Chapter 6

Hexapeptide revisited

Paul M. J. In der Rieden, Gaël Mainguy,
Joost Woltering, and Antony J. Durston

Hubrecht Laboratorium,
Nederlands Instituut voor Ontwikkelingsbiologie,
Uppsalalaan 8, 3584 CT, Utrecht, The Netherlands.
Introduction

The ParaHox cluster, an evolutionary sister of the Hox clusters, contains three genes, in chordates called *Gsh*, *Pdx1*, and *Cdx* (or *Cad*) respectively (Brooke et al., 1998; Coulier et al., 2000). An intact ParaHox cluster, containing the genes *Gsh1*, *Pdx1* and *Cdx2* is found in human and mouse, on chromosome 5 or 13, respectively. The other ParaHox genes, *Gsh2*, *Cdx1*, and *Cdx4* are not, or no longer, organised in a cluster, as they are all located on different chromosomes (Pollard and Holland, 2000), although it cannot be excluded that the genes were duplicated independently of complete cluster duplications. Amino acid sequence comparison studies using the homeodomains of the Hox and ParaHox proteins has revealed that Gsh, Pdx1, and Cdx/Cad proteins are most similar to Hox paralog groups 1 and 2, 3, and 9 through 13, respectively (Kourakis and Martindale, 2000; Yanze et al., 2001). However the homeodomains of Cdx1 and Cdx2 are closer to Hox paralog groups 8 and 9, and even to groups 1 and 2 than to the most posterior group (Van den Akker et al., 2002).

Studies concerning Hox target specificity have revealed that PBC-class co-factors, subfamily of the TALE-class of homeodomain proteins, are employed by most Hox proteins to enhance binding specificity and affinity (Pöpperl et al., 1995; Mann and Chan, 1996; Di Rocco et al., 1997). The Hox/PBC interaction is mediated via a so-called hexapeptide motif. NMR and X-ray crystallographic analysis of Lab/Exd and Hoxb1/Pbx1 fragments bound to a short DNA sequence has revealed that the tryptophan residue in the core of the hexapeptide binds a pocket formed by the atypical homeodomain of PBC family members (Jabet et al., 1999; Passner et al., 1999; Piper et al., 1999). This pocket is composed of the three amino acid loop extension of the PBC homeodomain, residues in the third helix of the homeodomain, and a residue in the C-terminal helix of PBC homeodomains (Piper et al., 1999). From known sequences the hexapeptide has been defined as a tryptophan residue surrounded by hydrophobic residues that is flanked by a lysine or arginine residue two to five amino acids C-terminally (Knoepfler et al., 1999). Previously we have shown that highly conserved amino acids are clustered around the hexapeptide sequence, and consistently identify Hox proteins as belonging to a particular paralog group (Morgan et al., 2000). We and others have suggested a recognition code. In contrast, the mechanism(s) in which ParaHox proteins achieve target specificity are largely unknown. Recently, the ParaHox gene *Pdx1* has been shown to depend on Pbx1 to fully employ its function in pancreatic development (Kim et al., 2002), raising the possibility that the “code” could be extended to
ParaHox genes. In addition, it has been reported that Cdx proteins contain a hexapeptide sequence as well (Van den Akker et al., 2002). To gain more insight into ParaHox/TALE-class co-factor interaction we undertook a search for hexapeptide sequences in all known ParaHox proteins, and if so whether conservation of flanking sequence can also be found in the ParaHox proteins. We found that all described Cdx and Pdx1 members contain a hexapeptide sequence and that in these factors hexapeptide-flanking sequences conservation exists. In contrast, Gsh members do not possess a hexapeptide sequence. More generally we searched for the presence of a hexapeptide sequence in all of the members of the Antp-class of homeodomain proteins, and found them to be widely distributed. This suggests a functional interaction between Antp-class homeodomain proteins and TALE-class co-factors early during evolution.

Sequences used in this study

**Gsh:** Ciona intestinalis Gsx (AF305500), Human Gsh1 (AB044158), Human Gsh2 (AB028838), Mouse Gsh1 (NM_008178), Mouse Gsh2 (S79041), Nephasoma minuta Gsx (AF363231), Oryzias latipes Gsh1 (AF035573), Phascolion strombi Gsx (AF363230), and Podocoryne carnea Gsx (AF268446). **Pdx1:** Human Pdx-1 (U35632), Mesocricetus auratus Pdx1 (U73854), Mouse Pdx-1 (XM_124700), Rat STF-1 (S67435), Xenopus laevis XlHbox8 (X16849), and Zebrafish Pdx1 (NM_131443). **Cdx:** Anopheles gambiae Cdx (AF119382), Bombyx mori Cdx (D16683), Carp Cdx1 (X80668), Caenorhabditis elegans Cad (NM_065590), Drosophila melanogaster Cad (NM_134301), Zebrafish Cad1 (NM_131109), Chicken CdxA (AB046532), Chicken CdxB (AF353624), Chicken CdxC (U80614), Halocynthia roretzi Cad (AB031032), Herdmania curvata Cad (AF242305), Human Cad1 (U51095), Human Cad2 (NM_001265), Human Cad4 (NM_005193), Mesocricetus auratus Cad3 (X81404), Mouse Cad1 (BC019986), Mouse Cad2 (NM_007673), Mouse Cad4 (L08061), Rat Cad2 (NM_023963), Tribolium castaneum, (AJ005421), Xenopus laevis Cad2 (U04302), Xenopus laevis Cad3 (U02034), Xenopus tropicalis Cad1 (AF417197), Xenopus tropicalis Cad2 (AF417198), and Xenopus tropicalis Cad3 (AF417199). **HB9:** Chicken HB9 (AF066861), Fugu rubripes (SINFR UG00000071220) Human HB9 (U07664), Mouse HB9 (NM_019944) and Xenopus laevis XHB9 (AF072382). **Msx:** Ambystoma mexicanum Msx1 (BAA11574), Ambystoma mexicanum Msx2 (AAD28493), Bos taurus Msx1 (BAA20367), Canis familiaris Msx2 (AJ277753), Chicken Gxh7 (D10372), Chicken Msx2 (S64478), Human Msx1 (M97676), Human Msx2 (S75361),
Hexapeptide revisited

Mouse Msx1 (BC016426), Mouse Msx2 (L11739), and Rat Msx1 (D83036). **Nkx**: *Fugu rubripes* Nkx6 (SINFRUG00000085230), *Fugu rubripes* Nkx6-1 (SINFRUG00000050033), Human Gtx (AF184215), Human Nkx6-1 (NM 006168), *Mesocricetus auratus* Nkx6-1 (CAA57166), Mouse Gtx (ENSMUSG0000041309), Mouse Nkx6 (ENSMUSG0000035187), and Rat Nkx6-1 (AF004431). **Emx**: *Fugu rubripes* Emx1 (SINFRUG00000074711), *Fugu rubripes* Emx2 (SINFRUG0000060256), Human Emx1 (ENSG00000135638), Human Emx2 (AF301598), *Oryzias latipes* Emx1 (AJ250402), *Oryzias latipes* Emx2 (AJ132403), Zebrafish Emx1 (D32214), Zebrafish Emx2 (D32215). **Hmx**: Human Hmx1 (M99587) and Mouse Hmx1 (AF009367). **Tlx**: Chicken Tlx1 (AF071874), Chicken Tlx3 (AF071875), Human Tlx1 (M62626), Human Tlx2 (BC006356), Mouse Tlx1 (S70632), Mouse Tlx2 (M75953), Mouse Tlx3 (AJ223801), *Xenopus laevis* Xhox11 (AF283694), *Xenopus laevis* Xhox11L2 (AF283693), and zebrafish Tlx3 (AY045753).

**Results and discussion**

The analysis of the complete protein sequences of Gsh members (all sequences used in this study and accession numbers are listed at the end) revealed that none of the Gsh proteins contains a hexapeptide, or a derived sequence. Therefore it is highly unlikely that they use PBC-class proteins as co-factors to enhance target specificity and/or affinity.

Human Pdx1 was shown to contain a hexapeptide (Goudet et al., 1999). Sequence alignments of the Pdx1 homologs revealed a conservation of the hexapeptide sequence in vertebrates (Fig. 1A). Sequence comparison of the homeodomain and 13 amino acids C-terminal to the homeodomain showed that Pdx1 members resemble closely the third Hox paralog group (Kourakis and Martindale, 2000), accordingly the conserved hexapeptide of Pdx1 family members resembles mostly the hexapeptide of the third Hox paralog group (FPWMK, Morgan et al., 2000). In addition, sequence conservation of the flanking regions can also be observed (Fig. 1A). The conservation of the sequences flanking the Pdx1 hexapeptide is in corroboration with the sequence conservation found in comparable sequences in Hox paralog group proteins (Morgan et al., 2000).

Phylogenetic analysis of the complete amino acid sequences of Cad/Cdx homologs reveals a division of the vertebrate caudal members into 3 groups (Fig. 2), that we named Cdx1, Cdx2, and Cdx3. We suggest reclassification of the Cdx/Cad members accordingly. A hexapeptide sequence can be found in all Cdx proteins (Fig 1B).
Because sequence conservation of hexapeptides and their flanking regions per paralog group is a feature described for Hox paralog groups 1 to 8

\[
\text{A} \quad \text{Pdx1} \quad \text{ggL}{E}_{\text{epnRvqlPPWMKSTKaHaW KgQWqGa}} \text{y} \\
\text{B} \quad \text{Cdx1} \quad \text{SPx}{A}_{qRrxpYEWMR}{R}_{s-\cdot-x7.10-}-GKTRT \\
\quad \text{Cdx2} \quad \text{QLSp}{x}_{GqRRx1cEwmRKPaq-x6.7-}-vKTRT \\
\quad \text{Cdx3} \quad \text{SPxxxxxxesYeWMKTVgSt---}-GKTRT \\
\text{C} \quad \text{HB9} \quad \text{DPIKlsAGTFQDLqWLRAstAGMIlPKMpDF} \\
\text{D} \quad \text{Msx1} \quad \text{AESPdkpeRtPWMQsPrFSPPpaRRLSPP} \\
\quad \text{Msx2} \quad \text{ASVKSEnSEdGaaWIQmpGRYSPPRHLSpT} \\
\quad \text{Nkx} \quad \text{PFWPGVmqspPWRDARLA} \\
\quad \text{Hmx} \quad \text{SDRDSPETGEEMGRAEgAWPRPGP} \\
\quad \text{Emx} \quad \text{FxssqqRDpltFYPWVhRyrlghRFQ} \\
\quad \text{Tlx} \quad \text{ltgl} \quad \text{tFPWmeSsRRfKdRfT} \\
\]

Figure 1. Conserved hexapeptide sequences and flanking regions. Capital letters denote conserved amino acids, in lower case predominant amino acids are depicted, and ‘x’ indicates that the amino acid at that position is not conserved (based on standard IUB codes). The amino acids that define the hexapeptide are shown in bold. (A) Pdx1 hexapeptide and flanking sequence conservation, the sequences compared are Human Pdx-1, Mesocricetus auratus Pdx1, Mouse Pdx-1, Rat STF-1, Xenopus laevis XllHbox8, and Zebrafish Pdx1. (B) Conserved hexapeptide and flanking region of the Cdx groups. The putative phosphorylation motif S-P is underlined. The sequences compared are for each Cdx group as follows. Cdx1: Chicken, Human, Mouse, Xenopus laevis, and Xenopus tropicalis. Cdx2: Chicken, Human, Mesocricetus auratus, Mouse, Rat, Xenopus tropicalis. Cdx3: Carp, Chicken, Human, Mouse, Xenopus laevis, and Xenopus tropicalis, Zebrafish. (C) Hexapeptide and flanking sequence conservation of Chicken HB9, Fugu rubripes HB9, Human HB9, Mouse HB9, and Xenopus laevis XHB9. (D) Putative hexapeptide and flanking sequence conservation in NKL genes. Msx: Ambystoma mexicanum Msx1 and Msx2, Bos taurus Msx1, Canis familiaris Msx2, Chicken Ghox7 and Msx2, Human Msx1 and Msx2, Mouse Msx1 and Msx2, and Rat Msx1. Nkx: Fugu rubripes Nkx6 and Nkx6-1, Human Gtx and Nkx6-1, Mesocricetus auratus Nkx6-1, Mouse Gtx and Nkx6, and Rat Nkx6-1. Emx: Fugu rubripes Emx1 and Emx2, Human Emx1 and Emx2, Oryzias latipes Emx1 and Emx2, and Zebrafish Emx1 and Emx2. Hmx: Human Hmx1 and Mouse Hmx1. Tlx: Chicken Tlx1 and Tlx3, Human Tlx1 and Tlx2, Mouse Tlx1, Tlx2, and Tlx3, Xenopus laevis Xho11 and Xho11L2, and Zebrafish Tlx3.
Figure 2. Phylogenetic tree of complete Cdx/Cad proteins, constructed using the sequences and abbreviations as follows: *Anopheles gambiae* (Ag) Cdx; *Bombyx mori* (Bm) Cdx; Carp Cdx1; *Caenorhabditis elegans* (Ce) Cad; *Drosophila melanogaster* (Dm) Cad; *Zebrafish* Cad1; Chick CdxA; Chick CdxB; Chick CdxC; *Halocynthia roretzi* (Hr) Cad; *Herdmania curvata* (Hc) Cdx; Human Cdx1; Human Cdx2; Human Cdx4; *Mesocricetus auratus* (Ma) Cdx3; Mouse Cdx1; Mouse Cdx2; Mouse Cdx4; Rat Cdx2; *Tribolium castaneum* (Tc); *Xenopus laevis* (Xl) Cad2; *Xenopus laevis* (Xl) Cad3; *Xenopus tropicalis* (Xt) Cad1; *Xenopus tropicalis* (Xt) Cad2; *Xenopus tropicalis* (Xt) Cad3.
(Morgan et al., 2000), and now also for Pdx1 ParaHox members, we analysed the sequences of the hexapeptides and flanking region of the different Cdx groups. The hexapeptide sequences of the three different Cdx groups are highly conserved (Fig. 1B), and distinct from other described hexapeptides. When the flanking sequences of the Cdx hexapeptides are compared to each other within each group, sequence conservation defining an individual Cdx group can also be observed (Fig. 1B).

Despite the differences mentioned above between the Cdx groups, all members of the Cdx family share a number of characteristics. Firstly, in all members of the Cdx1 and Cdx2 groups an acidic amino acid residue is present directly N-terminal to the tryptophan in the core sequence of the hexapeptide (Fig. 1B). In the Cdx3 group the Cad3 proteins of *Xenopus laevis* and *-tropicalis* also contain an acidic residue at that position. This acidic residue most likely influences the target specificity of the Cdx proteins in dimers with Pbx co-factors, and the fact that it is also found in all the known insect Cdx homologs suggests that this is an ancient characteristic of the Cdx proteins. Comparison of the Cdx hexapeptide to the other known hexapeptides reveals that the labial group Hox proteins have an acidic residue at the same position as the Cdx proteins (Morgan et al., 2000). Secondly, the Cdx/Cad hexapeptides contain an arginine or lysine residue at positions 2 and 3 (the latter with exception of Cdx3 proteins), C-terminal to the core tryptophan residue (Fig. 1B); this feature is only found in the hexapeptides of the fourth Hox paralog members (Morgan et al., 2000). Thirdly, an arginine or lysine residue is found at the −5 position to the core tryptophan residue in all known insect and ascidian Cdx homologs and an arginine at position −6 (and often also −5) in Cdx1, Cdx2, and Cdx3 groups, with the exception of the *Xenopus laevis* and *-tropicalis* Cad3 proteins, which have a serine residue at the −6 position instead of a basic residue (Fig. 1B). This is a feature shared with the Hox paralog group 3 proteins, where in the flanking sequences of the hexapeptide a conserved lysine is found at a position −5 to the core tryptophan residue (Morgan et al., 2000). Fourthly, in all described Cdx/Cad proteins, except those of *C. elegans* and cnidarians, an S-P sequence is found N-terminally to the hexapeptide (Fig. 1B). In the vertebrate Cdx proteins the serine residue is found at position -11 relative to the core tryptophan residue of the hexapeptide. In the insect Cad/Cdx proteins the S-P sequence is found between positions −4 to −9, depending on the species. We analysed the sequence flanking the serine residue (10 amino acids C- and N-terminally) of each Cdx/Cad member using NetPhos (Blom et al., 1999). The serine residues in the conserved S-P motifs score between 0.975 and 0.998, strongly suggesting that this indeed represents a conserved phosphorylation
site. Finally, the linker region between the hexapeptide and the homeodomain contains the amino acid sequence g-K-T-R-T in all Cdx/Cad members, directly N-terminal to the homeodomain (Fig. 1B), defining a third caudal specific motif in addition to the homeodomain and the hexapeptide. The N-terminal arm of the homeodomain has been implicated in the establishment of target specificity of Hox proteins by functional studies in *Drosophila* (Kuziora and McGinnis, 1989; Gibson *et al.*, 1990) and X-ray diffraction and NMR studies (Jabet *et al.*, 1999; Passner *et al.*, 1999; Piper *et al.*, 1999). The remarkably high sequence conservation between the Cdx family members suggests that amino acids in the linker region between homeodomain and hexapeptide contribute to target specificity. Taken together these results show that the Cdx family members contain a highly conserved hexapeptide that shares features with the first, third and fourth Hox paralog groups, but is Cdx specific. In addition, the strict sequence conservation in the linker region directly N-terminal to the homeodomain of Cdx proteins suggests a role for this region, and because it is located between the homeodomain and the Pbx interaction domain it is tempting to speculate about a role for the linker region in target specificity. Interestingly, Gsh family members do not have a hexapeptide in contrast to its closest Hox relatives (Fig 3A). *Pdx1* contains a hexapeptide, most similar in sequence to that of Hox paralog group 3 members, also the closest Hox group based on homeodomain sequence, further strengthening the suggested common ProtoHox origin of the Pdx1 and Hox paralog group 3 (Kourakis and Martindale, 2000). A scenario for successive duplication of genes and clusters has been proposed to account for the genomic organisation of the Antp-class of homeodomain proteins (Fig. 3) (Pollard and Holland, 2000). Accordingly, all four ProtoHox genes likely contained a hexapeptide. Interestingly, a hexapeptide sequence has been found in the Engrailed genes, raising the intriguing possibility that the hexapeptide origin could be even more widespread. To further investigate this we surveyed for putative hexapeptide sequences in all the Antp related homeodomain proteins.

In humans, the genes HB9, En2, and Gbx1 are linked, the same holds true for En1 and Gbx2 (Fig. 3A). These genes most likely arose by duplication of the so-called EHBox array (Fig. 3B). The Engrailed genes have been shown to contain a hexapeptide sequence (Peltenburg and Murre, 1996). Gbx1 and Gbx2 do not contain a hexapeptide sequence, but interestingly, HB9 does. The hexapeptide and its flanking sequences are conserved between chicken, *Fugu rubripes*, human, mouse, and *Xenopus laevis* HB9 (Fig. 1C). The other cluster proposed to have arisen from the ArcheHox cluster is the ProtoNKL cluster (Fig. 3B). We found that Msx, Nkx, Hmx, Emx, and Tlx homologs
contain a putative hexapeptide sequence (Fig. 1D and 3A). Under the assumption of divergence of the hexapeptide, as opposed to independent acquisition, this suggests that an ancient form of the hexapeptide was present in all of the ArcheHox cluster members, and therefore presumably in the ancestor of the Antp-related proteins. Interestingly, the TALE-class of homeodomain co-factors has been shown to be very ancient, as a TALE-class factor was present in the common ancestor of plants, fungi, and animals (Bürglin, 1997). Our findings suggest that hexapeptide mediated interactions between Antp superfamily members and TALE-class co-factors appeared early during evolution.

**Figure 3.** Distribution and proposed evolution of hexapeptide sequences in the Antp-class of homeobox genes. The schematic depiction of the clustal organisation is adapted from Pollard and Holland, 2000. (A) Overview of EHGbox, Extended Hox, ParaHox and NKL clusters. Green diamonds depict previously described hexapeptide sequences while red diamonds depict putative hexapeptides. (B) Suggested history of the hexapeptide, suggested hexapeptides are depicted by blue diamonds.
References


Gibson, G., Schier, A., LeMotte, P., and Gehring, W. J., (1990), The specificity of Sex combs reduced and Antennapedia are defined by a distinct portion of each protein that includes the homeodomain. Cell 62: 1087-1103.


Jabet, C., Gitti, R., Summers, M. F., and Wolberger, C., (1999), NMR Studies of the Pbx1 TALE Homeodomain Protein Free in Solution and


Hexapeptide revisited

