ACADEMIC PRESS

Molecular Phylogenetics and Evolution 24 (2002) 388–399

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.academicpress.com

An automated phylogenetic key for classifying homeoboxes

I. Neil Sarkar,[a,1] Joseph W. Thornton,[b,*,1] Paul J. Planet,[c] David H. Figurski,[d] Bernd Schierwater,[e] and Rob DeSalle[f]

[a] *Department of Medical Informatics, Columbia University College of Physicians and Surgeons, New York, NY, USA*
[b] *Center for Ecology and Evolutionary Biology, 5289, University of Oregon, Eugene, OR 97403, USA*
[c] *Department of Cellular and Molecular Biophysics, Columbia University College of Physicians and Surgeons, New York, NY, USA*
[d] *Department of Microbiology, Columbia University College of Physicians and Surgeons, New York, NY, USA*
[e] *Ecology and Evolution Division, Institute of Animal Ecology and Cell Biology, Hannover, Germany*
[f] *Division of Invertebrate Zoology, American Museum of Natural History, New York, NY, USA*

Received 26 May 2002

Abstract

When novel gene sequences are discovered, they are usually identified, classified, and annotated based on aggregate measures of sequence similarity. This method is prone to errors, however. Phylogenetic analysis is a more accurate basis for gene classification and ortholog identification, but it is relatively labor-intensive and computationally demanding. Here we report and demonstrate a rapid new method for gene classification based on phylogenetic principles. Given the phylogeny of a minimal sample of gene family members, our method automatically identifies amino acids that are phylogenetically characteristic of each class of sequences in the family; it then classifies a novel sequence based on the presence of these characteristic attributes in its sequence. Using a subset of homeobox protein sequences as a test case, we show that our method approximates classification based on full-scale phylogenetic analysis with very high accuracy in a tiny fraction of the time.

Keywords: Gene classification; Orthology; Gene nomenclature; Phylogenetics; Gene families; Homeobox proteins; Comparative genomics

1. Introduction

As sequence data have proliferated, methods for organizing and interpreting the diversity of genes has become a critical biological challenge. Within genomes, genes are organized hierarchically into gene families, due to repeated gene duplications followed by independent sequence divergence. The resulting nested hierarchy of genes is formally analogous to the naturally hierarchical organization of taxa produced by speciation. Molecular phylogenetic techniques can therefore be used to reconstruct historical relationships and classify genes within genomes, just as they are used to systematize organismal diversity. An evolutionary classification of genes provides the proper framework for identifying orthology and paralogy relationships, predicting the function of gene products, and reconstructing the evolutionary events by which particular gene family members have evolved their structures and functions. (Thornton and DeSalle, 2000) Phylogenetic analysis is computationally demanding and labor-intensive, however, and it is impractical to expect that a full-scale analysis will be conducted every time a novel gene family member is discovered, particularly in the age of whole-genome analysis.

Because phylogenetic analysis is demanding, most current methods for rapid gene classification and annotation are phenetic: they are based on aggregate measures of total similarity/difference among genes. These techniques include the widely-used Clusters of Orthologous Groups method (COG, Tatusov et al., 1997) and other BLAST-based approaches (Basic Local Alignment Search Tool, Altschul et al., 1997). Unfor-

* Corresponding author. Fax: 541-346-2364.
E-mail address: joet@darkwing.uoregon.edu (J.W. Thornton).
[1] These authors contributed equally to this work.

1055-7903/02/$ - see front matter
PII: S1055-7903(02)00259-2

tunately, phenetic techniques can sometimes result in erroneous classification because they do not distinguish shared derived from retained ancestral characters, and they count autapomorphies as informative sequence differences; they can also be undermined by differential gene loss, sequence gaps, and co-orthology relationships caused by additional gene duplications in some lineages (Thornton and DeSalle, 2000). These methods also give no indication of hierarchical relationships within gene families.

To address this problem, we have developed a simple, rapid, and automated method for classifying new gene sequences based on phylogenetic principles. Our method classifies novel sequences in the context of existing gene family information by approximating a full-scale parsimony analysis using a very efficient algorithm. This method identifies phylogenetically informative sequence elements that characterize each clade in a gene family tree and then classifies a novel gene according to the presence of these characters in its sequence. In this paper, we present the principles of our method and demonstrate its application to a subset of homeobox protein sequences.

1.1. Homeobox genes and classification

Since its discovery in 1984, the homeobox gene family has been a central subject of molecular, developmental, and evolutionary biology. Our work builds upon a body of previous analysis for the classification and interpretation of homeoboxes. In 1994 Duboule compiled a comprehensive description of nearly 400 homeobox genes found in 55 different organisms, which included a framework by Bürglin (1994) for classifying homeoboxes. Bürglin's classification of homeobox genes was based primarily on sequence similarity: genes were organized into superclasses based largely on the presence of specific sequence elements outside of the homeobox region, and a distance tree was then used to place genes within superclasses into classes (>55–57% identical amino acids), families (for which the cut-off was "less well-defined"), and finally paralog groups. A paralog group includes all genes that diversified due to chromosomal/genome duplications deep in the chordate lineage, which generated paralogous HOX clusters in vertebrates; for example, all HOXA9, HOXB9, HOXC9, and HOXD9 sequences belong to the HOX9 paralog group.

Since Bürglin's work, more than 5000 homeoboxes representing almost 1000 individual genes have been sequenced (Banerjee-Basu et al., 1999, 2000, 2001). In recent years many authors have used the growing database of gene sequences to expand upon Bürglin's (1994) initial framework and classify new sequences. In 1995, Balavoine and Telford (1995) classified new homeoboxes using phylogenetic trees. Newly generated sequences from platyhelminthes were aligned with previously discovered homeobox sequences, and these data were then analyzed phylogenetically using the parsimony criterion. Assignments to paralog groups were made based on the new gene's position in the tree. Others have used a similar approach for analysis of new sequences from other taxa (i.e., Ferrier and Holland, 2001).

More recently, some authors have used specific amino acid states that are common to genes within a class or family as a basis for classification of invertebrate sequences of uncertain affinity. The criteria used to establish these "characteristic residues" have varied across studies, however. Sharkey et al. (1997) defined them conservatively as amino acids that are conserved among all members of a paralog group but not conserved in any other group. de Rosa et al. (1999) considered a state diagnostic if it is found in more than 50% of the members of a group. These studies did not seek to identify characteristic amino acids that unite paralog groups into higher classes or to distinguish apomorphic, symplesiomorphic, and homoplastic states, although only the first of these carry phylogenetic information.

Homeobox classifications using whole genome data have presented new challenges, because they impose much greater computational demands than encountered with a few genes at a time. Most classification efforts in this category—and some using smaller numbers of sequences as well—have been based on aggregate similarity meausres. Ruvkun and Hobert (1998) and Callaerts et al. (2002), for example, used BLAST scores as the primary criterion to place new sequences into ortholog or paralog groups, although both analyses relied to some extent on tree-building and/or character analysis for additional analysis.

In our view, the variety of methods used—and the lack of a phylogenetic basis for some of them—demonstrates the need for an objective, automated, and phylogenetic approach to gene classification. We have previously used the distribution of sequence characters among clades to classify novel members of a gene family (see Planet et al., 2001). Here we have further developed, automated, and demonstated the method using a subset of homeobox genes. We focus here on homeoboxes belonging to the clade containing the proteins abdominal-B (ABD-B), caudal (CDX), and homeoboxes 9 (HOX9) through 13 (HOX13). Our analysis, consistent with previous work (Balavoine and Telford, 1995), indicates that these genes form a well-defined monophyletic group in the larger tree of all HOX cluster genes. We have focused on this subset—which we call the HOX9–13 class for convenience—because it is a well-studied clade of manageable size for demonstrating our method. We plan to extend our method in the near future to allow classification and annotation of any novel homeobox sequence.

2. Methods

2.1. Sequence acquisition

We obtained 818 homeobox protein sequences from the homeodomain resource website (http://genome.nhgri.nih.gov/homeodomain, Banerjee-Basu et al., 1999, 2000, 2001) and selected from this group 47 HOX9–13 sequences from the well-studied organisms *Drosophila melanogaster*, *Tribolium castaneum*, *Homo sapiens*, *Mus musuculus*, and *Brachiodanio rerio*. We then searched the Genpept database (www.ncbi.nlm.nih.gov/entrez) for additional ecdysozoan, lophotrochozoan, and deuterostome sequences annotated as members of this class (Appendix B) and retrieved 72 such sequences. We selected as outgroups the sequences of two closely related homeotic proteins from *D. melanogaster* that are not HOX9–13 members, for a working set of 121 sequences.

Identification of phylogenetically characteristic amino acids. To determine amino acid states on which rules for phylogenetic diagnosis can be based, we generated a sequence matrix and guide tree that comprises the 47 HOX9–13 sequences from the well-sequenced organisms listed above, plus Ultrabithorax (UBX) and Deformed (DFD) sequences from *D. melanogaster*, which were included as outgroups. Sequences were aligned using ClustalW (version 1.7, Thompson et al., 1994), requiring no insertion/deletion events. Heuristic tree searches were performed using the parsimony criterion in PAUP* 4.0b10 (Swofford, 2002). We used the parsimony ratchet, implemented via PaupMacRat (Sikes and Lewis, 2001) with the following parameters: two runs of 200 iterations each, 15% of characters upweighted per iteration, one tree saved per iteration. Exhaustive TBR branch swapping was then performed on the shortest trees found by this analysis. All parsimony trees were saved and a strict consensus tree generated. Branch support values (Bremer, 1995) were estimated using Autodecay (Eriksson, 1997) and PAUP*, with a heuristic search strategy of 10 random additions followed by TBR.

For each clade in the consensus tree at the level of paralog group or higher, we identified characteristic amino acids, which are defined as phylogenetically informative amino acid states that are found only in that clade but not in the alternate group that descends from the same node (Fig. 1). A characteristic attribute (CA) may be an amino acid state at a single position (a simple CA) or a combination of states at multiple positions (a

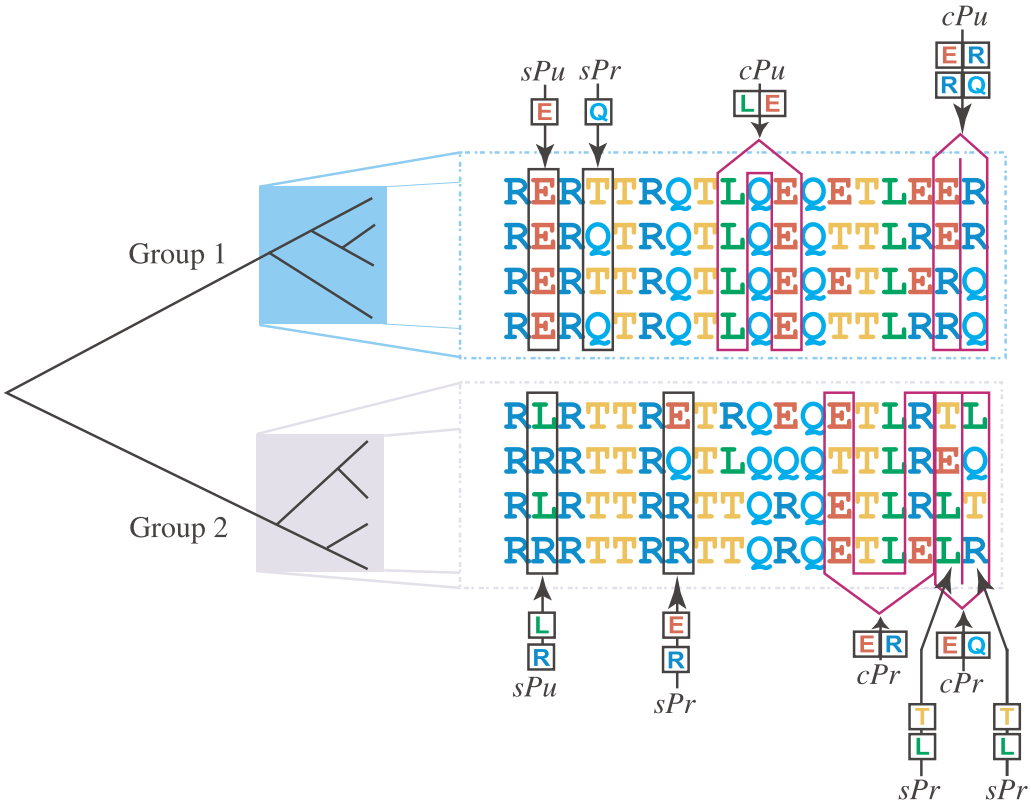


Fig. 1. Types of characteristic attributes (CAs). Pure (Pu) CAs are character states that exist in all elements of a given group but not in any members of the alternate group; private (Pr) CAs are present in only some members of a group but are absent from the alternate group. Simple (s) CAs are attributes at single positions. Compound (c) CAs are combinations of states at more than one position that occur only in the group they characterize.

compound CA); we limited this category to combinations of states at two positions. A computer application, written in C++, was prepared that automates the search for CAs. For each node at which classification will take place, the algorithm identifies amino acid states or binary combinations thereof that are unique to each descendant group. A symplesiomorphy filter then performs a Fitch optimization (Williams and Fitch, 1989) and removes any of these attributes that is a possible most parsimonious ancestral reconstruction at that node. Application of this procedure, detailed in Appendix A, provides a diagnostic rule set for classification of novel sequences into clades in the guide tree.

2.2. Classification

The remaining 72 HOX9–13 sequences that were not included in the guide tree constituted our "new" sequences for classification. We prepared a script, written in Perl 5.0, that uses the rule set generated from the guide tree and matrix to classify each novel sequence. For each node beginning at the tree's root, the algorithm searches the novel sequence for characteristic states that support classification in one of the two descendant clades; the clade with which the novel sequence shares more CAs is the one in which it is classified. This process is repeated for successive nodes until the new sequence has been classified into a paralog group. In the case of a tie or a lack of CAs at any node, no classification is made and the algorithm stops. The degree of support for classification of a sequence into a group at each node is calculated as the number of CAs that support that classification minus the number of CAs that support the alternate possible classification. This procedure is detailed in Appendix A.

2.3. Accuracy

To determine the accuracy of our method as a phylogenetic approximation, we inferred the phylogenetic tree of all 121 sequences, which includes those in the guide matrix as well as the "new" sequences, using the same search strategy described above. The tree was rooted on two related outgroup genes from *D. melanogaster*.

Each classification of a sequence at a node was categorized for accuracy as follows. A true positive (TP) classification places the sequence into the same clade that the full-scale phylogenetic analysis does. A false positive (FP) classification places the sequence into a clade in which it does not appear in the validation tree. A true negative (TN) excludes the sequence from a clade from which it is also excluded in the validation tree, and a false negative (FN) excludes a sequence from a clade in which it appears in the validation tree. Overall measures of accuracy reflect the frequency of these kinds of judgment among all classifications made in our diagnosis. Recall—the ability to correctly place sequences into the clade in which they belong—is calculated as TP/(TP + FN). Precision—the fraction of sequences classified in a clade that actually belong there—is calculated as TP/(TP + FP). Overall accuracy, or one minus the rate of errors of all types, is calculated as (TP + TN)/(TP + TN + FP + FN). For comparison, we also classified each "new" sequence by using it as query in a BLASTP search of our database of guide sequences; the sequence was classified in the same paralog group as its best hit under this method. Accuracy for BLASTP classification was calculated as above, but because this is a non-hierarchical approach, only paralog groups, not higher-level clades as well, were included.

3. Results and discussion

3.1. A genomic key for phylogenetic classification

The method described here provides a very rapid and easily implemented approximation of the parsimony algorithm for classifying members of a gene family. It takes the form of an automated "molecular key" based on attributes that are phylogenetically characteristic of each clade in a gene family tree. The method is analogous to the morphological keys that bird watchers and amateur botanists have used for decades to identify species in the field, but with several important differences. In this case, all characteristic amino acids for each clade in the tree are objectively identified, and a novel sequence is classified based on the net support from all CAs that provide evidence to include it in each clade. As with morphological keys, the classification occurs on a node-by-node basis, beginning at the root node and following a tipward path until the highest possible degree of supportable resolution is achieved (Fig. 3B).

This technique allows any novel amino acid sequence to be classified within a gene family, given an alignment of existing sequences of family members and a bifurcating "guide tree" of their evolutionary relationships. The algorithm (Appendix A) proceeds in two parts: identification of characteristic attributes for each clade in a gene family tree based on the alignment and guide tree, and then classification of the novel sequence based on the presence/absence of CAs in that sequence. The concept of a characteristic attribute—a character state found in one clade but not its sister group—is derived from population aggregation analysis, a theoretical framework for identifying attributes that define phylogenetic species or populations (Davis and Nixon, 1992). As Fig. 1 shows, a CA for a clade can be *pure* (shared by all members of the clade and absent from the other clade) or *private* (shared by some members and absent from the other clade). A combination of states at more than one amino acid po-

sition that are not phylogenetically characteristic in themselves can also be a CA, if the combination is restricted to the clade in question. Combinations are called *compound* characteristic attributes, while *simple* CAs are composed of a single position. Characteristic attributes that are symplesiomorphies—inherited ancestral states lost from the alternate clade—are filtered out, as these provide no useful phylogenetic information.

Once CAs have been identified, the novel gene is classified based on the presence of these characteristic elements in its sequence. Beginning at the root node of the tree, our method calculates the number of characteristic attributes that the novel sequence shares with each of the two possible clades descending from that node; it assigns the novel sequence to the clade with which it shares the greatest number of CAs. Our algorithm makes no distinction between private and pure CAs, and simple and compound CAs are also counted equally, since each type of CA implies a difference of a single step in an unweighted parsimony analysis (Fig. 2). This process is repeated iteratively across nodes, moving towards the tree tips, until the sequence has been classified at the finest possible level of resolution. In the case of a tie, including at score zero, the classification algorithm stops, and the sequence is placed in a polytomy at the base of the last clade to which it has been assigned.

The effect of this procedure is to approximate a full-scale parsimony analysis that includes the new sequence plus those in the guide tree (Fig. 2). Characteristic attributes are apomorphic character states that have no evidence of homoplasy due to convergence or reversal within the clades being examined. The result of classifying according to CAs is to place the novel sequence in the location in the tree that will minimize the total length of the tree. A sequence with amino acid state *i* at some position always requires an extra step to be added to the tree length unless the gene is placed in a clade in which state *i* can be inherited from a common ancestor; characteristic attributes of a clade are precisely those states for which descent without parallelism or reversal is possible. The use of compound CAs allows sequences to be classified according to unique combinations of amino acid states, minimizing the length of the tree over multiple characters. Although our classification method is based on the same logic that underlies the use of the parsimony criterion (Farris, 1983; Hennig, 1963), the guide tree can be generated using any approach, including likelihood or Bayesian methods.

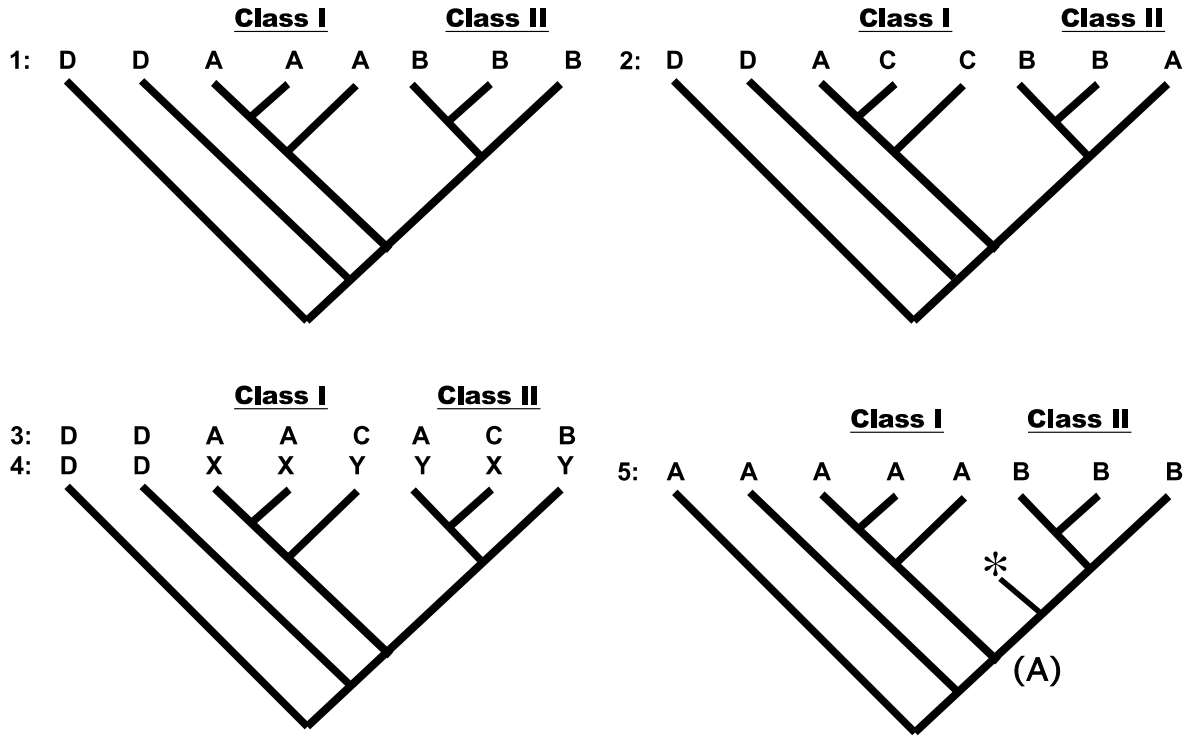


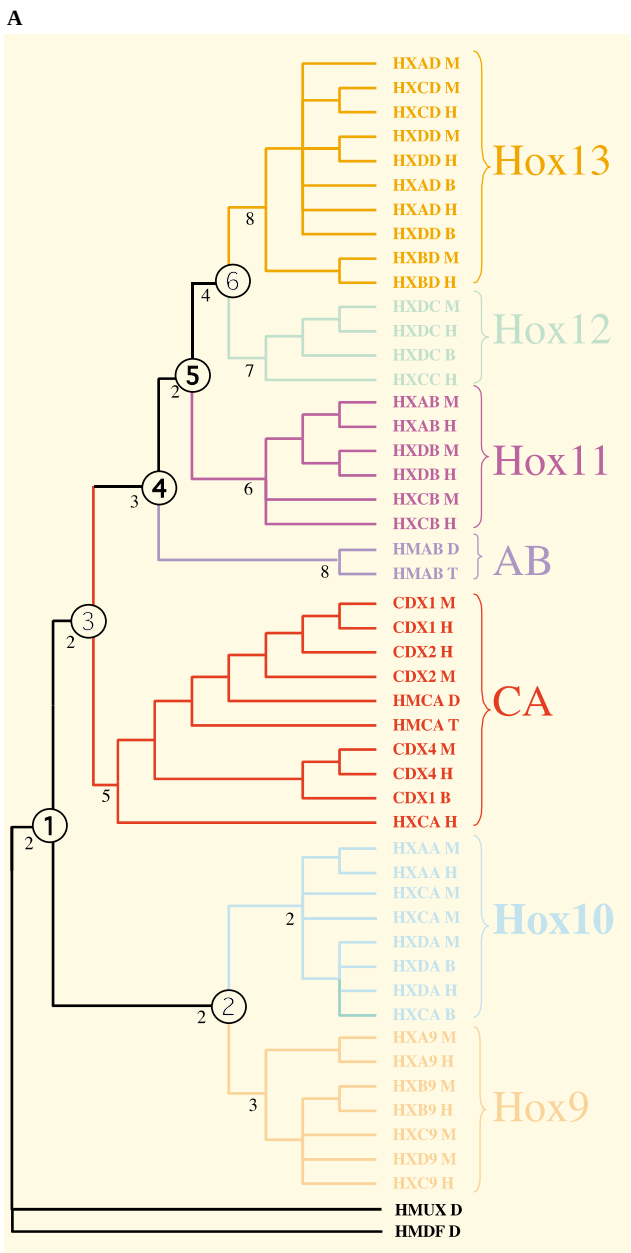
Fig. 2. Classification using pure, private, and compound characteristic attributes (CAs) approximates full-scale parsimony analysis. For character 1, states A and B are pure CAs of classes I and II, respectively. A novel sequence with state A can be placed in class I without adding any extra steps to the tree, but placement anywhere in class II would require one additional step. For character 2, B and C are private CAs for classes I and II. A novel sequence with state C can be placed in class I with no extra steps, but placement in class II would impose one additional step. For characters 3 and 4, the combination of states AX appears is a compound private CA for class I. A novel sequence with states AX can be placed in class I with no extra steps but requires 1 or 2 extra steps to be placed anywhere within class II. For character 5, state A is not phylogenetically characteristic of class I because it is symplesiomorphic, as shown by ancestral state at the node that connects classes I and II. A novel sequence with state A could be placed in class I or at the position labeled * in class II without adding extra steps to the tree.

3.2. Classification of HOX genes and accuracy of the method

To classify HOX 9–13 genes and evaluate the accuracy of our approach, we gathered available HOX9–13 members from vertebrates, ecdysozoans, and lophotrochozoans totaling 119 ingroup sequences. We divided this universe of sequences into two groups—one set was used to construct the guide tree and infer CAs, and the remainder constituted the "novel" sequences that were classified based on the rules generated from the first group. For the first group, we sought to demonstrate that our approach can be effective with minimal but broadly distributed taxon sampling. We therefore selected all 47 HOX9–13 sequences, caudal (CDX) and abdominal-B (ABD-B) from just five well-sequenced research organisms—human, mouse, zebrafish, fruitfly, and flour beetle. Sequences were aligned, and the guide tree inferred by parsimony-based analysis. Relationships among all paralog groups are fully resolved in the strict consensus of most parsimonious trees. Most paralog groups are well supported, but relationships among these groups have only weak character support (Fig. 3A).

We used this guide tree and alignment to infer CAs for each of the seven paralog groups ABD-B, CDX, and HOX9–13, as well as the more inclusive clades that unite these groups into higher classes. Both simple and compound CAs are distributed across the length of the homeodomain sequence (Fig. 4). Because compound CAs represent amino acid combinations that are unique to specific clades, at least some identified compounds are likely to represent biologically significant interactions among amino acid residues that are unique to specific functional and phylogenetic groups.

In total, 1339 CAs were identified, of which 501 were simple and 838 were compounds. All clades have a substantial number of characteristic residues, although the number of simple versus compound CAs varies considerably among groups. It is apparent that CAs of all types provide important information. For clade HOX9/HOX10, for example, there were no pure simples or compounds, but there were 11 private simples and 148



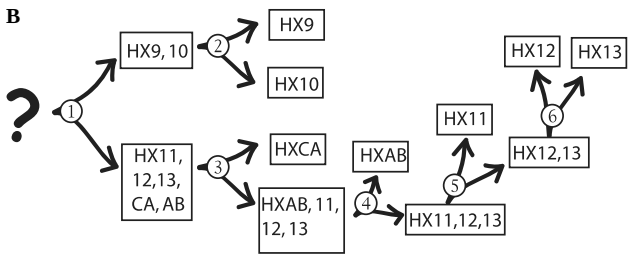


Fig. 3. Phylogenetic framework for discovery of HOX9–13 characteristic attributes. (A) HOX9–13 guide tree for identification of nested phylogenetic classes of proteins. This phylogeny is the strict consensus of 8961 most parsimonious trees (length = 225 steps, CI = 0.70, RI = 0.90) of all HOX9–13, caudal (CA or CDX), and abdominal-B (AB) sequences from *Homo sapiens* (H), *Mus musculus* (M), *Brachidanio rerio* (B), *Drosophila melanogaster* (D), and *Tribolium castaneum* (T). Small numbers below branches give branch support values (Bremer, 1995). Large numerical labels on nodes are gene duplications, with numbering corresponding to the tests in (B). Sequence accessions are given in Appendix B. (B) Overall flow of HOX9–13 classification algorithm. Each node in the guide tree that represents a gene duplication separates the gene family into alternate classes and represents a binary classification test. For each node, characteristic amino acid states are identified. A novel sequence is classified at each node into the descendant clade with which it shares the most characteristic attributes. The new sequence follows a path of tests from root node to paralog group assignment. If a decision cannot be made due to a tie or lack of CAs in the sequence, the algorithm stops at the last supported classification, placing the sequence in a polytomy at the relevant node.

private compounds, so classification at this node was entirely dependent on privates. Compounds were also important: several sequences could not be fully classified on the basis of simple CAs alone but were accurately classified when compounds were included (see below).

We used the CAs inferred from the guide tree to automatically classify the remaining 72 "novel" sequences that were not used to construct the guide tree. Classifications were deemed accurate if they were identical to those derived from full-scale phylogenetic analysis. The "validation tree" against which classifications were judged was the strict consensus of most parsimonious trees of all 121 sequences—the entire universe of 47 guide sequences, 72 novel sequences, and 6 outgroups (not shown). This tree was topologically consistent with the guide tree, although some of the deep relationships among paralog groups were not resolved.

Our method was very accurate. Overall, 71 of the 72 sequences were classified exactly as they were in the full-scale phylogenetic analysis. The one remaining sequence, accession AAF73210 from the arachnid *Achaeranea tepidarorum*, was a partial sequence only 25 amino acids in length. This sequence was classified correctly at the deepest node into ABD-B/CDX/HOX11–13, but it was not classified further, presumably because of the limited information available in its sequence. Numerous other members of our set of "novel sequences" were also truncated—to lengths as short as 27 amino acids—and all of these were classified correctly.

Statistical measures of the method's accuracy were very high. There were 196 classifications performed in all; the number of classifications is larger than the number of sequences because each sequence must be classified into multiple nested clades. The method's overall precision—the fraction of sequences classified in a clade that actually belong there—was 100%. Recall, or the ability to correctly place all members of a clade into that group, was 99.0%. The total error rate (false positives plus false negatives as a fraction of all classifications) was 0.5%, and the total accuracy rate was therefore 99.5%. When only complete sequences are considered, all measures of accuracy are 100%.

BLAST-based classification was also accurate but slightly less so than our character-based approach. Classifying the novel sequences based on their best hits in a BLASTP search of our guide sequence database resulted in correct assignments for all sequences except for the truncated sequence AAF73210 of *A. tepidarorum*, the same one that was not fully classified in our analysis. The BLAST method positively misclassified this sequence as an ABD-B rather than a CDX protein, as the phylogeny indicates it should. Overall accuracy measures for the BLAST classifications were 98.6% precision, 98.6% recall, and 98.6% total accuracy, for an overall error rate of 1.4%. The difference between the two methods' performance was that BLAST erroneously assigned this sequence to the wrong group—resulting in a false positive and a false negative error—

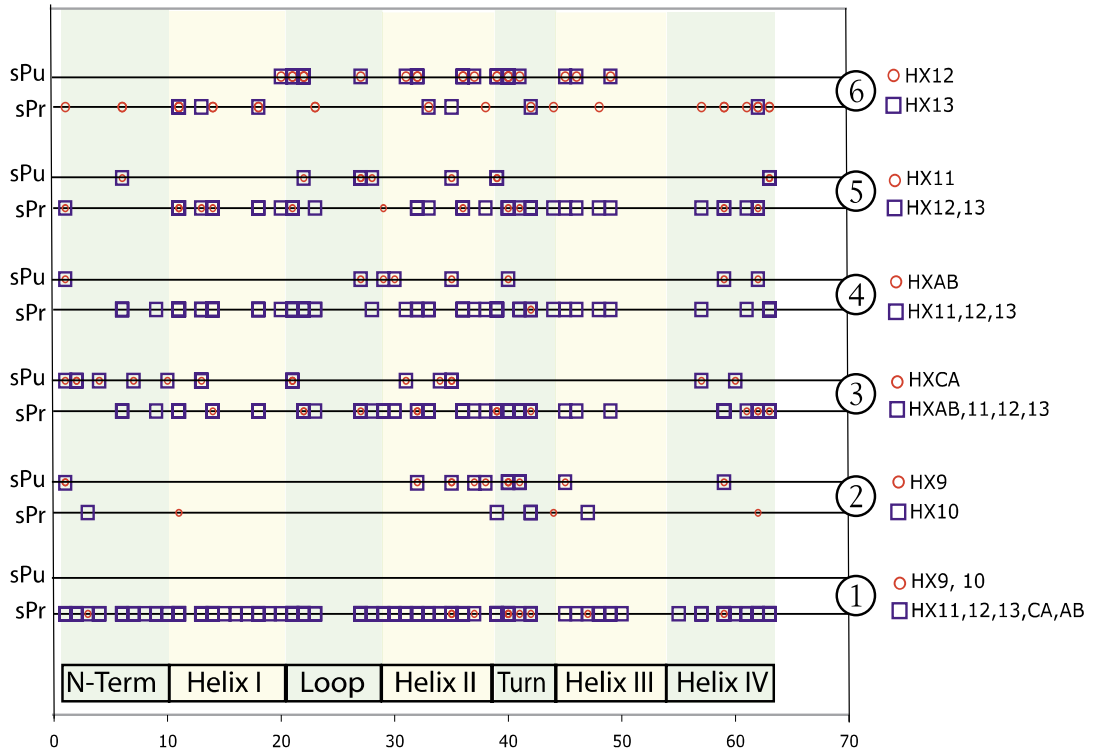


Fig. 4. Distribution of simple characteristic attributes in the HOX9–13 homeodomain. For each of the six classification tests, simple CAs for each clade are shown across the 63 amino acid homeodomain, with structural elements indicated.

whereas our technique, which did not classify the sequence at nodes where diagnostic character information was lacking, committed only a false negative. Indeed, one problem with BLAST is that it is non-hierarchical: it always selects a single sequence as the best hit, and any false negative is always a false positive as well, even if a sequence is equally related to two others generated by gene duplication or speciation.

Classifications of most individual sequences by our method were very well supported, but the degree of support varied among classifications, because the number of possible CAs differs across nodes, as does the number of CAs present in any one sequence. We measured support for each classification as the classification support index (CSI), calculated as the number of CAs that supported the classification minus the number that supported the alternate possible classification at that node. Averaged over all classifications performed, the mean CSI was 31.2, with 42% of the total support provided by simple CAs and 58% by compound CAs. Support for individual classifications ranged from CSI = 3 for classification of the partial sequence of the Cdx protein of the crustacean *Sacculina carcini* into its paralog group, up to a value of 72 for classification of HOXD9 of *Gallus gallus* into its paralog group.

All groups within the HOX9–13 class had ample support on average (Table 1). Classifications in certain groups, such as HOX10–13, received almost all their support from simple CAs, with minimal support from compound CAs. For other classes—such as CDX, ABD-B, and HOX11—compounds provided considerable supplementary support for classifications that were weakly established by simple CAs alone. Including compounds never reversed the classifications provided by simple CAs alone; two sequences, however, required compounds to achieve full resolution of their classification, as no pure CAs were present in their sequences for classification at specific nodes. Compounds do add some noise to the analysis, however: in about 5% of classifications, the use of compounds actually reduced by a small amount the support for classifications that was established by simple CAs alone. We conclude that compounds are necessary for classification in gene families in which simple CAs are not abundant, but in isolation they are not entirely reliable. For this gene family, the noise imposed by compounds was never greater than the signal provided by simples alone or simples and compounds combined.

Table 1
Average classification support index (CSI) and number of possible CAs for each class

Class	Total support	Simples	Compounds
HOX9, 10	42.8 (15.4)/159	2.2 (1.3)/11	40.8 (14.4)/148
HOX11–13, AB, CDX	19.6 (8.7)/198	17.0 (5.7)/141	2.6 (6.0)/57
CDX	49.0 (24.2)/203	12.3 (4.0)/26	36.7 (20.9)/177
ABD-B, HOX11–13	28.1 (10.7)/198	21.2 (6.6)/82	6.9 (8.3)/69
ABD-B	54.2 (29.8)/98	5.2 (2.4)/9	48.9 (30.1)/89
HOX11–13	23.8 (3.4)/73	24.0 (3.4)/73	0.0 (0.0)/0.0
HOX12, 13	35.9 (4.7)/130	18.0 (2.3)/48	17.9 (5.2)/82
HOX9	10.6 (1.0)/14	9.5 (1.0)/12	1.1 (0.3)/2
HOX10	10.8 (1.7)/21	9.4 (1.1)/19	1.4 (0.9)/2
HOX11	97.6 (6.1)/201	11.4 (0.5)/19	86.2 (5.7)/182
HOX12	29.2 (3.6)/60	21.6 (1.3)/39	7.6 (3.1)/21
HOX13	23.0 (1.0)/31	21.7 (0.6)/22	1.3 (1.2)/9

The mean support (CSI) for classification of "novel" sequences in each group is shown, with the standard deviation in parentheses and the total number of CAs identified for that class from the guide tree after the slash. CSI is calculated as the number of characteristic attributes that support the classification minus the number of CAs that support the alternate classification.

We were able to accurately classify not only "easy" sequences from taxa closely related to the insect, mammal, and teleost sequences used in our guide tree but also HOX9–13 members from taxa that are distantly related to the taxa from which characteristic amino acids were derived. Our method classified sequences from annelids, priapulids, onychophorans, elasmobranches, and tunicates with 100% accuracy. The AbdB gene from the priapulid *Priapulus caudatus*, for example, was classified with support values of 7, 21, and 69, into classes ABD-B/CDX/HOX11–13, ABD-B/HOX11–13, and ABD-B, respectively.

The posterior homeobox genes of lophotrochozoans present a significant classification challenge, because previous full-scale phylogenetic analysis has not resolved exact paralog group relationships to the HOX9–13 genes of ecdysyzoans or deuterostomes (de Rosa et al., 1999). These genes have been dubbed Post-1 and Post-2 to denote their relatedness to the posterior homeobox cluster without implying any specific classification. Our method was able to resolve the classification of these lophotrochozoan proteins to a greater extent than previous analyses. We analyzed six available Post-1 and Post-2 sequences from the brachiopod *Lingula anatina*, the annelid *Nereis virens*, and the cephalopod *Euprymna scolopes* and we found that all six were classified in the HOX11–13 clade (mean CSI = 5.7) and some were placed into the HOX12/13 class (Table 2). We cannot validate these classifications because of the sequences' uncertain phylogenetic placement in full-scale analysis. However, the presence of phylogenetically characteristic amino acids indicates that these proteins are most closely related to deuterostome proteins HOX11, HOX12, and/or HOX13.

Here we have evaluated our method's performance on only a single gene family. HOX9–13 sequences are very well-conserved within paralog groups and have no alignment ambiguity. On the other hand, they are only 60 amino acids in length, and they contain limited phylogenetic information. Although most paralog groups in our guide tree are well supported, many of the deeper nodes that determine the relationships among

Table 2
Classification of lophotrochozoan posterior homeobox proteins

Species	Protein	Terminal class	Support (CSI)
Euprymna scolopes	Post-1	HOX12/13	4
Lingula anatina	Post-1	HOX12	3
Nereis virens	Post-1	HOX12/13	2
Euprymna scolopes	Post-2	HOX11/12/13	7
Lingula anatina	Post-2	HOX11/12/13	5
Nereis virens	Post-2	HOX12	5

Posthox sequences were classified using characteristic attributes derived from sequences in our HOX9–13 guide tree. The most tipward classification possible for each protein is shown, along with the classification support index for that classification.

these groups are weakly supported and are not robust to changes in taxon sampling; indeed, many higher-level groupings collapsed in the validation tree when more taxa were included. Our ability to classify sequences at deeper nodes in the tree appears to have been due primarily to private CAs furnished by lower-level classes (such as paralog groups) within larger clades. This result suggests that—for the HOX9–13 group at least—character-based classification is not disrupted by changes in inferred phylogeny caused by more complete taxon sampling. That we were able to accurately classify sequences in a family with limited phylogenetic information and only weakly supported hierarchical structure appears to indicate that characteristic attributes provide a powerful and efficient basis for gene classification.

3.3. A web-based tool for homeobox classification

We have developed a web interface for our program that classifies a user-supplied HOX9–13 sequence, based on the tree and sequence-derived diagnostic rules described here (http://cpmcnet.columbia.edu/dept/figurski/homeo). The website automatically extracts the homeodomain from an amino acid sequence of any length and aligns it to other known homeobox sequences using the BLAST algorithm, and then proceeds to classify the sequence based on the rule sets we have developed (based on our guide tree). Although we have focused here on the precise classification of HOX9–13 proteins, diagnostic rule sets could be generated for every class in the homeobox family. Rules could also be generated to classify proteins within each HOX paralog group to the finer level of the chromosomal cluster to which each paralog belongs (i.e., A9, B9, C9, and D9). We anticipate that by the publication date of this paper, our web-based program will classify the full range of HOX sequences.

The method developed here can be extended to classify other gene families, as well, providing a rapid and efficient means for identifying and annotating new sequences. Given only a "canonical" alignment of gene family members and a tree of their relationships, new family members can be classified phylogenetically in a fraction of a second of computer time. Sequences and alignments for many gene families are available from protein family databases like PFAM (Bateman et al., 2002; http://www.sanger.ac.uk/Pfam) and SMART (Letunic et al., 2002; http://SMART.emblheidelberg.de/). The only curatorial effort required for classification by our method is to establish the guide tree and update it occasionally to improve the accuracy of the phylogeny and its CA-based classification rules as the sequence database grows. The accuracy of the method for other gene families remains to be validated. If our experience with the homeoboxes is not an aberration, however, this approach to rapid character-based phylogenetic classification may be a generally useful tool in genomics and gene family analysis.

Acknowledgments

Supported by the Columbia University Earth Institute and the Lewis B. and Dorothy Cullman Program for Molecular Systematic Studies at the American Museum of Natural History.

Appendix A. Algorithm for classification using simple characteristic sequence attributes

1. Identify characteristic attributes (CAs). Given an aligned sequence matrix C characters in length, and a phylogeny of those sequences with N nodes:
   (a) Identify each node in the phylogeny ($i = 1, 2, \ldots, N$, where $i = 1$ is the root node and its two descendant clades are i_a and i_b).
   (b) Beginning at the ingroup root node
      (1) For column 1 in the matrix, identify all character states found in all sequences in clade i_a. Gaps are missing data, not states.
      (2) For each state identified, determine if that state is present in column 1 of any sequence in clade i_b. If present, state is not a CA. If absent, state is a potential CA; store in CA set for clade i_a, column 1.
      (3) Determine all possible most parsimonious ancestral state reconstructions (MPRs) for column 1 at node i using Fitch optimization. Eliminate any stored CAs for i_a, column 1 that are MPRs at node i, column 1.
      (4) If a CA exists in all members of i_a, label it "pure." If it is present in only some members of i_a, label it "private."
      (5) Store CA states for column 1, clade i_a.
      (6) Move to the next column and return to step 1b1. If column is the last in the matrix, go to 1b7.
      (7) Move to clade i_b and return to step 1b1.
   (c) Move to the next node and repeat step (b). When all nodes have been evaluated, stop.

2. Classify a novel query sequence using the CAs identified in section A:

(a) Align query sequence to matrix, using BLASTP.

(b) Determine classification using only the aligned portion of the query sequence. Beginning at the root node

(1) Let $V(i_a)$ be the number of CAs for clade i_a that are found in the query sequence. Set $V(i_a) = 0$.

(2) For each column in the sequence, determine if the state in the query sequence is found in the list of CAs for that column for clade i_a. If yes, add 1 to Vi_a.

(3) Repeat 2b1–2b2 for all columns.

(4) Repeat 2b1–2b3 for clade i_b.

(5) Calculate classification support index (S) for each clade, such that $S(i_a) = V(i_a) - V(i_b)$, and $S(i_b) = -S(i_a)$.

(6) If $S(i_a) = 0$, store "stop" and skip to 2(c); otherwise continue.

(7) If $S(i_a) > S(i_b)$, then classify query in i_a and store "$i_a(S_{ia})$." If not, classify query in i_b and store "$i_b(S_{ib})$."

(8) Move tipward on the tree within the clade in which the query has been classified. If another node is present, go to step 2(b). If only tips are present, stop and store value "stop."

(c) Report all stored classifications and support indices.

Appendix B. Sequences used in this analysis

Abbreviations correspond to terminal labels in Fig. 3A. Sequences are categorized as follows: used in guide tree and matrix (G), "novel sequence" used in validation tree and classified using CAs from guide sequences (N), and outgroups (O).

Abbreviation	Accession	Species	Category
CDX1_B	CAA47380	*Brachydanio rerio*	G
HXAD_B	P79724	*Brachydanio rerio*	G
HXCA_B	CAA61029	*Brachydanio rerio*	G
HXDA_B	Q90469	*Brachydanio rerio*	G
HXDC_B	Q90471	*Brachydanio rerio*	G
HXDD_B	Q90472	*Brachydanio rerio*	G
HMAB_D	P09087	*Drosophila melanogaster*	G
HMCA_D	P09085	*Drosophila melanogaster*	G
CDX1_H	XP_003791	*Homo sapiens*	G
CDX2_H	Q99626	*Homo sapiens*	G
CDX4_H	AAB66319	*Homo sapiens*	G
HXA9_H	P31269	*Homo sapiens*	G
HXAA_H	P31260	*Homo sapiens*	G
HXAB_H	P31270	*Homo sapiens*	G
HXAD_H	P31271	*Homo sapiens*	G
HXB9_H	P17482	*Homo sapiens*	G
HXBD_H	Q92826	*Homo sapiens*	G
HXC9_H	P31274	*Homo sapiens*	G
HXCA_H	Q9NYD6	*Homo sapiens*	G
HXCB_H	O43248	*Homo sapiens*	G
HXCC_H	P31275	*Homo sapiens*	G
HXCD_H	P31276	*Homo sapiens*	G
HXDA_H	P28358	*Homo sapiens*	G
HXDB_H	P31277	*Homo sapiens*	G
HXDC_H	P35452	*Homo sapiens*	G
HXDD_H	P35453	*Homo sapiens*	G
CDX1_M	P18111	*Mus musculus*	G
CDX2_M	P43241	*Mus musculus*	G
CDX4_M	Q07424	*Mus musculus*	G
HXA9_M	P09631	*Mus musculus*	G
HXAA_M	P31310	*Mus musculus*	G
HXAB_M	P31311	*Mus musculus*	G
HXAD_M	Q62424	*Mus musculus*	G
HXB9_M	P20615	*Mus musculus*	G
HXBD_M	P70321	*Mus musculus*	G
HXC9_M	P09633	*Mus musculus*	G
HXCA_M	P31257	*Mus musculus*	G
HXCB_M	XP_111600.1	*Mus musculus*	G
HXCA_M	P31257	*Mus musculus*	G
HXCD_M	P50207	*Mus musculus*	G
HXD9_M	P28357	*Mus musculus*	G
HXDA_M	P28359	*Mus musculus*	G
HXDB_M	P23813	*Mus musculus*	G
HXDC_M	P23812	*Mus musculus*	G
HXDD_M	P70217	*Mus musculus*	G
CA_Tricas	CAA06527	*Tribolium castanaeum*	G
AB_Tricas	AAF36721	*Tribolium castaneum*	G
AB_Acankapu	AAB92404	*Acanthokara kaputensis*	N
CA_Acakap	AAB92405	*Acanthokara kaputensis*	N
AB_Achatepi	AAF73210	*Achaearanea tepidarorum*	N
HXA9_Amb	P50209	*Ambystoma mexicanum*	N
HXAD_Amb	P50210	*Ambystoma mexicanum*	N
HXBD_Amb	AAG27629	*Ambystoma mexicanum*	N
HxCA_Amb	AAG27631	*Ambystoma mexicanum*	N
CA_Anogam	AAD27585	*Anopheles gambiae*	N
CA_Bommor	BAA04086	*Bombyx mori*	N
HXA9_Cav	P51783	*Cavia porcellus*	N
CA_Chaevar	AAB16983	*Chaetopterus variopedatus*	N
AB_Cupsal	CAB40807	*Cupiennius salei*	N
CDX1_C	I50125	*Cyprinus carpio*	N
HXDA_Danaf	AAB38634	*Danio aff. albolineatus*	N

Appendix B (*continued*)

Abbreviation	Accession	Species	Category
HXDA_Dancf	AAB38636	*Danio cf. Tweediei*	N
HHDA_Dande	AAB38637	*Danio devario*	N
HXDA_Danfr	AAB38638	*Danio frankei*	N
HXDA_Danke	AAB38639	*Danio kerri*	N
HXDA_Danma	AAB38640	*Danio malabaricus*	N
HXDA_Danpa	AAB38642	*Danio pathirana*	N
HXDA_Danpu	AAB38641	*Danio pulcher*	N
CA_Distig	CAC19385	*Discocelis tigrina*	N
AB_Folcan	AAK52499	*Folsomia candida*	N
CA_Folcan	AAL78090	*Folsomia candida*	N
HXA9_F	O42506	*Fugu rubripes*	N
HXC9_F	O42506	*Fugu rubripes*	N
CDX_G	AAK38602	*Gallus gallus*	N
CDX1_G	S16417	*Gallus gallus*	N
CHOX_G	CAA48883	*Gallus gallus*	N
HXA9_G	Q98924	*Gallus gallus*	N
HXAB_G	P31258	*Gallus gallus*	N
HXAD_G	Q90X25	*Gallus gallus*	N
HXD9_G	P24340	*Gallus gallus*	N
HXDA_G	P24341	*Gallus gallus*	N
HXDB_G	P24342	*Gallus gallus*	N
HXDC_G	P24343	*Gallus gallus*	N
HXDD_G	P24344	*Gallus gallus*	N
CA_Halror	BAA85628	*Halocynthia roretzi*	N
CDX1_Hal	BAA85628	*Halocynthia roretzi*	N
HXA9_Het	Q9IA26	*Heterodontus francisci*	N
HXAA_Het	Q9IA27	*Heterodontus francisci*	N
HXD9_Het	Q9IA13.1	*Heterodontus francisci*	N
HXDA_Het	Q9IA14	*Heterodontus francisci*	N
HXDB_Het	Q9IA15	*Heterodontus francisci*	N
HXDC_Het	Q9IA16	*Heterodontus francisci*	N
HXDD_Het	AAF44637.1	*Heterodontus francisci*	N
CDX_Lin	P81193	*Lineus sanguineus*	N
AB_Lithatk	AAL36900	*Lithobius atkinsoni*	N
AB_Lithforf	AAK51950	*Lithobius forficatus*	N
CDX2_ME	Q04649	*Mesocricetus auratus*	N
HXA9_Mor	Q9PWD5	*Morone saxatilis*	N
HXAA_Mor	AAD46395	*Morone saxatilis*	N
HXDB_Not	P31263	*Notophthalmus viridescens*	N
HXA9_Ory	BAA85132	*Oryzias latipes*	N
HXC9_Sh	Q28601	*Ovis aries*	N
AB_Pachfer	CAB75738	*Pachymerium ferrugineum*	N
AB_PaurWye	AAK28143	*Pauropus* sp. Wye-1996	N
AB_Priapca	AAD40649	*Priapulus caudatus*	N
CA_Procla	AAK58679	*Procambarus clarkii*	N
HXDA_Pse	AAB38644	*Pseudorasbora cf. parva*	N
HXCA_Ras	AAB38647	*Rasbora paviei*	N
CDX1_R	Q05095	*Rattus norvegicus*	N
CA_Saccar	AAD00342	*Sacculina carcini*	N
CA_Schgre	AAK56940	*Schistocerca gregaria*	N
AB_Scutimma	AAG45178	*Scutigerella immaculata*	N
HXCA_Tan	AAB38648	*Tanichthys albonubes*	N
AB_Therdome	AAD50354	*Thermobia domestica*	N
HXA9_X	CAC44978	*Xenopus laevis*	N
HXAB_X	CAC44977	*Xenopus laevis*	N
HXB9_X	P31272	*Xenopus laevis*	N
HXCC_X	CAC44976	*Xenopus laevis*	N
HXd9_X	CAC44978	*Xenopus laevis*	N
HMDF_D	NP_477201	*Drosophila melanogaster*	O
HMUX_D	NP_536748	*Drosophila melanogaster*	O

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Balavoine, G., Telford, M.J., 1995. Identification of planarian homeobox sequences indicates the antiquity of most hox/homeotic gene subclasses. Proc. Natl. Acad. Sci. USA 92, 7227–7231.

Banerjee-Basu, S., Ferlanti, E.S., Ryan, J.F., Baxevanis, A.D., 1999. The homeodomain resource: sequences, structures and genomic information. Nucleic Acids Res. 27, 336–337.

Banerjee-Basu, S., Ryan, J.F., Baxevanis, A.D., 2000. The homeodomain resource: a prototype database for a large protein family. Nucleic Acids Res. 28, 329–330.

Banerjee-Basu, S., Sink, D.W., Baxevanis, A.D., 2001. The homeodomain resource: sequences, structures, DNA binding sites and genomic information. Nucleic Acids Res. 29, 291–293.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., Sonnhammer, E.L., 2002. The Pfam protein families database. Nucleic Acids Res. 30, 276–280.

Bremer, K., 1995. Branch support and tree stability. Cladistics 10, 295–304.

Bürglin, T.R., 1994. A comprehensive classification of homeobox genes. In: Duboule, D. (Ed.), Guidebook to the Homeobox Genes. Oxford University Press, New York, pp. 25–74.

Callaerts, P., Lee, P.N., Hartmann, B., Farfan, C., Choy, D.W., Ikeo, K., Fischbach, K.F., Gehring, W.J., de Couet, H.G., 2002. HOX genes in the sepiolid squid Euprymna scolopes: implications for the evolution of complex body plans. Proc. Natl. Acad. Sci. USA 99, 2088–2093.

Davis, J.L., Nixon, K.C., 1992. Populations, genetic variation, and the delimination of phylogenetic species. Syst. Biol. 41, 421–435.

de Rosa, R., Grenier, J.K., Andreeva, T., Cook, C.E., Adoutte, A., Akam, M., Carroll, S.B., Balavoine, G., 1999. Hox genes in brachiopods and priapulids and protostome evolution. Nature 99, 772–776.

Eriksson, T., 1997. Auto-Decay, version 2.9.9. Stockholm University, Stockholm.

Farris, J.S., 1983. The logical basis of phylogenetic analysis. Adv. Cladistics 2, 7–36.

Ferrier, D.E., Holland, P.W., 2001.

Hennig, W., 1963. Phylogenetic systematics. Annu. Rev. Entomol. 10, 97–116.

Kappen, C., 2000. Analysis of a complete homeobox gene repertoire: implications for the evolution of diversity. PNAS 97, 4481–4486.

Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P., Bork, P., 2002. Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res. 30, 242–244.

Planet, P.J., Kachlany, S.C., DeSalle, R., Figurski, D.H., 2001. Phylogeny of genes for secretion NTPases: identification of the widespread tadA subfamily and development of a diagnostic key for gene classification. Proc. Natl. Acad. Sci. USA 98, 2503–2508.

Ruvkun, G., Hobert, O., 1998. The taxonomy of developmental control in *Caenorhabditis elegans*. Science 282, 2033–2040.

Sharkey, M., Graba, Y., Scott, M.P., 1997. Hox genes in evolution: protein surfaces and paralog groups. Trends Genet. 13, 145–151.

Sikes, D.S., Lewis, P.O., 2001. PAUPRat: a tool to implement Parsimony Ratchet searches using PAUP*. University of Connecticut.

Swofford, D., 2002. PAUP* Software and Documentation, Version 4.0b10. Sinauer, Sunderland, MA.

Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families. Science 278, 631–637.

Telford, M.J., 2000. Turning Hox "signatures" into synapomorphies. Evol. Dev. 2, 360–364.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Thornton, J.W., DeSalle, R., 2000. Gene family evolution and homology: genomics meets phylogenetics. Annu. Rev. Genomics Hum. Genet. 1, 41–73.

Williams, P.L., Fitch, W.M., 1989. Finding the minimal change in a given tree. In: Fernholm, B., Bremer, K., Jurnvall, H. (Eds.), The Hierarchy of Life: Molecules and Morphology in Phylogenetic Analysis. Excerpta Medica, Amsterdam, pp. 453–470.