Sequence and evolutionary conservation of the murine *Gbx-2* homeobox gene

Gavin Chapman, Peter D. Rathjen*

Department of Biochemistry, The University of Adelaide, Adelaide 5005, Australia

Received 9 February 1995

Abstract The homeobox gene family is divided into classes based on similarity of sequence across the homeodomain. Representative members of discrete classes are often conserved widely throughout evolution and this can extend to their genomic organisation and biological function. In this paper we report the first complete cDNA sequence of a member of the *GBX* class of homeobox genes, the murine gene *Gbx-2*. Sequence comparisons indicate that this homeodomain class is highly conserved within vertebrates. The homeodomain differs at only three positions out of 60 and these can be used to subdivide the *GBX* class homeodomains into 2 sub-classes.

Key words: Homeodomain; Homeobox gene; Evolution; *GBX* class; *MMoxB*

1. Introduction

Homeobox genes are a developmentally important family of genes first identified as the basis of homeosis in *Drosophila*. The developmental importance of this family as control genes in axial patterning and the formation of different cell types during metazoan development has been inferred through analysis of loss and gain of function mutations in a range of species [1]. Homeobox gene products bind DNA specifically via the homeodomain [2] and have been shown to be transcription factors.

The homeobox encodes a protein motif called the homeodomain which is widespread and probably ubiquitous throughout eukaryotes [3]. The homeodomain is a conserved 60 amino acid DNA binding domain consisting of 3 α-helices with the second and third helix separated by a β-turn. The spacing and length of the α-helices are strictly conserved within the homeodomain. Comparison of homeodomains from many species has revealed that at the primary sequence level there is considerable conservation of specific amino acid residues at defined positions within the homeodomain [3]. This sequence conservation is most striking in helix 3 which contacts the DNA directly [2], and provides a 'trademark' for the identification of homeodomain proteins.

Within the broad consensus, homeodomain sequences can be divided into discrete classes on the basis of additional sequence conservation across the homeodomain and by the presence of regions of homology outside the homeodomain [4]. Individual members within a class are generally more than 75% related at the amino acid level, whereas sequence homology between classes rarely exceeds 55% [4]. The classification of homeodomain sequences into classes has proved useful in determining the evolutionary relationship between different homeobox genes, and in some cases has been shown to reflect equivalent developmental functions and/or chromosomal locations.

The *GBX* class of homeobox genes comprises 2 known members in humans and mice. *Gbx-1* and *Gbx-2* localise to murine chromosomes 5 [5] and 1C5-E1 [6], and are linked genetically to the two murine members of the *Engrailed* homeodomain class *En-2* and *En-1*, respectively [5]. This linked arrangement is also seen with the human homologues of these genes [7,8]. This has been interpreted as evidence for evolution of a homeobox gene cluster by locus duplication [8]. Members of the *GBX* and *EN* classes are specifically expressed in the developing nervous system of the mouse and in the case of *En-1* and *En-2* have been shown to be important for normal brain development [9]. Although the full sequence of the *En-1* and *En-2* genes are known, there has been no report of the full sequence of a member of the *GBX* class.

In this paper we report the full cDNA sequence of the murine homeobox gene *Gbx-2* (previously known as *MMoxB* [10]) and analyse the relationship between this gene and related homeobox genes from mouse, human and *Xenopus*.

2. Materials and methods

2.1. Library screening

A λZAPII library constructed from D3 ES cell cDNA (Clontech; [11]) was screened as described previously [11] with a [^{32}P]dATP oligonucleotide-labelled DNA probe (Gigaprime kit, Bresatec) derived from a partial homeobox fragment of *Gbx-2* isolated by RTPCR [12]. Nytran filters were washed twice in 2 × SSC/0.1% SDS at 42°C for 30 min, and then at 68°C for 65 min. Five positive λ clones were isolated from a total of 8×10^5 plaques screened.

2.2. DNA sequencing

Fragments of λ clone 7.1 were subcloned into pBluescript IIKS and sequenced in both directions and over all restriction sites using the Sanger dideoxy-chain termination procedure [13]. Sequencing was carried out using [^{33}P]dATP (Bresatec) and the Pharmacia T7 sequenase kit. Compressions were resolved with 7-Deaza-dGTP (USB) and dITP (Bresatec) sequencing kits. Sequence comparisons were performed using MacDNAsis software and the FastA database search software.

3. Results and discussion

3.1. cDNA Sequence of Gbx-2

A D3 ES cell cDNA library [11] was screened with a PCR amplified region of the *Gbx-2* homeobox [14]. Sequence analysis revealed that each of the five cDNA clones contained sequence identical to the partial homeobox sequence of *Gbx-2* described by Murtha et al. [10] as *MMoxB*. No *Gbx-1* (*MMoxA* [10]) cDNA clones were isolated despite the extensive sequence homology (84%) across the homeobox fragment used as a

*Corresponding author. Fax: (61) (8) 303-4348.

SSDI 0014-5793(95)00410-6

probe. The longest *Gbx-2* cDNA clone (λ clone 7.1, 2117 bp) was subcloned and sequenced in both orientations.

The sequence and deduced amino acid sequence of the *Gbx-2* cDNA are presented in Fig. 1. The estimated size of the *Gbx-2* transcript by Northern blot is 2.3 kb [12] and is consistent with the 2.12 kb cDNA representing most or all of the *Gbx-2* mRNA. A consensus polyadenylation site (AATAAA) at position 2082 was located 20 bases upstream of the poly(A) tail. This arrangement fits the common positioning of the polyadenylation signal 10–30 bases upstream of the poly(A) tail and indicates that the cDNA is complete at the 3′ end [15].

A 1017 nucleotide open reading frame extended from position 423 to position 1440. The *Gbx-2* protein deduced from this sequence is 339 amino acids long and is predicted to have a molecular weight of 36.4 kDa. The identity of this open reading frame was confirmed by the presence of the *Gbx-2* homeodomain [5] (Fig. 1). Although the position of the ATG initiation codon for translation cannot be established unequivocally, translation appears to commence at 3 successive ATG initiation codons which are the first ATG codons in the open reading frame of the *Gbx-2* homeodomain.

The transcript has an unusually long 5′ untranslated region of 423 bases. 5′ untranslated regions of this length in some cases contain sequences which appear to influence translation efficiency [16].

A proline-rich region (38% proline) extended over 57 amino acids near the amino-terminus of the *Gbx-2* protein (Fig. 1, bold overline). Regions rich in proline residues have been implicated in the activation of transcription by transcription factors such as *AP-2* [17]. This region of the *Gbx-2* protein includes a polymeric repeat of 8 consecutive prolines which is also seen in other homeodomain proteins and may represent a domain of specific importance for activation. This possibility is consistent with the known role of homeodomain proteins as transcription factors [18].

3.2. Comparisons with other homeobox genes

Alignment of *Gbx-2* with 4 related partial length homeodomain sequences is shown in Fig. 2A. *Gbx-2* was most closely related to the human homeobox gene *GBX-2* [8] sharing sequence identity over the 60 amino acids of the homeodomain. *Gbx-2* also showed strong conservation across the homeodomain with other homeobox genes including the human gene *GBX-1* (97% [7]), the chicken gene *CHox7* (97% [19]) and the *Xenopus* gene *XlHox7a* (98% UO4867). The close relationships between these homeodomains indicate that each gene is a member of the *GBX* homeodomain class.

The *GBX* class can be subdivided based on conservative amino acid substitutions at positions 1 and 59 within the *GBX* homeodomain (Fig. 2A). Residue 1 can be either a serine or an asparagine and residue 59 can be either a valine or isoleucine. Thus the mouse *Gbx-2*, human *GBX-2* and *XlHox7a* homeo-

```
                                                                                              cg    2
   3 gttcgcgcgctccccagccgcccgtgggagacgacggactgcaggacggacagacggacggacgcacgcactgctggtagctgggccgagacctgcaga  101
 102 cgcgaagcgcgcgcgttccgacgggctagcagccgggagcggttccggcgaacgcgaggtccccggcgccctctccatcccgggcgcacggcctcatcg  200
 201 cgctcggtggctgcacccgtgcaggccgggtatgctgcgggtcctgcactgagggggcggcggcggcgtctgcgctctgggcaccgctcagcccctgccc  299
 300 ggagttcccggtggtactcccagcgtccggagggcttccaggcggcggcgggacacggggacttttcgtctctccctggcttctcccgaacgcgtctat  398

                                 M  M  M  Q  R  P  L  G  S  S  T  A  F  S  I  D  S  L  I  G  S  P  P  Q  P    25
 399 gagcgcacgttcccgccgtcgctgATGATGATGCAGCGCCCGCTGGGGAGTAGTACCGCCTTCAGCATAGACTCGCTGATCGGCAGCCCGCCGCAGCCC  497

  26 S  P  G  H  F  V  Y  T  G  Y  P  M  F  M  P  Y  R  P  V  V  L  P  P  P  P  P  P  P  P  A  L  P  Q    58
 498 AGTCCCGGCCATTTCGTCTACACCGGCTACCCCATGTTCATGCCCTACCGGCCGGTGGTGCTGCCGCCACCGCCGCCACCGCCTCCCGCGCTGCCCCAG  596

  59 A  A  L  Q  P  A  L  P  P  A  H  P  H  H  Q  I  P  S  L  P  T  G  F  C  S  S  L  A  Q  G  M  A  L    91
 597 GCAGCGCTGCAGCCCGCTCTGCCGCCCGCGCACCCTCACCACCAGATCCCCAGCCTGCCCACCGGCTTCTGCTCCAGCCTGGCGCAGGGCATGGCGCTC  695

  92 T  S  T  L  M  A  T  L  P  G  G  F  S  A  S  P  Q  H  Q  E  A  A  A  A  R  K  F  A  P  Q  P  L  P   124
 696 ACCTCCACGCTCATGGCCACTCTGCCCGGCGGCTTCTCTGCGTCGCCCCAGCACCAAGAGGCGGCGGCTGCCCGCAAGTTCGCTCCACAGCCACTGCCC  794

 125 G  G  G  N  F  D  K  A  E  A  L  Q  A  D  A  E  D  G  K  A  F  L  A  K  E  G  S  L  L  A  F  S  A   157
 795 GGAGGCGGCAACTTCGACAAAGCCGAGGCGCTCCAAGCGGATGCGGAAGACGGCAAAGCCTTCTTGGCCAAGGAGGGCTCGCTGCTCGCTTTCTCTGCG  893

 158 A  E  A  V  Q  A  S  L  V  G  A  V  R  G  Q  G  K  D  E  S  K  V  E  D  D  P  K  G  K  E  E  S  F   190
 894 GCCGAAGCGGTGCAGGCGTCGCTCGTCGGGGCTGTCCGAGGGCAAGGGAAAGACGAGTCAAAGGTGGAAGATGACCCGAAGGGCAAGGAGGAGAGCTTC  992

 191 S  L  E  S  D  V  D  Y  S  S  D  D  N  L  P  G  Q  T  A  H  K  E  E  D  P  G  H  A  L  E  E  T  P   223
 993 TCTCTGGAGAGCGATGTGGATTACAGCTCAGATGACAATTTGCCTGGTCAGACTGCTCATAAGGAAGAAGACCCCGGCCACGCACTGGAGGAGACCCCG 1091

 224 Q  S  G  G  A  A  G  S  T  T  S  T  G  K  N  R  R  R  R  T  A  F  T  S  E  Q  L  L  E  L  E  K  E   256
1092 CAGAGCGGCGGTGCAGCAGGCAGCACCACGTCCACAGGCAAGAACCGGCGGCGGCGGACTGCCTTCACCAGCGAACAGCTGCTGGAGCTGGAGAAAGAA 1190

 257 F  H  C  K  K  Y  L  S  L  T  E  R  S  Q  I  A  H  A  L  K  L  S  E  V  Q  V  K  I  W  F  Q  N  R   289
1191 TTCCACTGCAAAAAGTACCTCTCCCTGACCGAGCGCTCACAGATCGCCCACGCCCTCAAACTCAGCGAGGTGCAAGTAAAAATCTGGTTCCAGAACCGC 1289

 290 R  A  K  W  K  R  V  K  A  G  N  A  N  S  K  T  G  E  P  S  R  N  P  K  I  V  V  P  I  P  V  H  V   322
1290 CGGGCCAAGTGGAAACGTGTCAAGGCAGGCAACGCCAATTCCAAGACGGGGGAGCCCTCTCGGAACCCCAAGATTGTCGTCCCCATCCCTGTTCACGTT 1388

 323 S  R  F  A  I  R  S  Q  H  Q  Q  L  E  Q  A  R  P  *
1389 AGCAGGTTCGCTATTCGAAGTCAACACCAGCAGCTGGAGCAGGCCCGACCCTGAgggcccagccagggtggacaccaccatggagaggccctaacacct 1487
1488 gaaggaacccgtgtcagactccactgtttaaagcagaactcagaggcacctcctagatgtggacatccaagcaaattgagagtatattcattagacggg 1586
1587 cttaaagggggctacagccacaccagaagatacactaaatatttattatacaaccctactttgtacatagatctctatctagactggatctcagctgca 1685
1686 gtgttttgaaagggataggatcaaatctctctccttacactttccagatgaaacaaaaccaccactgtactctgccgccttcagcgaccaccttcccat 1784
1785 acgtgcccaaaggtaaacaggagacacaaggacaaacactgcggtttaaccttcttattgttatttttgtttgtttgtttctattttggcttttgag 1883
1884 ttgagtgcagagcagtcggttgatttggagcaatctagaactctgctctgctcgtttgctttgcctggctgtctttcctggccttgtatatcctatct 1982
1983 gagccatgtcggaagcacgtctccgttgtgtttaatttatttcaatgtagcttattttcatataagttatgacatttaaacaatctcagtcttgtgaat 2081
2082 aataaaaaaaaaggcaaggaaaaaaaaaaaaaaaaa                                                                 2117
```

Fig. 1. Nucleotide and deduced amino acid sequence of *Gbx-2* cDNA clone 7.1. The putative initiation and termination codon are shown in bold, the proline rich region and the homeodomain are indicated by a bold overline and a box, respectively. The polyadenylation signal is underlined.

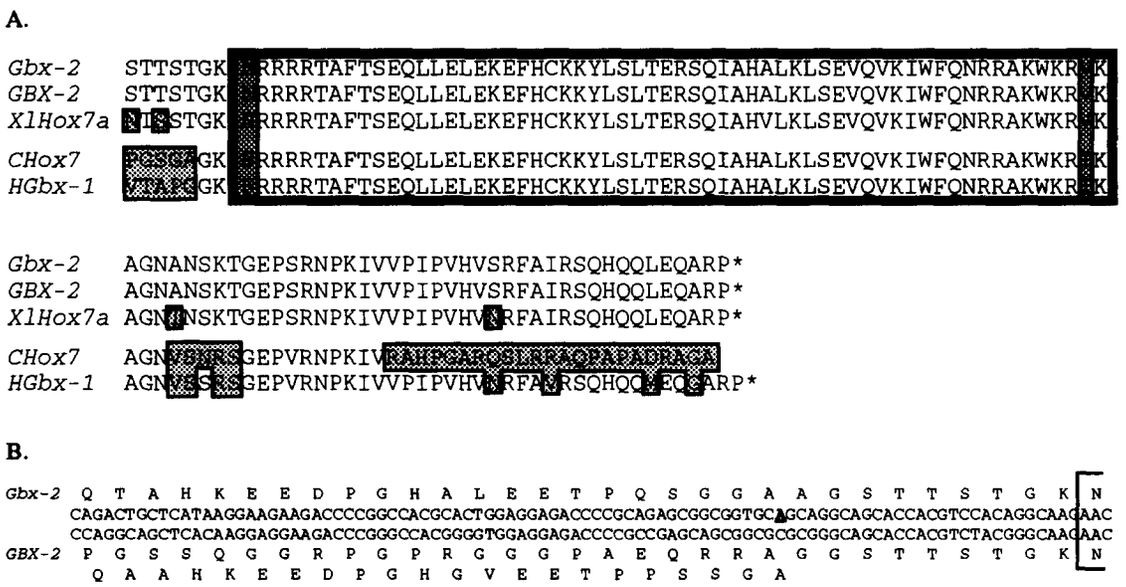

Fig. 2. Protein sequence comparison of *GBX* class homeodomain proteins over the homeodomain and flanking regions. (A) The homeodomain is boxed (bold) and variant amino acids are shaded. (B) DNA and amino acid sequence comparison between mouse *Gbx-2* and human *GBX-2* 5′ of the homeodomain (indicated by the line). The additional A residue in *Gbx-2* is indicated in bold. The alternative reading frame induced in *GBX-2* as a consequence of this frameshift is indicated below the sequence.

domains contain Asn-1 and Val-59, while the *GBX-1* and *CHox7* homeodomains contain Ser-1 and Ile-59. The partial homeodomain sequence for the murine *Gbx-1* gene also contains Ile-59 [5]. These variant positions can therefore be used to identify two subclasses of the *GBX* class. It will be interesting to examine the distribution of these class members in lower vertebrates.

Sequence homologies within the *Gbx-2* subclass extended beyond the homeodomain and were identical beginning 9 amino acids upstream of the homeodomain and ending at the C-terminus, 42 amino acids downstream of the homeodomain. The murine and human proteins were identical over this region although *XlHox7a* contains 5 substitutions, one of which is within the homeodomain (Fig. 2A). This is probably a reflection of the evolutionary distance between *Xenopus* and mammals compared to that separating mouse and human. This suggests that the amino acid sequence downstream and within the homeodomain is crucial to the functionality of *Gbx-2* proteins.

Upstream of this point there is substantial divergence between the mouse and human *GBX* genes. At the nucleotide level this is manifest as a frameshift resulting from the insertion of an adenine residue at position 1113 of the murine *Gbx-2* sequence (Fig. 2B). The presence of this extra residue is confirmed by the fact that in its absence a TAA stop codon at position 1052 of the *Gbx-2* sequence would be in the same frame as the homeodomain. Translation of the homeodomain could therefore not occur. This confirms the presence of this extra base in the murine sequence. Addition of this base to the human *GBX-2* sequence restores homology for the remaining 5′ 22 amino acids to 72% (Fig. 2B).

The expression patterns of *GBX* class members implicates these genes in development of the nervous system [20]. It will be interesting to trace the evolutionary conservation of the *GBX* class, and the different *GBX* class members among lower vertebrates, and to examine the existence, expression pattern and function of the various class members in organisms with more primitive nervous systems.

Acknowledgements: We would like to thank Paul Thomas for the *Gbx-2* homeobox fragment and for helpful advice. Our thanks also to Thomas Schulz for support and technical advice and Timothy Sadlon for helpful comments. This work was supported by grants from the Australian Research Council (ARC).

References

[1] McGinnis, W. and Krumlauf, R. (1992) Cell 68, 283–302.
[2] Gehring, W.J., Qian, Y.Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A.F., Resendez-Perez, D., Affolter, M., Otting, G. and Wuthrich, K. (1994) Cell 78, 211–223.
[3] Scott, M.P., Tamkun, J.W. and Hartzell III, G.W. (1989) Biochim. Biophys. Acta 989, 25–48.
[4] Burglin, T.R. (1994) in: Guidebook to the Homeobox Genes. (Duboule, D. ed.) A Comprehensive Classification of Homeobox Genes, pp. 27–71, Sambrook and Tooze, Oxford University Press, Oxford.
[5] Frohman, M.A., Dickinson, M.E., Hogan, B.L.M. and Martin, G.R. (1993) Mouse Genome 91, 323–325.
[6] Webb, G.C., Chapman, G., Bottema, C.D.K. and Rathjen, P.D. (1995) in preparation.
[7] Matsui, T., Hirai, M., Wakita, M., Hirano, M. and Kurosawa, Y. (1993) FEBS Lett. 322, 181–185.
[8] Matsui, T., Hirai, M., Hirano, M. and Kurosawa, Y. (1993) FEBS Lett. 336, 107–110.
[9] Guidebook to the Homeobox Genes (Duboule, D.) (1994) Descriptions of the Homeobox Genes, pp. 75–246, Sambrook and Tooze, Oxford University Press, Oxford.
[10] Murtha, M.T., Leckman, J.F. and Ruddle, F.H. (1991) Proc. Natl. Acad. Sci. USA 88, 10711–10715.

[11] Thomas, P.Q., Johnson, B.V., Rathjen, J. and Rathjen, P.D. (1995) J. Biol. Chem. (in press).
[12] Thomas, P.Q., Chapman, G. and Rathjen, P.D. (1995) in preparation.
[13] Sanger, F., Nicklen, S. and Coulson, A.R. (1977). Proc. Natl. Acad. Sci. USA 74, 5463–5468.
[14] Thomas, P.Q. (1995) Ph.D. Thesis, University of Adelaide.
[15] Goulding, M., Chalepakis, G., Deutsch, U., Erselius, J., R. and Gruss, P. (1991) EMBO J. 10, 1135–1147.
[16] Kozak, M. (1987) Nucleic Acids Res. 15, 8125–8148.
[17] Williams, T. and Tjian, R. (1991) Genes and Dev. 5, 670–682.
[18] Okamoto, K., Okazawa, H., Okuda, A., Sakai, M., Muramatsu, M. and Hamada, H. (1990) Cell 60, 461–472.
[19] Fainsod, A. and Greunbaum, Y. (1989) FEBS Lett. 250, 381–385.
[20] Bulfone, A., Puelles, L., Porteus, M.H., Frohman, M. A., Martin, G.R. and Rubenstein, J.L.R. (1993) J. Neurosci. 13, 3155–3172.