proteins is found. The side chains conformations in model building are those typically found in proteins [24].

The basic region α-helices were docked in the major groove of the cognate DNA so that the side chains of the conserved residues point toward the major groove of the DNA. The conformations of the basic (Arg and Lys), Asn 8 or Thr 8 and Glu 11 side chains in the basic region were changed to improve hydrogen bonding and charge complementarity of protein-DNA residues. The Dreiding forcefield [25] was used for the minimization of protein-DNA interaction energy, while fixing the DNA and α-helix backbone coordinates. The 0.1 kcal/mol change in energy and 0.01 unit change in RMS value were used for energy minimization. The changes in strain energy are not reported as they have no significance in these types of studies.

In this study, we consider the dyad-symmetry of the cognate DNA and the dimeric peptide, and therefore the interaction between one monomer of bHLH motif and half of the binding site will be described. The base pairs within each half site from the 5'-3' direction (GACAG) are given numbers 1-5, while the complementary bases (CTGTC) are given negative numbers (-1, -5). The peptides are numbered from the



Sources for sequences: MyoD (26), Myogenin (27), Myf-5 (28), E12/E47 (5), da (29), c-myc (30) N-myc (31), Max (10), Myn (32), SCL (33), lyl-1 (34), AS-C/T3 & T8 (35).

Figure 1. Relationships between the amino acid sequences of some bHLH proteins. The conserved amino acids are in bold characters. The Ψ indicates hydrophobic residues. The B1 region is not conserved in all proteins and therefore is not included in the bHLH motif. The numbering of the sequences in the bHLH motif is from the residues close to the B2 region and is given at the top.

amino to carboxy terminal direction. The sequences for some of the bHLH proteins are given in Figure 1.

Results and Discussion

Sequence Specific Recognition of DNA by bHLH Motif

The basic region α-helices were aligned in the major groove of the cognate DNA as mentioned before; then we looked for the various possible recognition elements. Conservation of the Asn 8 and Thr 8 suggests that they could be used as a unique recognition element for at least one of the consensus DNA base pairs. Two possibilities were considered, namely hydrogen bonding between the side chains of either Asn 8 or Thr 8 with adenine 4 or guanine -3. In the first case, many of the conserved residues in the basic region were placed out of the major groove of DNA while in the second case all of the conserved residues came in contact with base pairs or backbone phosphates. This required the participation of the hydroxy group of Thr 8, or one of the amide protons of the Asn 8, in a bifurcated hydrogen bond to the guanine -3 (Figure 2). The Thr methyl group in this position has van der Waals interaction with the methyl group of thymine -4 and together with the hydrophobic side chain at position 9 produces a hydrophobic pocket for recognition of the thymine -4 methyl group. The carboxylate side chain of Glu 11 accepts a hydrogen bond from one of the NH2 protons of adenine 4 and acts as the second specific recognition element for the consensus base pairs. In this model, all of the Arg side chains interact with the backbone phosphates and there is less possibility of them acting as a recognition element with respect to any of the base pairs. Interactions of Arg side chains to the backbone phosphates are: Arg 3 (or Lys 3)/phosphate 1, Arg 4/phosphate -3, Arg 12/phosphate -5 and Arg 14/phosphate 3 (Figure 2).

Insertion of Ala between the basic region and the helix I disrupts DNA binding but not dimerization [7]. This was modeled and it was found that the conserved residues in the basic region were placed out of the major groove of DNA when the rest of the protein was kept as dimer. Therefore, the conserved residues in the basic region were not at the right position to interact with the consensus base pairs or backbone phosphates.

The Hydrophobic Interaction and Dimerization of Proteins

Alignment of the basic regions of the two monomeric peptides in the successive major groove of DNA as described above, brought the helix I-helix I close together so that the hydrophobic side chains of the two helices were able to interact with each other upon a very small distortion of the idealized helix structure in the region of residues 14-17. In this way, the highly conserved Phe 22 from the first monomer will sit in a hydrophobic pocket produced by the residues 18 (V, I, L or M), Ala 21 (or CH2 side chain of Ser), Phe 22 and highly conserved Leu 25 of the other monomer (Figure 3). The Leu's 25 are also placed in van der Waals proximity. Mutation of the conserved Phe 22

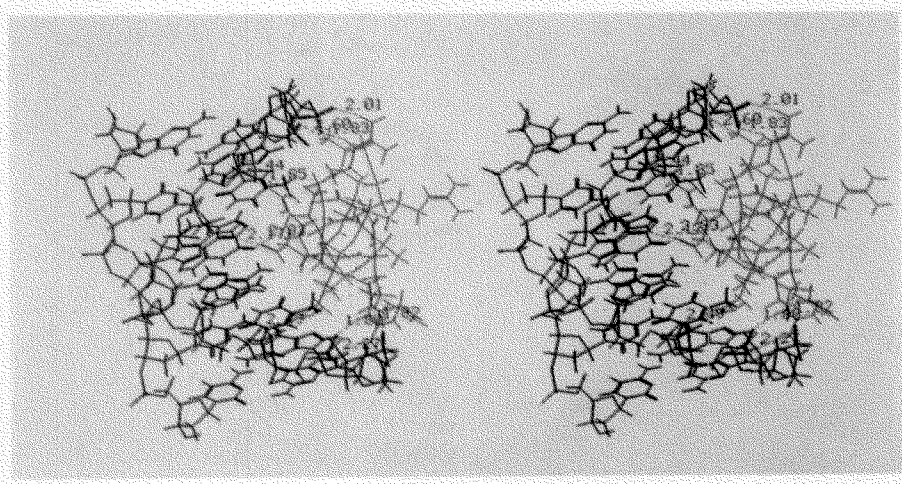


Figure 2. Stereo view of the basic region of MyoD (modeled as an α-helix) docked to the GACAGCTGTC base pairs. The view is from the C-terminus of the basic region after model building and energy minimization. The hydrogen bonds to the consensus DNA base pairs by Thr 8 and Glu 11 and Arg's interactions to phosphates are shown. In MyoD, in addition to the Arg/phosphates interactions mentioned in the text, two more interactions are possible, namely Lys 5/phosphate-4 and Arg 10/phosphate 2.

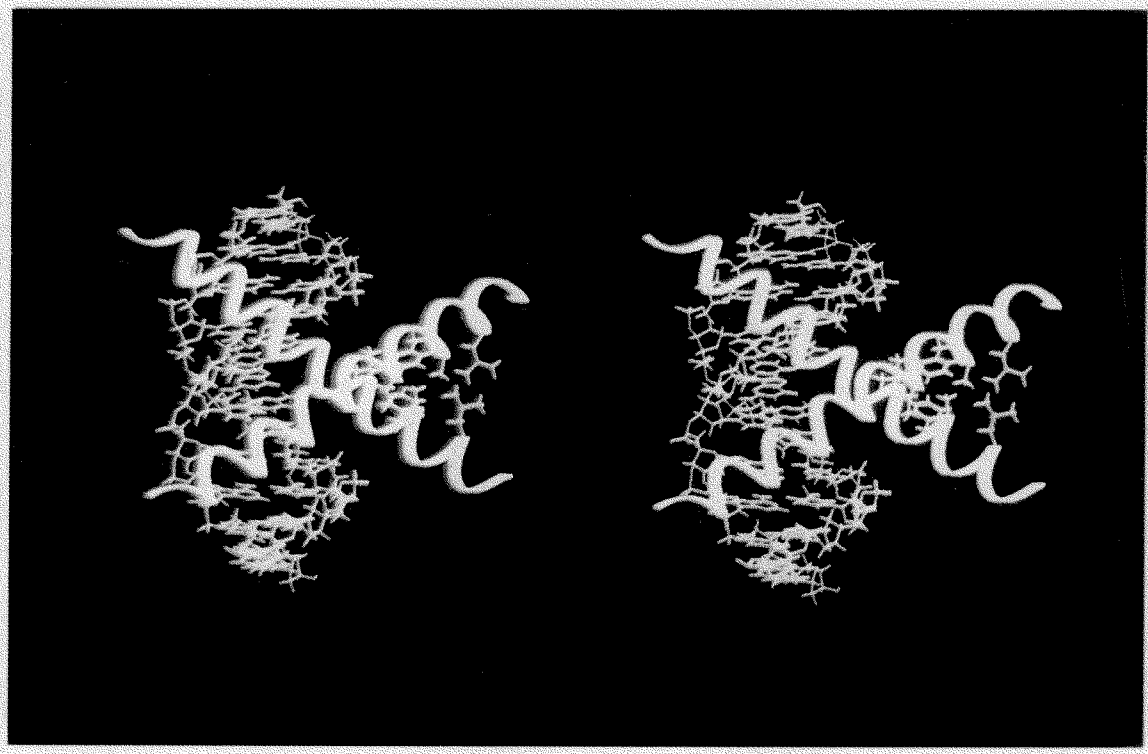


Figure 3. Stereo view of the helix I-helix I hydrophobic side chains interactions. The basic-helix I is modeled as a continuous α-helix and displayed in ribbons. The Phe 22 from the first monomer is located in a hydrophobic pocket produced by the hydrophobic residues of the other monomer. The Leu's 25 side chains from the two monomers are also in van der Waals proximity and contribute to the helix I-helix I hydrophobic interactions.

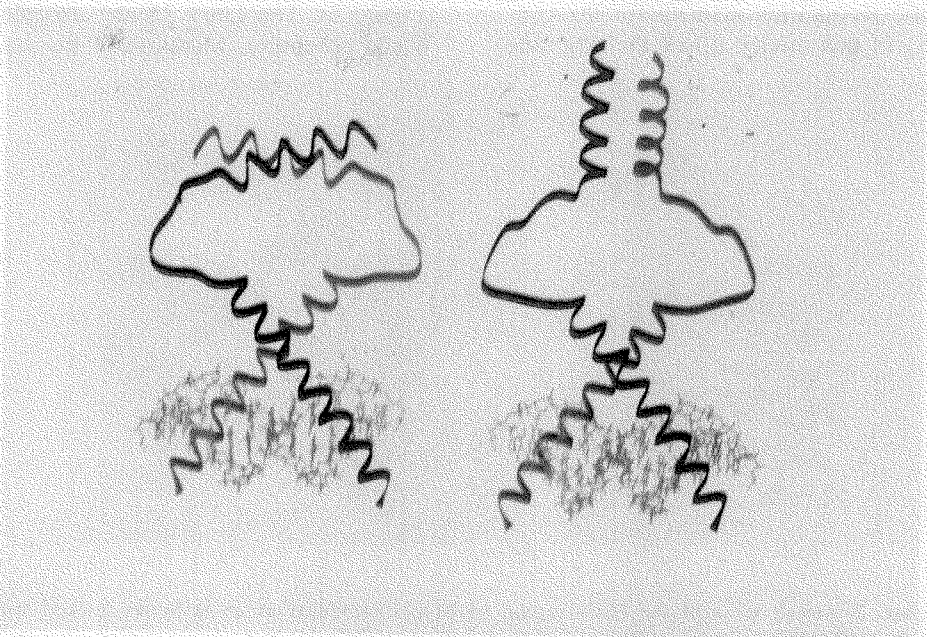


Figure 4. Two possible models for the bHLH motif with the helix II-helix II anti parallel (left) and parallel coiled coil (right) alignments. The conformation of the side chains of the protein dimer was minimized on DNA while keeping the DNA and the protein backbone coordinates constant.

and Leu 25 residues to the Asp and Glu disrupted dimerization [13] indicating that the hydrophobic interactions are crucial for helix I-helix I interactions.

In this model, the helix I-helix I axes are approximately perpendicular to each other and there is no possibility for the formation of a helix I-helix I coiled coil structure [5] or helix I-helix II hydrophobic type of interaction [7].

Perpendicular alignment of helix I-helix I axes puts the end residues of these helices apart from each other so that a "loop" is needed to bring the helix II-helix II in reasonable proximity for dimerization. The dimerization is mediated by the hydrophobic forces as mutation of hydrophobic residues in helix II to hydrophilic ones disrupted dimer formation [13].

Apart from the necessity of a "loop", the second helices could dimerize in two different ways, i. e. parallel (whether coiled coil or not) and antiparallel. The loop size (at least 9 residues) allows for both types of dimerization (Figure 4). In myc subfamily of the bHLH proteins, helix II is part of a leucine zipper, therefore a parallel coiled-coil structure for that protein is expected. In the MyoD and E12/E47 subfamilies, the helix II is not a part of a leucine zipper and may form a different packing. The mutational studies of the loop region [7, 13] and highly conserved Lys at the beginning of helix II might be better explained with an antiparallel alignment of helices II in MyoD and E47. The positive charge on Lys could be stabilized efficiently with the negative end of the dipole of the other helix, therefore explaining why mutation of Lys to Ala disrupted dimerization [13]. A mutational study by deleting residues from the loop region could be performed to test the anti versus parallel alignment of helices II. From modeling, an anti alignment is predicted to be more amenable to dimer disruption when the loop is shortened. Also, mutation of Lys to Leu could demonstrate if the positive charge of Lys is important, for dimer formation of Lys serves as a substituent of Leu.

Acknowledgements

This work was done at the Molecular Graphic Laboratory, Department of Organic Chemistry, University of Gothenburg, S-412 96 Gothenburg, Sweden under a grant given by Professor Per Ahlberg. The author is indebted to Professor Per Ahlberg for this grant and the interest he has shown during the progress of this work.

References

1. Frankel, A. D. and Kim, P. S. *Cell*, **65**, 717, (1991); Mitchell, P. J. and Tjian, R. *Science*, **245**, 371, (1989); Abel, T. and Maniatis, T. *Nature*, **341**, 24, (1989); Busch, S. J. and Sassone-Corsi, P. *Trend Genet*, **6**, 36, (1990); Jones, N. *Cell*, **61**, 9, (1990).
2. Anderson, J. E., Ptashne, M. and Harrison, S. C. *Nature*, **326**, 846, (1987); Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F. and Sigler, P. B. *Nature*, **335**, 321, (1988); Jordan, S. R. and Pabo, C. O. *Science*, **242**, 893, (1988); Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M. and Harrison, S. C. *Science*, **242**, 899, (1988).
3. Pavletich, N. P. and Pabo, C. O. *Science*, **252**, 749, (1991).
4. Saudek, V., Pasley, H. S., Gibson, T., Gausepohl, H., Frank, R. and Pastoré, A. *Biochemistry*, **30**, 1310, (1991).
5. Murre, C. A., McCaw, P. S. and Baltimore, D. *Cell*, **56**, 777, (1989).
6. Benezra, R., Davis, R. L., Lockshon, D., Turner, D. L. and Weintraub, H. *Ibid.*, **61**, 49, (1990).
7. Davis, R. L., Cheng, P. F., Lassar, A. B. and Weintraub, H. *Ibid.*, **60**, 733, (1990).
8. Murre, C., McCaw, P. S., Vaessin, H., Caudy, M., Jan, L. Y., Jan, Y. N., Caberera, C. V., Buskin, J. N., Hauschka, S. D., Lassar, A. B., Weintraub, H. and Baltimore, D. *Ibid.*, **58**, 537, (1989).
9. Sun, X. H. and Baltimore, D. *Ibid.*, **64**, 459, (1991).
10. Blackwood, E. M. and Eisenman, R. N. *Science*, **251**, 1211, (1991).
11. Ellis, H. M., Spann, D. R. and Posakony, J. W. *Cell*, **61**, 27, (1990); Garrell, J. and Modolell, J. *Cell*, **61**, 39, (1990).
12. Christy, B. A., Sanders, L. K., Lau, L. F., Copeland, N. G., Jenkins, N. A. and Nathans, D. *Proc. Natl. Acad. Sci. USA*, **88**, 1815, (1991); Peterson, C. A., Gordon, H., Hall, Z. W., Paterson, B. M. and Blau, H. M. *Cell*, **62**, 493, (1990).
13. Voronova, A. and Baltimore, D. *Proc. Natl. Acad. Sci. USA*, **87**, 4722, (1990).
14. Lassar, A. B., Buskin, J. N., Lockshon, D., Davis, R. L., Apone, S., Hauschka, S. D. and Weintraub, H. *Cell*, **58**, 823, (1989).
15. Blackwell, T. K. and Weintraub, H. *Science*, **250**, 1104 (1990).
16. Blackwell, T. K., Kretzner, L., Blackwood, E. M., Eisenman, R. N. and Weintraub, H. *Ibid.*, **250**, 1149, (1990).
17. Prendergast, G. C. and Ziff, E. B. *Science*, **251**, 186, (1991).
18. Silicon Graphic Personal IRIS 4D35 TG workstation.
19. Insight II. A molecular modeling package available from Biosym Inc.
20. Biograf. A molecular modeling and force field calculations package available from Biodesign Inc.
21. Anott, S. and Hukins, D. W. L. *Biochem. Biophys. Res. Commun.*, **47**, 1504, (1972).
22. Patel, L., Abate, C. and Curran, T. *Nature*, **347**, 572 (1990); Weiss, M. A., Ellenberger, T., Wobbe, C. R., Lee, J. P., Harrison, S. C. and Struhl, K. *Nature*, **347**,

575, (1990).
23. Berg, J. A., Van Opheusden, J. H. J., Burgering, M. J. M., Boelens, R. and Kaptein, R. *Ibid.*, **346**, 586, (1990).
24. McGregor, M. J. *et al.*, *J. Mol. Bio.*, **198**, 295, (1987).
25. Mayo, S. L., Olafson, B. D. and Goddard III, W. A. *J. Phys. Chem.*, **94**, 8897, (1990).
26. Davis, R. L., Weintraub, H. and Lassar, A. B. *Cell*, **51**, 987, (1987).
27. Wright, W. E., Sassoon, D. A. and Lin, V. K. *Ibid.*, **56**, 607, (1989).
28. Braun, T., Buschhausen-Denker, G., Bober, E., Tannich, E. and Arnold, H. H. *EMBO J.*, **8**, 701, (1989).
29. Caudy, M., Vassin, H., Brand, M., Tuma, R., Jan, L. Y. and Jan, Y. N. *Cell*, **55**, 1061, (1988).
30. Bernard, O., Cory, S., Gerondakis, S., Webb, E. and Adams, J. M. *EMBO J.*, **2**, 2375, (1983).
31. Kohl, N. E., Legouy, E., DePinho, R. A., Nisen, P. D., Smith, R. K., Gee, C. E. and Alt, F. W. *Nature*, **319**, 73, (1986).
32. Prendergast, G. C., Lawe, D. and Ziff, E. B. *Cell*, **65**, 395, (1991).
33. Begely, C. G., Aplan, P. D., Denning, S. M., Haynes, B. F., Waldmann, T. A. and Kirsch, I. R. *Proc. Natl. Acad. Sci. USA*, **86**, 10128, (1989).
34. Mellentin, J. D., Smith, S. D. and Cleary, M. L. *Cell*, **58**, 77, (1989).
35. Alonso, M. C. and Cabrera, C. V., *EMBO J.*, **7**, 2585, (1988).