Figure S1. Residue conservation of bilaterian homeodomain sequences from the Anterior homology group. A. The homeodomain motif is represented as sequence Logos. They were obtained with TeXshade [1], from the alignments of unique bilaterian sequences of various phyla, belonging to PG1 (56 sequences), PG2 (24 sequences) and PG3 (27 sequences). B. Differential logos showing sequence deviation between the PG1, PG2 and PG3 groups, obtained with Subfamily [2]. The upper part of the logos shows residues that are characteristic for the indicated subfamily of sequences, compared to the remaining sequences whose typical residues are indicated in the lower part, in the form of a shaded and reversed logo. Residues conserved in all sequences are not shown. The stars indicate positions that have been identified by Subfamily as significantly deviating from the other subfamily. The groups PG2/3 share 7 characteristic residues when compared to PG1 (top logo), whereas PG1/2 only have 3 residues in common when compared to PG3 (bottom logo). The positions of the 7 characteristic residues between PG2 and PG3 are 7-8 (PG1 = NF, PG2/3: AY), 14 (PG1=T, PG2/3= L or V), positions 27-29 (PG1 = TRA, PG2/3 = CRP), 41 (PG1=N, PG2/3= T or S) and 43 (PG1= T, PG2/3=R or K).


Figure S2. Phylogenetic analyses of the deuterostome posterior Hox genes. Both trees were constructed from a multiple alignment of homeodomain sequences (60 residues). The sequence names are colored according to the classification by HoxPred. The vertebrate PGs clusters and the ambulacraria groups 9/10 and 11/13 are indicated. The amphioxus Hox15 sequence is indicated by a star. A. The tree is a Bayesian inference tree, which has been rooted using the Gsx gene of various deuterostomes as an outgroup. The statistical support values are posterior probabilities, indicated in blue. B. The tree is a maximum-likelihood tree, which has been rooted using the Gsx gene of various deuterostomes as an outgroup. The statistical support values are bootstrap support (100 bootstrap replicates). Support values >50 are indicated in blue.

Methods:

Maximum likelihood (ML) analyses were performed with PHYML [1]. PHYML analyses were performed using the WAG amino-acid substitution model [2], the frequencies of amino acids being estimated from the data set, and rate heterogeneity across sites being modelled by two rate categories (one constant and eight γ-rates). Statistical support for the different internal branches was assessed by bootstrap resampling (150 bootstrap replicates), as implemented in PHYML [1]. Bayesian inference was performed using the Markov chain Monte Carlo method as implemented in the MRBAYES (version 3) package [3,4]. We used the WAG substitution frequency matrix [2] with among-sites rate variation modelled by means of a discrete distribution with four equally probable categories. Two independent Markov chains were run, each containing 3,000,000 Monte Carlo steps. One out of every 200 trees was saved. The trees obtained in the two runs were meshed and the first 25% of the trees were discarded as ‘burnin’. Posterior probabilities at each internal branch were taken as a measure of statistical support.

References:


**Figure S3. Permutation tests for Hox classes.** The labels of the training dataset were permuted to evaluate whether specific classes are favored for the predictions. As permutation tests always return 100% classification into the CTL group (due to its high prior probability), only the Hox class labels were permuted here (100 independent permutation tests performed). Half of the training set are CTL sequences, here correctly classified into CTL class. The Hox sequences are classified in all PGs with a similar percentage, showing a random classification rather than a bias towards one particular class.