Nucleic Acids Research, 1997, Vol. 25, No. 21 **4173–4180**

Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals

Thomas R. Bürglin

Department of Cell Biology, Biozentrum, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland

Received September 8, 1997; Revised and Accepted September 24, 1997

DDBJ/EMBL/GenBank accession no. AJ00053

ABSTRACT

A new *Caenorhabditis elegans* homeobox gene, *ceh-25,* is described that belongs to the TALE superclass of atypical homeodomains, which are characterized by three extra residues between helix 1 and helix 2. ORF and PCR analysis revealed a novel type of alternative splicing within the homeobox. The alternative splicing occurs such that two different homeodomains can be generated, which differ in their first 25 amino acids. *ceh-25* is an orthologue of the vertebrate Meis genes and it shares a new conserved domain of 130 amino acids with them. A thorough analysis of all TALE homeobox genes was performed and a new classification is presented. Four TALE classes are identified in animals: PBC, MEIS, TGIF and IRO (Iroquois); two types in fungi: the mating type genes (M-ATYP) and the CUP genes; and two types in plants: KNOX and BEL. The IRO class has a new conserved motif downstream of the homeodomain. For the KNOX class, a conserved domain, the KNOX domain, was defined upstream of the homeodomain. Comparison of the KNOX domain and the MEIS domain shows significant sequence similarity revealing the existence of an archetypal group of homeobox genes that encode two associated conserved domains. Thus TALE homeobox genes were already present in the common ancestor of plants, fungi and animals and represent a branch distinct from the typical homeobox genes.

INTRODUCTION

The group of developmentally important transcription factors encoded by the homeobox genes has been known since 1984 (for reviews see, for example, 1,2). Typical homeobox genes encode the 60 amino acid long homeodomain. The structure of several homeobox genes has been determined by NMR and X-ray crystallography; it consists of three α helices which pack around a hydrophobic core (for review, see 3).

A particular subset of homeobox genes distinguish themselves from typical homeodomains by having more or fewer than 60 amino acids in the homeodomain when the sequences are aligned (4). Structural studies of such genes, i.e., yeast MATα2 (5) and the mammalian transcription factor LFB1 (6,7) have shown that the extra amino acids are accommodated either between helix 1 and helix 2, or helix 2 and helix 3. Several types of atypical homeodomains have been observed (for review see 2,4). One particular group has emerged that has three extra amino acids between helix 1 and helix 2 and has been given the name TALE (three amino acid loop extension; 8). Members of this group are yeast MATα2 (9), maize Knotted-1 (10), the human protooncogene PBX1 (11,12), and the transcription factors TGIF (8) and MEIS1 (13), and the fly Iroquois complex genes (14). A search of the *Caenorhabditis elegans* database ACeDB revealed an EST with weak similarity to *ceh-20*, a PBX1 orthologue. Full sequencing of the cDNA revealed that this gene, *ceh-25*, encodes a homeodomain and is an orthologue of mouse *Meis1*. Given that yeast is completely sequenced and *C.elegans* is sequenced to a large extent, TALE homeobox genes were compiled and analyzed to determine their relationships; this study shows that previous analysis and classifications are incomplete or even incorrect. A new classification and novel highly conserved domains are described as a consequence of the analysis.

MATERIALS AND METHODS

Sequencing and PCR

cm12d8 was subcloned as two fragments into Bluescribe$^+$ using the pRATII polylinker restriction sites and an internal *Bam*HI site. Sequencing was carried out with M13 forward and reverse and sequence-specific primers (Microsynth Co.) using Sequenase (USB) according to the manufacturer's instructions. To test alternative splicing, PCR was performed using *Taq* polymerase (Boehringer) according to instructions on a 1 µl aliquot of an embryonic λgt11 library (generous gift of P.Okkema). 25mer primers from the indicated positions (Fig. 1A) were used at an annealing temperature of 60°C (30 s) and extension temperature of 72°C (1 min) for 35 cycles in the first round. Aliquots of 0.5 µl of the first reaction were used with the nested primers (Fig. 1A) under the same cycling conditions.

Sequence analysis

Blast searches were performed using BLAST at the NCBI (15). For initial sequence extraction and analysis the GCG package (16) was used. Sequences were aligned using MSE (generous gift of W.Gilbert). *Caenorhabditis elegans* sequence searches were performed at http://www.sanger.ac.uk/DataSearch/. Phylo-

*To whom correspondence should be addressed. Tel: +41 61 267 2066; Fax: +41 61 267 2078; Email: burglin@ubaclu.unibas.ch

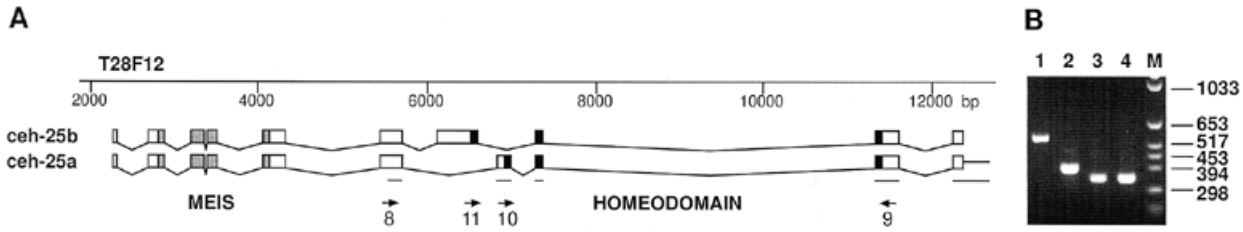


Figure 1. *ceh-25* ORF analysis. (**A**) Schematic representation of the *ceh-25* ORFs. Two different ORFs (a and b) are found, that distinguish themselves in the first exon of the homeodomain. The underlined portion of *ceh-25* marks the extend of the cDNA cm12d8. The ORFs are indicated by boxes, black regions mark the homeodomain, grey regions the MEIS domain. 8, 9, 10 and 11 denote the primer positions used for PCR. (**B**) PCR analysis of the alternative splicing analyzed on a 2% agarose gel. Lane 1: PCR performed with primers 8 and 9 on an aliquot of embryonic cDNA library (30 cycles), expected sizes: 556 and 868 bp; the upper band was not detected. Lane 2: aliquot of the 1. PCR reaction reamplified using primers 10 and 9 (20 cycles), expected band: 396 bp. Lane 3: reamplification of 1. reaction using primers 11 and 9 (20 cycles), expected band: 351 bp. Lane 4: same as 3, but 30 cycles. Restriction digestion of the products of lanes 2 and 4 with *Hpa*I yielded the appropriate sizes (data not shown).

genetic analyses were carried out using the programs ClustalW 1.6 (ftp://ftp.ebi.ac.uk/pub/software/mac/clustalw.sea.hqx) and PHYLIP 3.572 by J.Felsenstein (17) (http://evolution.genetics.washington.edu/phylip.html) on a Macintosh, trees were visualized using TreeView for Macintosh V1.2 by R.D.M.Page (http://taxonomy.zoology.gal.ac.uk/rod/treeview.html), and NJPLOT by M.Gouy (in ClustalW). PUZZLE (18) and PROTML by J.Adachi and M.Hasegawa were used on a SUN SPARCStation5. ORFs of unfinished *C.elegans* cosmid sequences were analyzed using Genefinder within ACeDB (19).

Species codes: c: chicken; Ce: *C.elegans*; d: *Drosophila melanogaster*; Hs: *Homo sapiens*, Mm: *Mus musculus*; Xl: *Xenopus laevis*. Fungi: fCc: *Coprinus cinereus* (inky cap fungus); fUm: *Ustilago maydis* (smut fungus); fSc: *Schizophyllum commune* (bracket fungus); fy: *Saccharomyces cerevisiae*; fSp: *Schizosaccharomyces pombe*. Plants: pAt: *Arabidopsis thaliana* (thale cress); pBn: *Brassica napus* (rape); pGm: *Glycine max* (soybean); pHv: *Hordeum vulgare* (barley); pLe: *Lycopersicon esculentum* (tomato); pOs *Oryza sativa* (rice); pSt: *Solanum tuberosum* (potato); pZm: *Zea mays*.

Accession numbers: c AKR (U25353); *ceh-20* (U01303); d *ara* (*araucan*) (X95179); d *caup* (*caupolican*) (X95178); d *exd* (*extradenticle*, *Dpbx*) (S29960, Z18864, P40427, L19295); fCc β1-1 (β1-1 mating type protein) (X62336); fSc AαZ3 (M97180, M80824); fSc AαZ4 (M97181); fSc AαZ5 (U22049); fSp mat1-Pi (X07643); fUm bE1 (M58553, M30648); fUm bE2 (M58554, M30649); fUm bE3 (M58555, M30650); fUm bE4 (M58556, M30651); fUm bE5 (X54069); fUm bE6 (X54071); fUm bE7 (X54070); fy CUP9 (YPL177c) (L36815, Z73533); fy MATα2 (P01367, L00059); fy YGL096w (Z72168); Hs IRX2a (U90304, U90309); Hs PBX1 (prl) (M86546); Hs PBX2 (G17) (X59842); Hs PBX3 (X59841, P40426); Mm Pbx1 (L27453); Hs TGIF (X89750); Mm mTGIF (X89749); Mm Meis1 (U33629, U33630); Mm Meis2 (U57343); Mm Meis3 (U57344); Mm Mrg1a (Meis1-related protein 1a) (U68383), Mm Mrg1b (C-terminal alternative splice of Mrg1a) (U68384); Hs MRG2 (Meis1-related protein 2) (U68385); pAt ATH1 (X80126); pAt BEL1 (BELL1) (U39944); pAt KNAT1 (U14174); pAt KNAT2 (U14175) same as pAt ATK1 (X81353, X81354); pAt KNAT3 (X92392); pAt KNAT4 (X92393); pAt KNAT5 (X92394); pAt STM (Shootmeristemless) (U32344); pBn hd1 (Z29073, S41980); pGm Sbh1 (L13663); pHv knox3 (Hooded) (X83518); pLe TKn1 (U32247); pOs OSH1 (D16507, JQ2379); pOs OSH45 (D49703, D49704); pSt POTH1 (U65648); pZm Kn-1 (Knotted-1) (X61308); pZm Rs1 (Rough sheath1) (L44133); pAT Z35398 (Z35398); Xl XMeis1-1 (U68386); Xl XMeis1-2 (U68387); The following sequences were taken from (20,21): pZm knox1 (Zmh1); pZm knox10; pZm knox11; pZm knox2 (Zmh2); pZm knox3; pZm knox4 (P11); pZm knox5 (B15); pZm knox6 (R6); pZm knox7 (R7); pZm knox8 (P15); pZm *lg3* (*liguleless3*). *Caenorhabditis elegans* sequences were obtained by ftp from ftp.sanger.ac.uk in /pub/C.elegans_sequences/ (www: http://www.sanger.ac.uk/), and from www: http://genome.wustl.edu/gsc/gschmpg.html. Several partial sequences and most ESTs were not included.

RESULTS

ceh-25 is an orthologue of mouse Meis1

A text search of the *C.elegans* database ACeDB (19) for the keyword 'prl' (original name of *PBX1*) revealed a new cDNA, cm12d8, annotated with a marginal blast score similarity to prl, a homologue of *ceh-20* that was described previously (22). This cDNA was completely sequenced and found to encode a new atypical homeodomain that had not been properly identified due to four separate frameshifts and other errors of the EST sequence within the homeobox. This gene was named *ceh-25* and searches of the databases revealed several mammalian ESTs with high similarity, which were grouped together under a new class name, HAC (2). However, this analysis was incomplete and this group of genes has now been identified as the Meis genes (13,23,24).

An unfinished cosmid sequence (T28F12, Genome Sequencing Center, personal communication) matching *ceh-25* was found in the *C.elegans* genome project. Analysis of the *ceh-25* region by Genefinder revealed that the ORF can be extended at the 5′ end for an additional five exons (Fig. 1A). Furthermore, an internal exon (*ceh-25b*) different from that of the cDNA (*ceh-25a*) was predicted. PCR analysis confirmed that alternative splicing occurs and both exons are used (Fig. 1B). The highly unusual feature of these two exons is that they both encode an N-terminus of the homeodomain. Thus each ORF can produce a protein with a distinct homeodomain that differs in the first 25 residues (Fig. 2). The homeodomain of CEH-25 is 75% identical to that of the vertebrate Meis genes, a value typical for homeobox genes orthologous between vertebrates and nematodes.

Classification of the TALE superclass homeobox genes

To better understand the relationships of the TALE superclass homeobox genes, comprehensive searches of the sequence

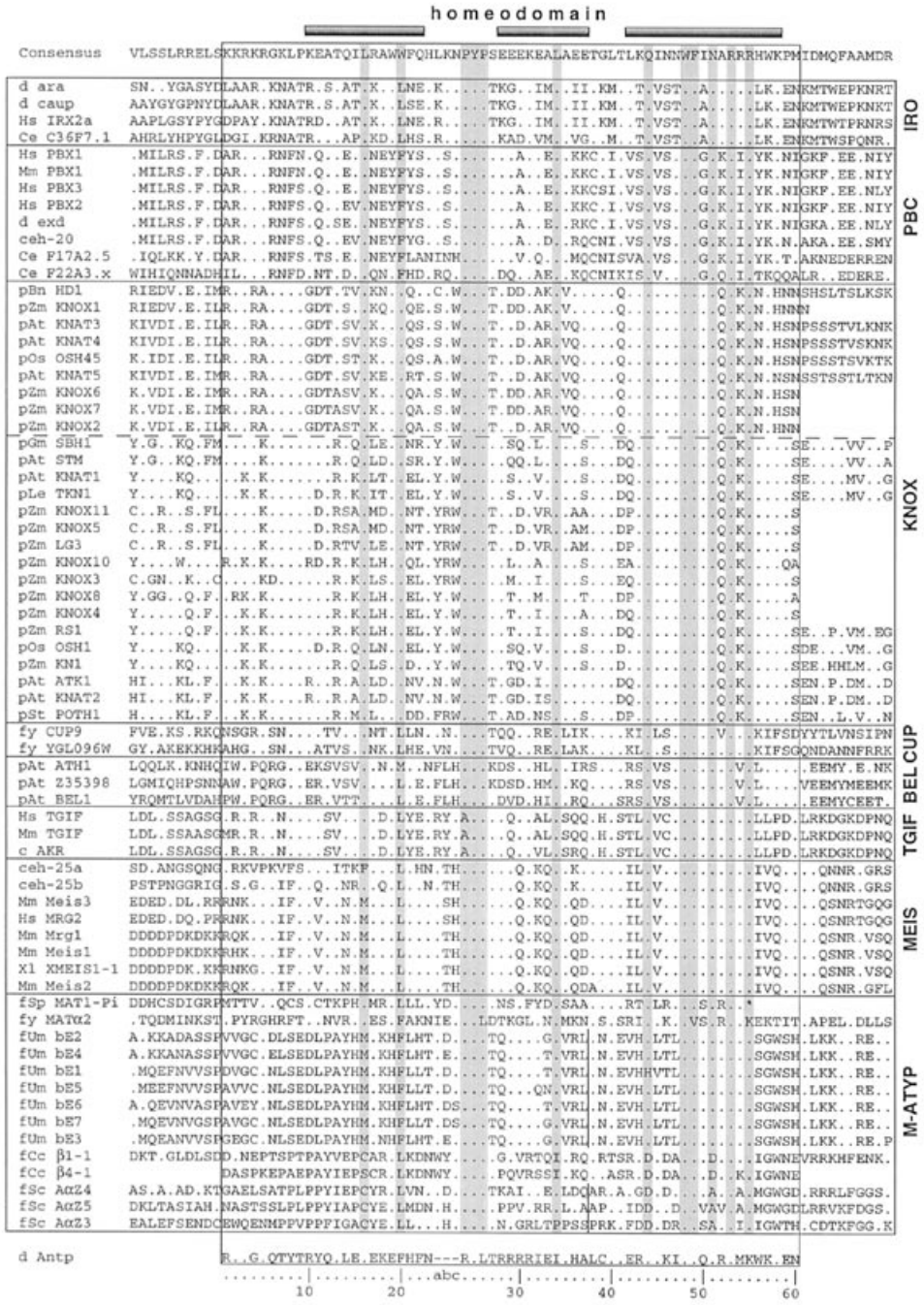


Figure 2. Compilation of TALE superclass homeodomain sequences. The homeodomain as well as the different classes are framed and labeled. For comparison, Antennapedia (Antp) is shown at the bottom. Dots represent identities to the consensus. The numbering scheme is according to (5). Grey bars highlight particular positions.

databases (GenBank and EMBL), as well as of the unfinished sequences of the *C.elegans* genome project for TALE homeobox genes were performed. Given that the complete yeast genomic sequence is available and that—including unfinished sequences—a large part of the *C.elegans* genome is now available (~80% of the genomic sequence, ~ 90% of the genes; Steve Jones, personal communication), an overview of this group of genes becomes feasible. More than 60 sequences were retrieved, and were classified based on their homeodomain sequences (Figs 2 and 3A and B). Some of the classes have already been defined previously, such as PBC (22), KNOX (21), the fungal mating type genes M-ATYP (2) and MEIS (24). The genes of the KNOX class can be grouped into two families (Figs 2 and 3), called family 1 and family 2 (21). The M-ATYP genes are highly divergent. In addition, the *Ustilago maydis* and the *Schizophyllum commune* genes have extra residues between helix 2 and helix 3 of the homeodomain, which were removed for all phylogenetic analyses in this study. Nevertheless, because they are clearly related functionally (mating type genes) as well as structurally, they have been grouped together into the M-ATYP class (2). It has been proposed that the fly Iroquois complex genes form a new class (14,25). This is now confirmed by the existence of *C.elegans* and vertebrate orthologues.

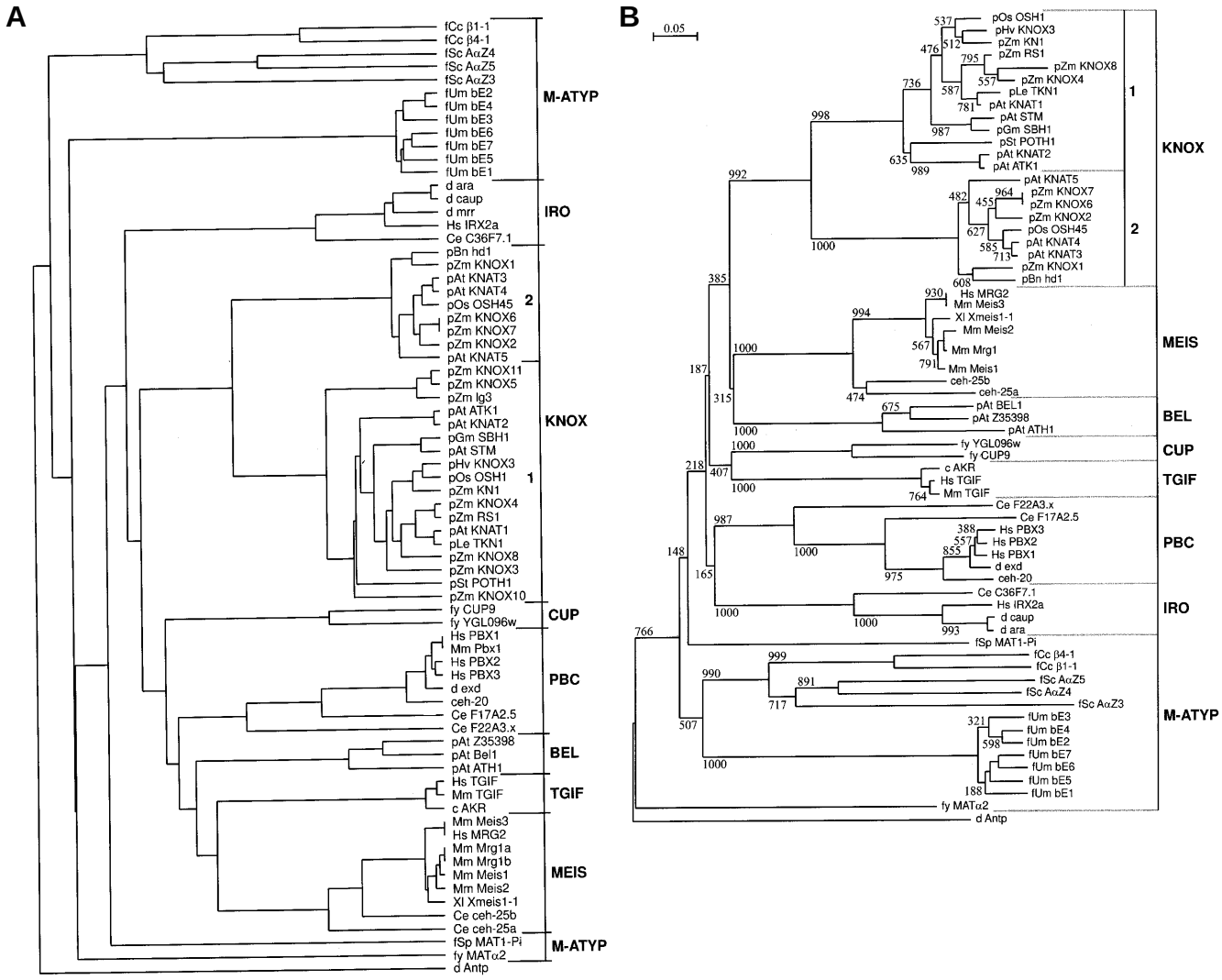


Figure 3. Comparative and evolutionary trees of TALE homeodomain sequences. (**A**) Comparative UPGMP tree generated by PILEUP. (**B**) Neighbor Joining tree generated by ClustalW. Numbers at the branches indicate bootstrap values for 1000 trials. The different classes are indicated. In all cases, the typical homeodomain of Antp was used as an outgroup.

Three additional new groups, TGIF, CUP and BEL, can be identified (Fig. 3), although they do not yet fully satisfy the criteria for a new class (4). The TGIF transcription factors (8,26) have thus far only been described in vertebrates. However, they form a distinct group, and orthologues in flies and worms might exist. Similarly, the existence of several *Arabidopsis* genes [BELL1 and ATH1 (27,28), as well as the EST pAT Z35398] that are very different from the KNOX genes seems to indicate that this BEL group could be a new class remaining to be found in other plants. This is supported by the fact that the homeodomain intron positions of the KNOX and BEL genes are different (Fig. 5). The yeast CUP9 and YGL096w genes must have arisen through a duplication event. Orthologues from other fungi are not yet known, but I refer to them as the CUP group.

Searches of EST databases revealed mammalian members of the MEIS, TGIF and IRO class, as well as plant KNOX and BEL members. These partial cDNAs, apart from the *Arabidopsis* EST Z35398, were not included in the present analysis as they do not add much additional information, but they do demonstrate that in vertebrates several members of each class exist, consistent with the view of large scale genome duplications in chordate evolution (see for example, 29,30).

Features of the TALE homeodomain

The most characteristic feature of TALE homeodomains is that they have three extra residues in the loop between helix 1 and helix 2 of the homeodomain. Furthermore, this loop is much more conserved than in typical homeodomains: positions 24–26 are virtually always proline–tyrosine–proline, except in the TGIF group, which has an alanine at position 24 (Fig. 2). This turn is often followed by a serine or threonine and several acidic residues. Other differences are at residues 16 and 20, which are very highly conserved in typical homeodomains (leucine and phenylalanine or tyrosine, respectively; 4). In TALE homeodomains position 16 can be a leucine, methionine, phenylalanine, even a cysteine, or serine and position 20 can be a phenylalanine, tryptophane, leucine or methionine. Residue 50

in the DNA-binding helix 3 of the TALE homeodomains is in many cases a small, non-polar residue. In the IRO class it is an alanine, in the PBC class it is a glycine, in most of the other genes it is an isoleucine. Position 50 is very critical for the DNA binding specificity of the homeodomain (for example, 31), and in many typical homeodomains polar residues such as glutamine, lysine, cysteine, histidine or serine are found. The fact that in TALE homeodomains a small, non-polar residue is at that position suggests that the DNA–protein interactions of TALE genes could be of a very different nature. In the case of the PBC class with a glycine, there might not even be a strong interaction with the DNA, and additional specificity might be conferred by other parts of the protein, for example the N-terminal region of the homeodomain. The characteristic differences between typical homeobox genes and the TALE class demonstrates that the TALE genes constitute a distinct separate group.

Conserved motifs outside the homeodomain

The PBC domain, a large bipartite domain upstream of the homeodomain of PBC class genes, has been described previously (22). In addition to *ceh-20*, two other genes with similarity to the PBC class were discovered in the *C.elegans* genome. F17A2.5 contains a conserved PBC domain upstream of the homeodomain (Fig. 4A). F17A2.5 is, however, in both the homeodomain and the PBC domain, less similar to the fly and vertebrate genes than CEH-20 suggesting that F17A2.5 might be the founder of a new family of PBC class genes.

Analysis of the cosmid sequence of F22A3 revealed no PBC domain, only a PBC-like homeodomain. The ORF as predicted by Genefinder did not splice the homeodomain properly; in Figure 2 the corrected splice is shown that results in a standard homeodomain. Given the lack of a PBC domain, the divergent homeobox sequence, and the poor splice acceptors in the homeodomain, it is possible that F22A3.x is not a functional gene.

Extensive sequence conservation has been observed between the three fly IRO genes (14,25). A comparison of the fly, human and worm IRO sequences (Fig. 4B) revealed that the sequence similarity is mainly restricted to the homeodomain region. In particular, an acidic patch downstream of the homeodomain is noteworthy, which might serve as a transcriptional activation domain. In addition, a short motif (25) has not only been conserved in flies, but also in worms; IRO box is proposed here as a name (Fig. 4B). Searches of the C36F7.1 ORF and cosmid C36F7 did not reveal any obvious similarity to a second motif described in the fly genes with similarity to the *Notch* genes (25).

The maize gene *Knotted 1* (10) has been the founding member of a large group of similar genes in plants. The ELK domain, just upstream of the homeodomain, has been described (20). While sequence comparisons of various Knotted-like genes have shown extensive conservation further upstream, the KNOX domain of about 100 amino acids, has previously not been defined (Fig. 4C). At least one intron position has been conserved within the KNOX domain between KNOX family 1 and family 2. A smaller, less conserved element, the GSE box, is present between the KNOX domain and the ELK domain.

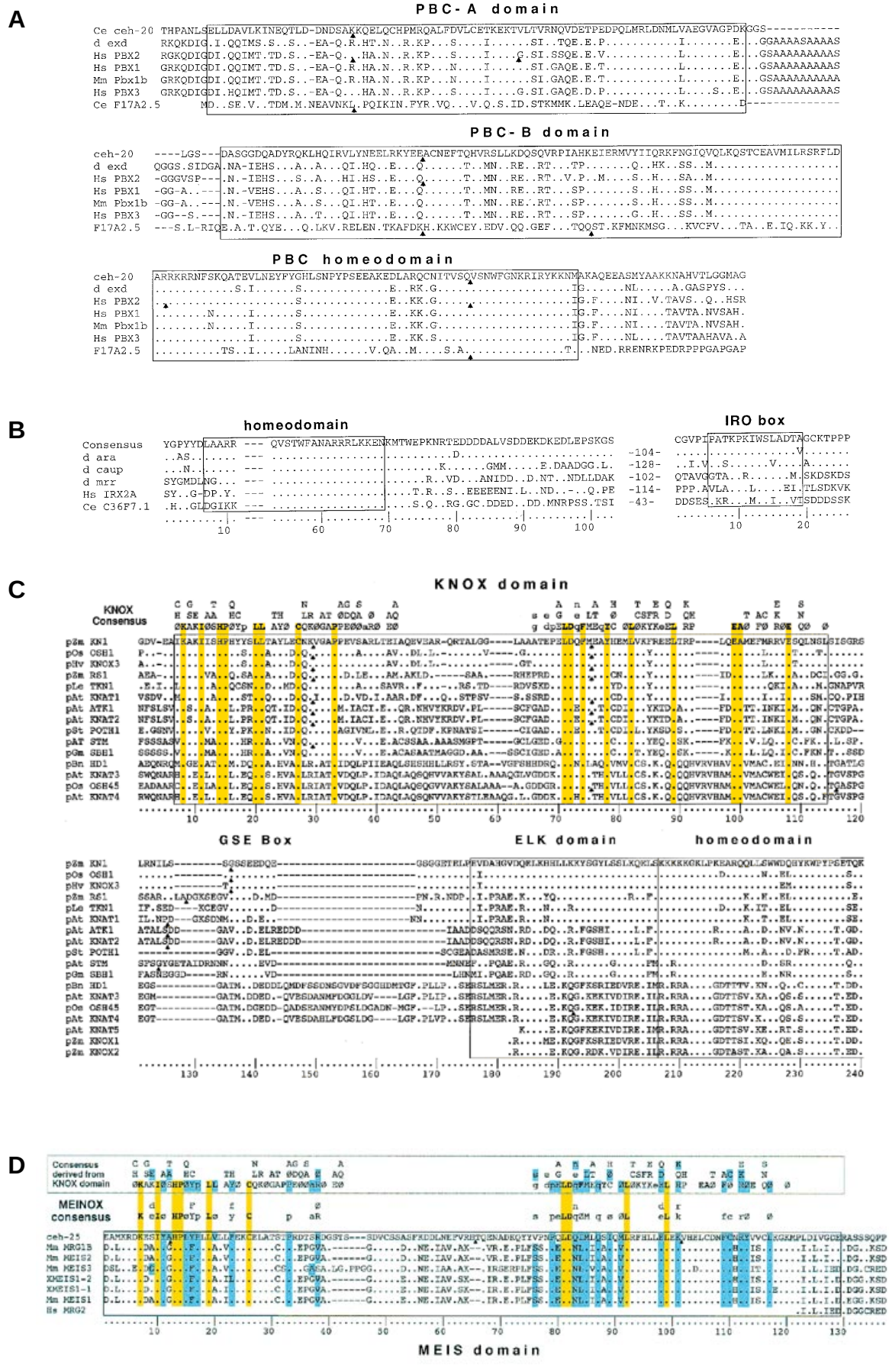
Comparison of the full *ceh-25* ORF with the Meis genes revealed a novel, highly conserved domain upstream of the homeodomain, termed MEIS domain (Fig. 4D). The presence of a second conserved domain supports the notion that *ceh-25* is the orthologue of the vertebrate Meis genes. The domain is about 130 amino acids long and bipartite, as there is a more variable region in the middle. It is separated from the homeodomain by a long variable region rich in glycine and serine residues.

During multiple sequence alignments, similarities between the Meis and Knox genes were observed outside of the homeodomain. A consensus of the KNOX domain was established, and compared to the MEIS domain (Fig. 4D). Out of 17 absolutely conserved positions in the KNOX domain, 10 are also absolutely conserved in the MEIS domain. Many additional positions share the same residues, though not always perfectly conserved, and some positions have similar residues. Clearly, the MEIS domain and the KNOX domain are both derived from the same common ancestral domain, the MEINOX domain.

Evolution of TALE homeobox genes

Three TALE superclass homeobox genes are found in the completely sequenced genome of *S.cerevisiae* that can be grouped into the M-ATYP and the CUP classes. In animals four different TALE groups have been found, PBC, MEIS, TGIF and IRO, and not many more are expected to surface. In plants the KNOX and BEL groups can be defined so far. A clear relationship exists between the MEIS and KNOX classes because of their conserved MEINOX domain. The question arises as to whether it is possible to determine how the different classes have evolved from each other. Several different methods of evolutionary tree construction were used on the homeodomain sequences to elucidate that question (see Materials and Methods). A simple UPGMA analysis clearly differentiates the different groups with exception of the fungal mating type genes (M-ATYP, Fig. 3A), which show high sequence divergence (sometimes <20% identity in the homeodomain). The MEIS, TGIF, BEL, CUP and KNOX classes are marginally more similar to each other than to the PBC, IRO and M-ATYP classes. A Neighbor Joining tree analysis using ClustalW generated a similar picture (Fig. 3B). In that analysis, the different groups are clearly demarcated, and the KNOX, MEIS and BEL genes may be most similar to each other, followed by CUP and TGIF. The bootstrap values indicate, however, that the branching pattern of the different classes from each other cannot be significantly determined. Maximum-likelihood analysis of selected sequences using Puzzle resulted in a tree which clearly clustered all the groups (again with exception of the M-ATYP genes), but the groups all branched from the root (data not shown). A Puzzle analysis that excluded the M-ATYP genes resulted in a tree in which KNOX, CUP, TGIF, BEL and MEIS were more similar to each other than to IRO and PBC (data not shown). But again the branching pattern of the different classes was not statistically significant (being only ~50%, data not shown). Finally, parsimony analysis was performed using Protpars (data not shown). Of the eight best trees generated by this method, seven produced trees in which the BEL, TGIF, MEIS and KNOX classes were most closely associated. CUP was clustered with some M-ATYP genes, while IRO and PBC grouped together.

Overall, the trees suggest that KNOX and MEIS are more closely related, although TGIF, CUP and BEL are about equally closely related to MEINOX. IRO and PBC are consistently more distantly related, in some cases they are a little more related to each other, suggesting they could be derived from a common ancestor. Interestingly, this grouping is supported by the DNA-binding characteristics: KNOX, CUP, BEL, TGIF and MEIS share an isoleucine at position 9 of helix 3, while PBC and IRO have a glycine or alanine, respectively. The M-ATYP are virtually


A
PBC-A domain
PBC-B domain
PBC homeodomain
B
homeodomain
IRO box
C
KNOX domain
GSE Box
ELK domain
homeodomain
D
MEINOX consensus
MEIS domain

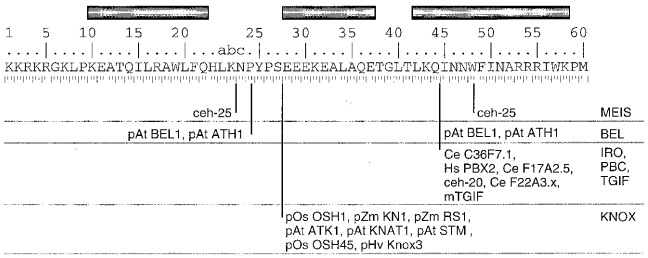


Figure 5. Intron positions in the homeodomain are indicated under the TALE consensus. Sequences are grouped according to classes.

impossible to classify due to their high variability and as a consequence they are mostly in the position of an outgroup. An analysis of the intron positions (Fig. 5) does not shed much further light into the evolutionary history. It supports the notion that the KNOX and BEL genes are distinct groups, but the intron positions between KNOX and MEIS appear not to be conserved. Interestingly, several TALE homeobox genes have an intron at position 44/45, the same position where many typical homeobox genes have an intron (4). Perhaps this intron position is extremely ancient, being already present in a common ancestor of TALE and typical homeobox genes.

DISCUSSION

Alternative splicing of *ceh-25*

The type of alternative splicing observed in the homeodomain of *ceh-25*, where two different exons can both encode part of the homeodomain, has previously not been seen in any other homeobox gene. The POU homeobox gene tI-POU produces an alternatively spliced variant where two amino acids are missing in the N-terminus of the homeodomain giving rise to I-POU, which is incapable of DNA binding (32). Alternative splicing is also seen in HOX cluster genes: the first exon of human HOX3C can splice over the homeobox of HOX3C into that of HOX3E (33), thus different homeodomain products can be produced from the same promoter. Alternative splicing that gives rise to transcripts lacking a homeodomain is also known (34). Within the TALE superclass, alternative splicing has been observed in PBC (35), MEIS (13,24) and KNOX (36). However, these alternative splices occur outside of conserved regions, in most cases giving rise to differences in the C-termini of the proteins.

The two alternative homeodomain exons in *ceh-25* have most likely arisen through a duplication event from a single ancestral exon. The *ceh-25b*-specific exon is more similar to the vertebrate Meis genes, suggesting that *ceh-25a* might have altered DNA-binding properties given the importance of the N-terminal region for DNA binding (see for example, 3). The possibility of duplicating exons containing only parts of conserved domains suggests novel ways of tinkering with motifs and creating diversity.

MEINOX, a homeodomain-associated domain conserved between plants and animals

The conservation of a homeodomain-associated domain between plants and animals clearly demonstrates that the TALE superclass of homeobox genes is very ancient and must have existed in the common ancestor of plants, fungi and animals. Searches with this new motif have not revealed any other obvious sequence matches. The function of the MEINOX domain is not known. Examination of the conserved residues suggests that it is perhaps not a DNA-binding domain, since it contains few conserved basic residues. The domain is split into two subdomains, joined by a flexible linker. Secondary structure predictions suggest that the MEINOX domain is constituted of α helices, some of which appear to be of amphipathic nature. Hydrophobic residues, which are likely to be relevant for the structure, constitute the major portion of conserved positions. Perhaps it functions in protein–protein interaction for homodimer or heterodimer formation.

Evolution of TALE homeobox genes

The existence of a MEINOX TALE gene at the origin of plants and animals provides a clear anchor point for evolutionary considerations. A further consideration is that in yeast, two groups of TALE genes exist, M-ATYP and CUP, while in animals four groups, PBC, MEIS, IRO and TGIF, have been identified. It seems likely that few, if any, further groups will be discovered in animals, since the *C.elegans* genome project has sequenced a large part of the worm genome by now. In plants the situation is less clear; two groups, KNOX and BEL, have been identified so far, but the *Arabidopsis* genome project should give a much better overview in the future. Given that fungi have only two groups, it seems highly likely that the ancestral organism of plants, fungi and animals did not have more than two TALE homeobox genes. Thus, the four animal TALE genes must have evolved from not more than two homeobox genes, perhaps only from one. The various phylogenetic analyses suggest that TGIF, MEIS, KNOX, CUP and BEL are more closely related to each other than to IRO, PBC and M-ATYP. Thus, a likely hypothesis is that MEIS, KNOX, TGIF, CUP and BEL all evolved from a common ancestral MEINOX gene, with MEIS and KNOX staying most similar to that ancestral state.

The relationships of the PBC, IRO and M-ATYP classes are more difficult to evaluate. PBC and IRO might be derived from each other. The M-ATYP class genes are highly divergent, making any assignment of that group to other classes virtually impossible.

Figure 4. Conserved sequence motifs outside of the homeodomain. Arrowheads mark intron positions, dots represent identities to the uppermost sequence, dashes indicate gaps. (**A**) PBC class genes with PBC domain and homeodomain. (**B**) The IRO class genes show extended conservation downstream of the homeodomain, in particular an acidic region. The IRO box is located further C-terminal, the numbers indicate the number of omitted residues. (**C**) KNOX class genes. The KNOX domain, the GSE box and the ELK domain are indicated. Above the KNOX domain, a consensus derived from the KNOX domain is shown. Bold capital letters, highlighted with a yellow bar, indicate absolutely conserved positions, capital letters indicate positions with three or fewer residues occurring at a particular position (Note: hydrophobic residues, marked by Ø, i.e., I, V, L, M, F, Y, W, count as 'one' residue). Small letters indicate frequently occurring residues at a particular position that are not perfectly conserved. (**D**) MEIS domain of the MEIS class genes. At the top of the panel, the consensus derived from the KNOX domain (Fig. 3C) is shown. Comparison of the KNOX consensus and the MEIS domain gives a consensus (shown in the middle) of those positions that have been conserved between KNOX and MEIS, termed MEINOX consensus; similar conventions to derive the consensus as in Figure 3C were applied. Yellow bars indicate absolutely conserved positions, blue shading marks conserved or similar residues (similar residues: Ø = I, V, L, M, F, Y, W; K, R; E,D).

Biochemical and genetic data of the fungal mating type genes shows that they interact with typical homeobox genes, which are also part of the mating type locus (for review see 37,38). Biochemical interaction between a TALE homeobox gene and a typical homeobox gene has also been documented for PBC class genes. For example, the human PBX genes interact with typical homeobox genes of the HOX cluster (for review see 39). Since both in fungi and animals TALE homeobox genes interact with typical homeobox genes, it is feasible that this interaction is an ancient conserved feature and that the M-ATYP and PBC (and possibly IRO) class homeobox genes are derived from a common ancestral gene and the ancestral organism might have had a locus similar to a mating type locus. However, whether the putative common ancestral gene of PBC/IRO/M-ATYP was the MEINOX gene, or a separate, second TALE gene present in the common ancestral organism, cannot be determined at present. The limited length of the homeodomain, together with the long evolutionary distances involved, makes the proper resolution of the deep branch points very difficult, irrespective of the computational method used. More data from other species such as sponges and coelenterates, from lower fungi and lower plants, as well as the complete sequence of *Arabidopsis*, should help to unravel the evolutionary history of the TALE homeobox genes. Biochemical studies of the MEINOX genes could provide additional helpful information; for example, are there other TALE homeobox genes that interact with MADS box genes like MATα2 (for review see 40)? Or could some of the TALE homeobox genes, such as *TGIF* or *Meis1*, be partners for typical homeobox genes, in particular those which have been shown not to interact with *PBX*/*exd*? Indeed, genetic evidence suggests that *Meis1* could interact with posterior members of the HOX cluster (41).

Nevertheless, several points can presently be made: the TALE homeobox genes have undergone much less diversification and radiation in animals than the typical homeobox genes, for which many more classes can be defined. The MEINOX genes represent an extremely archetypal form of homeobox gene which must have been present in the last common ancestor of plants, fungi and animals; this ancestral organism might have had even two different types of TALE genes. This establishes the TALE homeobox genes as an old, distinct group, which separated long ago from typical homeodomains. Thus the separation of TALE and typical homeobox genes from a common Urhomeobox gene seems have occurred at some point in protozoa evolution.

ACKNOWLEDGEMENTS

I would like to thank Prof. K.Ikeo for valuable advice with the phylogenetic programs, Drs R.Clerc and S.Hake for sharing information, and M.Naegeli and G.Niklaus for technical help. I wish to thank the Genome Sequencing Center, Washington University, St Louis, for communication of DNA sequence data prior to publication. This work was supported by grants NF. 3130-038786.93 and NF. 3100-040843.94 from the Swiss National Science Foundation and the Kanton Basel-Stadt.

REFERENCES

1 Gehring, W. J. (1994) In Duboule, D. (ed.), *Guidebook to the Homeobox Genes*. Oxford University Press, Oxford, pp. 1–10.
2 Bürglin, T. R. (1995) In Arai, R., Kato, M. and Doi, Y. (eds), *Biodiversity and Evolution*. The National Science Museum Foundation, Tokyo, pp. 291–336.
3 Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A. F., Resendez-Perez, D., Affolter, M., Otting, G. and Wüthrich, K. (1994) *Cell*, **78,** 211–223.
4 Bürglin, T. R. (1994) In Duboule, D. (ed.), *Guidebook to the Homeobox Genes*. Oxford University Press, Oxford, pp. 25–71.
5 Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D. and Pabo, C. O. (1991) *Cell*, **67,** 517–528.
6 Ceska, T. A., Lamers, M., Monaci, P., Nicosia, A., Cortese, R. and Suck, D. (1993) *EMBO J.*, **12**, 1805–1810.
7 Leiting, B., De Francesco, R., Tomei, L., Cortese, R., Otting, G. and Wüthrich, K. (1993) *EMBO J.*, **12**, 1797–1803.
8 Bertolino, E., Reimund, B., Wildt-Perinic, D. and Clerc, R. G. (1995) *J. Biol. Chem.*, **270**, 31178–31188.
9 Astell, C. R., Ahlstrom-Jonasson, L., Smith, M., Tatchell, K., Nasmyth, K. A. and Hall, B. D. (1981) *Cell*, **27**, 15–23.
10 Vollbrecht, E., Veit, B., Sinha, N. and Hake, S. (1991) *Nature*, **350**, 241–243.
11 Kamps, M. P., Murre, C., Sun, X.-H. and Baltimore, D. (1990) *Cell*, **60**, 547–555.
12 Nourse, J., Mellentin, J. D., Galili, N., Wilkinson, J., Stanbridge, E., Smith, S. D. and Cleary, M. L. (1990) *Cell*, **60**, 535–545.
13 Moskow, J. J., Bullrich, F., Huebner, K., Daar, I. O. and Buchberg, A. M. (1995) *Mol. Cell. Biol.*, **15**, 5434–5443.
14 Gómez-Skarmeta, J.-L., Diez del Corral, R., de la Calle-Mustienes, E., Ferrés-Marcó, D. and Modolell, J. (1996) *Cell*, **85**, 95–105.
15 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.*, **215**, 403–410.
16 Devereux, J., Haeberli, P. and Smithies, O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
17 Kuhner, M. K. and Felsenstein, J. (1994) *Mol. Biol. Evol.*, **11**, 459–468.
18 Strimmer, K. and von Haeseler, A. (1996) *Mol. Biol. Evol.*, **13**, 964–969.
19 Durbin, R. and Thierry Mieg, J. (1991) Code and data available from anonymous FTP servers lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk, and ncbi.nlm.nih.gov.
20 Vollbrecht, E., Kerstetter, R., Lowe, B., Veit, B. and Hake, S. (1993) *Evolutionary Conservation of Developmental Mechanisms*. Wiley-Liss, Inc., pp. 111–123.
21 Kerstetter, R., Vollbrecht, E., Lowe, B., Veit, B., Yamaguchi, J. and Hake, S. (1994) *Plant Cell*, **6,** 1877–1887.
22 Bürglin, T. R. and Ruvkun, G. (1992) *Nature Genet.*, **1**, 319–320.
23 Nakamura, T., Jenkins, N. A. and Copeland, N. G. (1996) *Oncogene*, **13**, 2235–2242.
24 Steelman, S., Moskow, J. J., Muzynski, K., North, C., Druck, T., Montgomery, J. C., Huebner, K., Daar, I. O. and Buchberg, A. M. (1997) *Genome Res.*, **7**, 142–156.
25 McNeill, H., Yang, C.-H., Brodsky, M., Ungos, J. and Simon, M. A. (1997) *Genes Dev.*, **11,** 1073–1082.
26 Ryan, A. K., Tejada, M. L., May, D. L., Dubaova, M. and Deeley, R. G. (1995) *Nucleic Acids Res.*, **23**, 3252–3259.
27 Reiser, L., Modrusan, Z., Margossian, L., Samach, A., Ohad, N., Haughn, G. W. and Fischer, R. L. (1995) *Cell*, **83**, 735–742.
28 Quaedvlieg, N., Dockx, J., Rook, F., Weisbeek, P. and Smeekens, S. (1995) *Plant Cell*, **7**, 117–129.
29 Holland, P. W. H., Garcia-Fernàndez, J., Williams, N. A. and Sidow, A. (1994) *Development*, 1994 Supplement, 125–133.
30 Sharman, A. C. and Holland, P. W. H. (1996) *Netherlands J. Zool.*, **46**, 47–67.
31 Hanes, S. D. and Brent, R. (1989) *Cell*, **57**, 1275–1283.
32 Treacy, M. N., Neilson, L. I., Turner, E. E., He, X. and Rosenfeld, M. G. (1992) *Cell*, **68**, 491–505.
33 Simeone, A., Pannese, M., Acampora, D., D'Esposito, M. and Boncinelli, E. (1988) *Nucleic Acids Res.*, **16**, 5379–5390.
34 Wright, C. V. E., Cho, K. W. Y., Fritz, A., Bürglin, T. R. and De Robertis, E. M. (1987) *EMBO J.*, **6**, 4083–4094.
35 Monica, K., Galili, N., Nourse, J., Saltman, D. and Cleary, M. L. (1991) *Mol. Cell. Biol.*, **11**, 6149–6157.
36 Tamaoki, M., Tsugawa, H., Minami, E., Kayano, T., Yamamoto, N., Kano-Murakami, Y. and Matsuoka, M. (1995) *Plant J.*, **7**, 927–938.
37 Duboule, D. (ed.) (1994) *Guidebook to the Homeobox Genes.* Oxford University Press, Oxford.
38 Kahmann, R. and Bölker, M. (1996) *Cell*, **85**, 145–148.
39 Mann, R. S. and Chan, S.-K. (1996) *Trends Genet.*, **12**, 258–262.
40 Treisman, R. (1995) *Nature*, **376**, 468–469.
41 Nakamura, T., Largaespada, D. A., Shaughnessy, J. D. Jr, Jenkins, N. A. and Copeland, N. G. (1996) *Nature Genet.*, **12**, 149–153.