Phylogenetic Reconstruction of Vertebrate Hox Cluster Duplications

Wendy J. Bailey, Junhyong Kim, Günter P. Wagner, and Frank H. Ruddle
Department of Biology, Yale University

In invertebrates and the cephalochordate, amphioxus, the closest vertebrate relative, Hox genes are linked in a single cluster. Accompanying the emergence of higher vertebrates, the Hox gene cluster duplicated in either a single step or multiple steps, resulting in the four-cluster state present in teleosts and tetrapods. Mammalian Hox clusters (designated A, B, C, and D) extend over 100 kb and are located on four different chromosomes. Reconstructing the history of the duplications and its relation to vertebrate evolution has been problematic due to the lack of alignable sequence information. In this study, the problem was approached by conducting a statistical analysis of sequences from the fibrillar-type collagens (I, II, III, and V), genes closely linked to each Hox cluster which likely share the same duplication history as the Hox genes. We find statistical support for the hypothesis that the cluster duplication occurred as multiple distinct events and that the four-cluster situation arose by a three-step sequential process.

Introduction

Hox genes encode a family of helix-turn-helix transcription factors which operate in a regulatory cascade to specify, among other events, character identity along the anterioposterior body axis in the developing embryo (Shashikant et al. 1991; Hunt and Krumlauf 1992; McGinnis and Krumlauf 1992). Mammalian Hox genes are homologous to clustered homeobox genes found in the Antennapedia and the Bithorax complexes of Drosophila melanogaster (Duboule and Dolle 1989; Gra- ham, Papalopulu, and Krumlauf 1989). In both mouse and human, Hox genes are organized into four clusters (termed HOX A, B, C, and D, each extending over 100 kb) that are located on four different chromosomes (reviewed in Ruddle et al. 1987; Bürglin 1994). Linked genes have been assigned to 13 paralogous groups (referred to as cognate groups 1–13) based on the fact that Hox genes show more sequence similarity to paralogous genes on other clusters than to genes linked on the same cluster. A model of mammalian Hox gene evolution hypothesizes that the gene family arose through a series of gene duplications resulting in 13 paralogs linked in a cluster, followed by cluster duplication, which generated the current four-cluster state (Kappen, Schughart, and Ruddle 1989; Schughart, Kappen, and Ruddle 1989). Under this proposed pattern of evolution, cognate genes present on different clusters share a more recent common ancestor than genes within the same cluster, thus accounting for the higher level of sequence similarity observed among cognates on different clusters than among genes on the same cluster. The actual Hox gene complement of each cluster is unique, most likely due to gene loss after cluster duplication. The objective of this study is to reconstruct the historical events which led to the four-cluster situation in higher vertebrates.

Among the characters which distinguish the vertebrates from their closest invertebrate relatives is a pronounced differentiation of the axial organs. Cephalochordates, i.e., amphioxus and relatives, show very little axial differentiation. In contrast, all vertebrates have at least a strongly differentiated anterior end of the body axis, the head, in addition to the trunk and tail region already found in cephalochordates. Furthermore, higher vertebrates like mammals have at least six clearly differentiated axial body regions, which have been shown to be under the developmental control of Hox genes (Hunt and Krumlauf 1992; Burke et al. 1995). It is thus likely that the origin and elaboration of the vertebrate bauplan is intimately linked to the evolutionary history of Hox clusters. To test this possibility, it is necessary to reconstruct the evolutionary history of the Hox gene clusters.

To date, all invertebrates examined, including a cephalochordate (Kenyon and Wang 1991; Stuart et al. 1993; Garcia-Fernández and Holland 1994; Averof and Akam 1995), are known to possess a single Hox cluster, with the possible exception of Limulus polyphemus (Cartwright, Dick, and Buss 1993). Therefore, the four-cluster state must have arisen in the lineage leading from primitive chordates to higher vertebrates. There are two avenues to reconstruct the phylogenetic history of the Hox clusters: (1) comparison among the major groups of vertebrates to detect evidence of two- or three-cluster animals and (2) comparison among the four clusters of higher vertebrates.

There is little information about the existence and extent of Hox cluster and Hox gene number variation among recent vertebrate lineages. All tetrapods examined have four clusters with largely identical Hox gene complements (for a recent review, see Holland and Garcia-Fernández 1996). Recently, it was shown that teleosts also are four-cluster organisms (zebrafish: Hoeven et al. 1996; Misof, Blanco, and Wagner 1996; killifish: Misof and Wagner 1996). Based on sequence comparison, the four Hox clusters of the teleosts seem to be homologous to those in the tetrapods (Misof and Wagner 1996). Only fragmentary information is available for lower vertebrates like jawless vertebrates and cartilaginous fishes (Pendelton et al. 1993). Hence, the available

1 Present address: Department of Bioinformatics, Merck and Co. Inc., West Point, Pennsylvania.

Key words: molecular evolution, Hox, homeobox, collagen, vertebrate.

Address for correspondence and reprints: Junhyong Kim, Department of Biology, Yale University, P.O. Box 208103, New Haven, Connecticut 06520-8103. E-mail: junhyong_kim@quickmail.yale.edu.


843
information on interspecies variation is not sufficient to reconstruct the history of Hox cluster duplications.

The alternative approach, sequence comparison of paralogous genes from the same cognate group and the comparison of gene loss among Hox clusters, has been attempted repeatedly (Kappen and Ruddle 1993; Bürglin 1994; Meyer 1996; Zhang and Nei 1996). Most sequence comparisons only used the highly conserved homeobox sequence. But the 183 bp of the homeobox do not provide sufficient information for resolving the cluster duplication events (see Discussion).

In this paper, we extend the data by including alignments of sequences outside the homeobox including exon 1 sequences from groups four and nine. In addition, we take advantage of the fact that a number of multigene families are closely linked to the four mammalian Hox clusters (fig. 1). These genes are thought to share the same evolutionary history as the Hox clusters (Ruddle et al. 1994). A survey of these linked genes revealed that large orthologous chromosomal domains are conserved between mouse and human (Lundin 1993). Members of the fibrillar-type collagen gene family are among these linked genes and were included in our analyses with the Hox genes for the following reasons. Five fibrillar-type collagen genes are closely linked to each of the four Hox clusters. The collagen genes are approximately 4 kb in length, providing much more sequence information than the Hox genes, which have approximately 1 kb of coding sequence. The fibrillar-type collagens are divided into major (I, II, III) and minor (V, XI) types and are more closely related to each other than to the basement membrane types (IV) based on protein sequence, intron-exon structure, and function (Sakurai, Sullivan, and Yamada 1986; MacFie, Light, and Bailey 1988). Thus, the fibrillar-type collagen genes are classified as a distinct family of related proteins encoded by paralogous genes (Miller 1985; Ramirez et al. 1985; Vuorio and de Crombrugghe 1990).

In the following, we first present preliminary analyses which show the monophyly of the five Hox-linked collagen genes, lending support to the hypothesis that these genes duplicated in concert with the Hox genes. Next, in an attempt to reconstruct the mammalian cluster duplication event or multiple events, we test a null hypothesis which states that the four vertebrate clusters arose from a single cluster in a single event or several closely timed events. This would be expected if Hox genes evolved through a polyploidization event or through a series of rapid successive tetraploidizations. Although two or three independent events of cluster duplications seems to be the favored hypothesis of most authors (e.g., Holland and Garcia-Fernandez 1996), the alternative hypothesis of a star phylogeny has not been formally rejected. Here we present the first statistical test rejecting this possibility. In the next step, we attempt to reconstruct the tree topology with parsimony, neighbor-joining, and minimal-evolution tests. The results support an ((A,D)(B,C)) relationship among the clusters. Finally, we attempt to root the cluster topology by outgroup comparison to sponge and sea urchin fibrillar-type collagen sequences as outgroup using parsimony, neighbor-joining, and maximum-likelihood methods.

Materials and Methods

DNA Sequence Accession Numbers and Alignments

Mouse Hox 4, Hox 9 and human, sea urchin (*Strongylocentrotus purpuratus* pro-alpha2(1) collagen; Exposito et al. 1992), and sponge (*Ephydatia muelleri* alpha (COLFl); Exposito and Garrone 1990) collagen sequences were obtained through GenBank and/or EMBL. The accession numbers are as follows: HoxA4—X66861; HoxB4—X65950; HoxC4—S62287; HoxD4—J03770; HoxA9—M28449; HoxB9—S66855; HoxC9—X55318; HoxD9—X62669; Collagen1A1 M20789; Collagen1A2—J03464; Collagen2A1—...
TAL W (Thompson, Higgins, and Gibson 1994). All multiple sequence alignments were generated using CLUSTAL W (Thompson, Higgins, and Gibson 1994). All alignments were done based on the amino acid sequence. In the analysis based on nucleotide data, we used the nucleotide sequence corresponding to the aligned amino acids (the original sequence, not a translated sequence). The total number of nucleotides included in each alignment are as follows: chicken Hox 4 = 408 nt; mouse Hox 4 = 408 nt; mouse Hox 9 = 234 nt; human collagen = 2,595 - 3,648 nt (depending on the analysis); sea urchin and sponge collagen = 3,456 nt.

Human collagen genes were included in this study, because mouse collagen type I alpha 1 (COL1A1) and type III alpha 1 (COL3A1) sequences were not complete. However, the entire coding regions for all five human collagen genes linked to the Hox clusters were available from GenBank or EMBL. Since the human and mouse orthologs are more similar to each other than are the paralogs (results not shown), the collagen genes must have duplicated prior to the speciation event, resulting in separate human and mouse lineages. Therefore, sequences from either human or mouse provide the same information for the purpose of reconstructing the Hox cluster duplication event. The initial collagen alignment was generated using all available fibrillar-type collagen genes COL1A1 (linked to HOXB), COL1A2 (linked to HOXA), COL2A1 (linked to HOXC), COL3A1 and COL5A2 (linked to HOXD; Emanuel et al. 1985), COL5A1 (unlinked), COL11A1 (unlinked), and COL11A2 (unlinked). Because of the distance of the unlinked collagen genes, a very conservative alignment was chosen, deleting the variable regions and indel regions. About 2/3 of the nucleotide positions were variable in these alignments. A second alignment was constructed using only COL1A1, COL1A2, COL2A1, and COL3A1 for the purpose of inferring the Hox cluster duplication events. A final alignment included COL1A2, COL1A1, COL2A1, COL3A1, sea urchin, and sponge sequences. All alignments included the triple helical collagenous domain, the C-telopeptide domain, and the C-propeptide domain. The alignments are deposited in EMBL accounts DS 28913 and DS 29386.

Monophyly of Hox-Linked Collagen Genes and the Selection of COL3A1

If the Hox-linked collagen genes arose through the same duplication events that led to the Hox cluster duplications, we would expect the Hox-linked collagen genes to form a monophyletic clade. This possibility was tested using phylogenetic analyses of a total of eight different fibrillar-type collagen sequences. These sequences were analyzed using: protein pasimony with codon-based step matrix (PAUP version 3.1; Swofford 1991) and a modified step matrix in PROTPARS program of PHYLIP 3.572 (Felsenstein 1993); unweighted unordered parsimony on the nucleotide sequences; neighbor-joining analysis using PAM matrix derived categorical protein distance (PHYLIP and MEGA1.01; Kumar, Tamura, and Nei 1994); and a maximum-likelihood analysis using a Hidden Markov model with three rate categories and the relative rates estimated using codon positions (PHYLIP). Each of these analyses included 1,000 bootstrap replicates except for the Hidden Markov model maximum-likelihood analysis, which used 250 replicates.

The linkage of COL1A2, COL1A1, and COL2A1 can be uniquely assigned to Hox clusters A, B, and C, respectively. However, both COL3A1 and COL5A2 are linked to Hox cluster D. The monophyly test indicated that COL3A1 was the best candidate for representing Hox cluster D (see Results), but we also tested to see whether a four-cluster tree using COL3A1 or a four-cluster tree using COL5A2 had more robust support. This was done using the same alignment as the monophyly test, but using only the COL1A2, COL1A1, COL2A1, and COL3A1 genes or COL1A2, COL1A1, COL2A1, and COL5A2 genes (that is, two four-taxon trees). The same phylogenetic analyses as above were done with 1,000 replicates for all methods.

Likelihood Analysis for the Hox Cluster Duplication History

A log-likelihood ratio test was applied to determine the significance of the length of the internal branch separating the four clusters. The null hypothesis (H0), that the internal branch length is zero, implies a simultaneous multiplication of all four clusters, while the alternative hypothesis (H1), that the internal branch length is positive, implies a sequential duplication of the four clusters. For each data set, the maximum-likelihood (ML) values were computed for H0 and H1 using a numerical optimization program (using Powell's method; Press et al. 1986) under a four-state six-parameter symmetric transition model (the most general time-reversible model; Swofford et al. 1996). The ancestral frequency of nucleotides was set to the average frequency of the current clusters. For the H1 model, the ML value for each of the three possible trees was computed, and the highest value was selected. For the H0 model, numerical optimization was carried out with the internal branch length constrained to zero.

The test statistic is the log odds ratio -2log(L(H0)/L(H1)). This statistic is approximately chi-square distributed when the parameter space for the alternative hypothesis contains the parameter space for the null hypothesis (Lehmann 1986, pp. 480-490). In our case, this condition is not exactly met because we do not allow negative branch length. However, this makes the significance value conservative. This is because the chi-square approximation gives a probability value of x for a test statistic value y, assuming negative-length branches are possible. But when the branch lengths are constrained to be positive, the actual probability density of the statistic value y is lower than the reported x because of this constraint.
For a four-cluster analysis under the $H_0$ model, there are five total branches. Nucleotides from each branch are allowed to change, unconstrained, with six free parameters; therefore, the model has $6 \times 5 = 30$ total parameters. Under the $H_0$ model, we assume the internal branch length is zero; therefore, the model has $6 \times 4 = 24$ free parameters. The difference in the two numbers $(30 - 24 = 6)$ is the appropriate degrees of freedom for the chi-square approximation.

Phylogenetic Analysis of the Hox Clusters

Three different methods were employed to estimate phylogenies of the four clusters: a distance-based algorithm; neighbor-joining (Saitou and Nei 1987), implemented in MEGA version 1.01 (Kumar, Tamura, and Nei 1994) and PHYLIP 3.572 (Felsenstein 1993); and the maximum parsimony method, implemented in PAUP version 3.1 (Swofford 1991). The neighbor-joining analysis included bootstrapping with 1,000 replications. Hox 4 and Hox 9 were analyzed using the Jukes and Cantor correction for superimposed mutations, because they had no transversion/transversion (TV/TS) bias and an even GC ratio (Jukes and Cantor 1969). The collagen distance values were corrected using Tamura-Nei, because they had a $2:1$ TV/TS and a GC ratio bias (Tamura and Nei 1993). A gamma correction was applied to the combined data sets, allowing for varying rates of evolution at different positions. This was done by applying a range of gamma parameters (0.2, very variable rates, 0.5, 1.0, 1.5, and infinity, i.e., no variation). The entire range of gamma corrections yielded the same topologies. For maximum-parsimony, all gaps were excluded from the analysis, and each position was weighted equally. An exhaustive search was run to generate all possible trees for all analyses. The parsimony analysis included bootstrap resampling with 1,000 replications. The amino acid analysis was run using a weighted protein parsimony step matrix. The tree with the two outgroup sequences was analyzed with the maximum-parsimony, neighbor-joining, and maximum-likelihood methods. For the amino acid maximum-parsimony method, a codon-based protein parsimony step matrix was used, while the neighbor-joining method used PAM weighted distances and protein categories distances (PROTDIST program in PHYLIP). The maximum-likelihood tree was obtained using the Hidden Markov model in DNAML with three categories (for each codon position) with the relative rates estimated from the number of inferred parsimony changes for each codon position. Each rate category was considered to be independent of the position.

Minimal-Evolution Tests of the Hox Clusters

The minimal-evolution (ME) method assumes that the phylogenetic tree with the smallest estimated branch length ($S$) represents the true evolutionary relationships. The ME method allows a test of whether the topology with the best fit is actually significantly better than an alternative topology by comparing the branch length estimates of the two topologies. We used the four-taxon analysis described by Rzhetsky, Kumar, and Nei (1995) as implemented in the PHYLTEST v.2.0 program (Kumar and Rzhetsky 1995). The statistical significance of the branch length differences $S_1 - S_2$ was evaluated using a normal distribution table as described by Kumar and Rzhetsky (1996). All data sets with the exception of the combined MHox4 + MHox9 data set was tested with both the Jukes and Cantor (1980) distance correction for superimposed mutations which allows for rate differences between transitional and transversional substitutions. The MHox4 + MHox9 data set was tested with both the Jukes and Cantor (1969) and the Kimura corrections.

Results

Monophyly of Hox-Linked Collagen Genes and the Selection of COL3A1

As mentioned above, the collagen genes COL1A2, COL1A1, and COL2A1 are linked to Hox clusters A, B, and C, respectively, while COL3A1 and COL5A2 are linked to the D cluster. In every analysis, the monophyly of these five Hox-linked genes was supported by 100% bootstrap support, unambiguously establishing this clade (fig. 2). In addition, all of the methods except for one resulted in support for the monophyly of the COL1A2, COL1A1, COL2A1, and COL3A1 clade (fig. 2; from 92% bootstrap support for protein distance neighbor-joining to 63% support for Hidden Markov maximum-likelihood. Only the codon-based step matrix parsimony (using PAUP) yielded the tree topology (((COL1A1, COL2A1, COL3A1), COL5A2), COL1A2) with 100% bootstrap support for all branches. When the two four-taxon trees (COL1A2, COL1A1, and COL2A1 with either COL3A1 or COL5A2) were analyzed, the tree with COL3A1 had higher bootstrap support for all of the analyses except for the Hidden Markov maximum-likelihood method, which had higher support for the tree with COL5A2. Despite this, given the results from the monophyly analyses and the better support in the four other methods, the results indicated that COL3A1 is a better representative of Hox cluster D evolution than COL5A2 (see also Discussion).
Maximum-Likelihood Test for an Internal Branch Length >0

Two different hypotheses for Hox cluster evolution were tested: \( H_0 \); All four clusters of the vertebrate genome originated as a single event or a rapid succession of more events, and \( H_1 \); The four clusters originated from multiple duplication events separated by significant amounts of time. The null hypothesis (\( H_0 \)) implies that the gene tree of the clusters would be a "star-like" topology, while \( H_1 \) implies that the gene tree would have resulted in a significant non-zero-length internal branch. Since \( H_1 \) allows three possible alternative trees for four clusters, ML values were computed for each.

ML values were calculated for \( H_0 \) and \( H_1 \) using five independent sequence alignments, chicken HoxA4-D4, mouse HoxA4-D4, mouse HoxA9-D9, and human collagens (COL1A1 = HOXB; COL1A2 = HOXA; COL2A1 = HOXC; COL3A1 = HOXD) (table 1). The result of the mouse Hox9 log-likelihood ratio test could not reject the null hypothesis at a 95% confidence level (\( P = 0.115 \); table 1), but chicken Hox 4 and Mouse Hox 4 were both significant, with \( P = 0.012 \) and \( P = 0.037 \), respectively. Furthermore, analysis of the collagen data resulted in the rejection of the \( H_0 \) hypothesis with a highly significant \( P \) value whether the COL3A1 gene or the COL5A2 gene was used to represent the HOXD cluster (\( P = 9.8 \times 10^{-12} \), \( P = 1.6 \times 10^{-12} \), respectively).

Next, the ML values were combined in two ways. First, the separate ML values for each gene (MHOx4, MHOx9, and COL3A1) were summed, and the log-likelihood ratios were calculated based on the summed value. This approach assumes that each gene evolves with independent rates of change. The summed ML value calculation resulted in significant support for the \( H_1 \) model, with \( P = 2.6 \times 10^{-9} \). Next, the data sets (MHOx4, MHOx9, HCOL3A1) were concatenated, and a single ML value was computed. This approach assumes that the genes share the same rate of evolution. Combining the mouse Hox 4 and Hox 9 sequences resulted in rejection of the null hypothesis in favor of the \( H_1 \) model with \( P = 6.9 \times 10^{-3} \) (table 1). Combining the separated MHOx4, MHOx9, and HCOL3A1 alignments into one data set gave significant support for the \( H_1 \) model (\( P = 1.9 \times 10^{-14} \)) as well (table 1).

Maximum-Parsimony Analysis of Hox Cluster Topology

Maximum-parsimony analysis was conducted applying the exhaustive search in PAUP version 3.1. The analysis was conducted on both amino acid sequences using the codon-based step matrix as well as with unweighted nucleotide sequences. All of the Hox data sets failed to provide resolution either by two minimum-length trees being generated or by the minimum-length tree differing from the next longest tree by a single mutational change (data not shown). When the D cluster was represented by the human collagen COL3A1, a consistent pattern arose. Both the nucleotide and the amino acid data yielded an ((A,D)(B,C)) minimum-length tree, with 58% bootstrap support for the amino acid data and 96% bootstrap support for the nucleotide data (table 2). The minimum-length tree for the combined Hox 4, Hox 9, and COL3A1 sequences also strongly supported the ((A,D)(B,C)) tree.

Neighbor-Joining Analysis

Neighbor-joining analysis was run on the individual Hox 4, Hox 9, collagen, and combined alignments (table 2). The favored topology is indicated by the symbols in the corresponding topology columns for each data set in table 2. Bootstrap support with 1,000 replications for the Hox data was weak and supported various topologies (MHOx4 = 44%; MHOx9 = 52%; MHOx4+9 = 46%–50%). As with the maximum-parsimony analysis, the Hox data failed to provide significant resolution. In contrast, the bootstrap values were 99% and 88%–100% for an ((A,D)(B,C)) topology with the collagen COL3A1 and combined MHOx4 + MHOx9 + COL3A1 data sets, respectively.

Minimal-Evolution Tests

The minimal-evolution test of Rzhetsky, Kumar, and Nei (1995) was conducted to see whether the support for the ((A,D), (B,C)) topology is significantly bet-
ster than for either alternative topology (table 3). The results confirm the results from the previous section in that only the collagen genes contain sufficient information to yield a significant result. The analysis of human collagens with COL3A1 representing cluster D yields the ((AD),(B,C)) topology as the best fit, and shows significant differences from both alternative topologies. The combined data set MHox4 + MHox9 + COL3A1 also resulted in significant support for the ((AD),(B,C)) topology. Hence, data sets with COL3A1 are the only data sets which favor only one topology over both alternatives in any data set. From this, we conclude that the ((AD),(B,C)) topology is the best model for the phylogenetic relationships among the mammalian Hox gene clusters.

Outgroup Comparison and Rooting
In this analysis, we included two additional collagen sequences, alpha2(I) from a sea urchin (Strongylocentrotus purpuratus) and alpha (COLF1) from a sponge (Ephydatia muelleri) as outgroups to estimate the root position of the four clusters. The sequences were aligned from the amino acids and the phylogeny was estimated using codon-based step matrix maximum-parsimony, the neighbor-joining method using PAM protein distance and Categories model protein distance (PHYLIP 3.572), and the Hidden Markov maximum-likelihood model with three rate categories (corresponding to three codon positions) and independent probability of rate variation (PHYLIP 3.572). The three rate categories parameters were estimated using empirical observed changes from the parsimony reconstructions. All three methods gave the same topology shown in figure 3. The results of 1,000 bootstrap analyses for protein parsimony and neighbor-joining are also shown in figure 3, which supports the D-(A, B, C) internal branch at 96%-100% and A-(B, C) branch at 85%-99.5%. The estimated tree supports a sequential cluster expansion model where the ancestral lineage initially gives rise to the lineages leading to clusters D and A, followed by another duplication from the A lineage to the ancestor of clusters B and C, followed by a third duplication event leading to the B and C lineages.

Discussion
Orthologous DNA Sequences
The purpose of this study was to reconstruct the Hox cluster duplication events that occurred along the descent of vertebrates. Although all vertebrates examined to date possess four Hox gene clusters, only two cognate groups, Hox 4 and Hox 9 (Hox 13 has recently been shown to have four cognates; Zeltser et al. 1996), are known to have all four genes present in extant mammals, HoxA4, B4, C4, D4, and HoxA9, B9, C9, D9 (fig. 1). Because the Hox cluster duplication event is more

Table 2
Phylogenetic Comparison of Alternative Branching Orders Among Hox Clusters

<table>
<thead>
<tr>
<th>Data Set</th>
<th>BEST TREE</th>
<th>ALTERNATIVE TREES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse Hox4</td>
<td>((A,B),(C,D))</td>
<td>I CP 77 (%) II CP 61 (%)</td>
</tr>
<tr>
<td>Mouse Hox9</td>
<td>((A,B),(C,D))</td>
<td>((A,C),(B,D))</td>
</tr>
<tr>
<td>MHox4 + MHox9</td>
<td>((A,B),(C,D))</td>
<td>((A,C),(B,D))</td>
</tr>
<tr>
<td>MHox4 + 9*</td>
<td>((A,B),(C,D))</td>
<td>((A,C),(B,D))</td>
</tr>
<tr>
<td>MHox4 + 9*</td>
<td>((A,B),(C,D))</td>
<td>((A,C),(B,D))</td>
</tr>
</tbody>
</table>

# Table 3
Comparison of Alternative Branching Orders Among Hox Clusters Using Four-Cluster Analysis

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>BEST TREE</th>
<th>I CP 77 (%)</th>
<th>II CP 61 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHox4</td>
<td>((A,B),(C,D))</td>
<td>((A,C),(B,D))</td>
<td>77 (%) ((A,D),(B,C))</td>
</tr>
<tr>
<td>MHox4</td>
<td>((A,B),(C,D))</td>
<td>((A,C),(B,D))</td>
<td>7.2 (%) ((A,D),(B,C))</td>
</tr>
<tr>
<td>MHox9</td>
<td>((A,B),(C,D))</td>
<td>((A,C),(B,D))</td>
<td>47 (%) ((A,D),(B,C))</td>
</tr>
<tr>
<td>MHox4 + 9*</td>
<td>((A,B),(C,D))</td>
<td>((A,C),(B,D))</td>
<td>51 (%) ((A,D),(B,C))</td>
</tr>
<tr>
<td>MHox4 + 9*</td>
<td>((A,B),(C,D))</td>
<td>((A,C),(B,D))</td>
<td>48 (%) ((A,D),(B,C))</td>
</tr>
</tbody>
</table>

a HOXA cluster (A), HOXB cluster (B), HOXC cluster (C), HOXD cluster (D).
b Confidence probabilities (CP) are expressed as 100 × (1 - F)% where F is the probability value obtained in a two-tailed normal deviate test.
c Jukes and Cantor correction factor.
d Kimura correction factor.
ancient than the divergence of amniotes (250 Myr), we used available sequences from human, mouse and chicken. The entire coding sequence for the mouse and chicken Hox 4 genes, approximately 1 kb, was available, but only about 400 bp of these sequences could be accurately aligned. The complete sequence for mouse HoxA9 was unavailable, so 245 bp total aligned sequence was used from the mouse Hox 9 group for the analysis.

The fibrillar-type collagen I (COL1A1, COL1A2), II (COL2A1), III (COL3A1), and V (COL5A1, COL5A2, COL5A3) genes are very similar in structure and function and are considered to be evolutionarily closely related (Solomon and Cheah 1981; Bernard et al. 1983; Vuorio and Crookshagge 1990). Homology of the collagen genes is based on biochemical criteria (Pikkarainen and Kulonen 1969; MacFie, Light, and Bailey 1988). On DNA and protein sequence comparisons (Hofmann, Fietzek, and Kuhn 1980; Sakurai, Sullivan, and Yamada 1986; Weil et al. 1987), and on structural and functional similarities (McLachlan 1976; Bornstein and Sage 1980; Eyre 1980). All members of the major fibrillar-type collagen genes (I, II, and III) and one member of the minor type V (5A2) are each linked to one of the Hox clusters in both human and mouse (table 4). In humans, four out of five fibrillar-type collagen genes remain in close linkage to the Hox clusters with the exception of collagen type I alpha 2 (COL1A2). COL1A2 has been moved by chromosomal inversion to the opposite arm of chromosome 7 from the Hox A cluster; all other fibrillar-type collagen genes (COL1A1, COL2A1, COL3A1, COL5A2) map within the same chromosomal band as the Hox clustered genes (table 4). In addition, as mentioned previously, phylogenetic analyses of all available fibrillar-type collagen sequences (the Hox-linked collagens and COL5A1, COL11A1, and COL11A2) strongly support the monophyly of Hox-linked collagens (100% bootstrap support).

The inversion breakpoint on the p arm of chromosome 7 is likely proximal to the EVX1 locus which remains in linkage to the Hox A cluster located at 7p15–p14 (fig. 1). The mouse collagen type I alpha 2 (COL1A2) ortholog is linked to the Hox A cluster on chromosome 6 (table 4) along with a number of other genes that in human, have remained in linkage to the COL1A2 locus on chromosome 7q21.3–q32 (fig. 1) (Nadeau et al. 1992). This suggests that the inversion found in humans occurred more recently than the duplication event which

![Image of phylogenetic analysis](image_url)

**Fig. 3**—A rooted phylogenetic analysis of the Hox-cluster-linked collagen genes. The Hox clusters were represented using COL1A2 = cluster A, COL1A1 = cluster B, COL2A1 = cluster C, COL3A1 = cluster D, and two outgroup fibrillar-type collagen genes from a sea urchin (*Strongylocentrotus purpuratus*) and a sponge (*Ephydatia muelleri*). The tree was estimated using protein parsimony, the neighbor-joining method using protein distances, and a Hidden Markov model maximum-likelihood method. All three methods gave the same topology. The tree is drawn using branch length estimated with the maximum-likelihood method. Bootstrap values on 1,000 replicates for protein parsimony and neighbor-joining (in parentheses) are indicated by the numbers at the internal branches.

### Table 4

Map Locations of Collagen and Hox Genes

<table>
<thead>
<tr>
<th>Human Chromosomes</th>
<th>Reference(s)</th>
<th>Mouse Chromosomes</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collagen 1A*</td>
<td>17q21–q22</td>
<td>11</td>
<td>Buchberg et al. (1989); Munke et al. (1986)</td>
</tr>
<tr>
<td>Hox B</td>
<td>17q21–q22</td>
<td>11</td>
<td>Joyner et al. (1985); Do and Lonai (1988)</td>
</tr>
<tr>
<td>Collagen 1A2</td>
<td>7q21–q22a</td>
<td>6</td>
<td>Irving et al. (1989); Sundaresan and Francke (1989)</td>
</tr>
<tr>
<td>Hox A</td>
<td>7p15–p14</td>
<td>6</td>
<td>Rabin et al. (1986)</td>
</tr>
<tr>
<td>Collagen 2A1</td>
<td>12q12.13</td>
<td>15</td>
<td>Cheah et al. (1991)</td>
</tr>
<tr>
<td>Hox C</td>
<td>12q12–13</td>
<td>15</td>
<td>Rabin et al. (1985)</td>
</tr>
<tr>
<td>Collagen 3A1</td>
<td>2q31–q32.3</td>
<td>1c</td>
<td>Schurr et al. (1990)</td>
</tr>
<tr>
<td>Collagen 5A2</td>
<td>2q32–q33</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td>Hox D</td>
<td>2q31</td>
<td>2</td>
<td>Stubbs et al. (1990)</td>
</tr>
</tbody>
</table>

*NOTE.*—The unlinked collagen genes' human map locations are as follows: COL5A1 (unknown), COL11A1 (1p21), and COL11A2 (6p21.2).

*COL1A1 is known to be linked within 180 kb of the Hoxb13 gene (Zelser et al. 1996).

*COL1A2 is located on the opposite chromosomal arm (7q) from the HOXA cluster locus in human, but the mouse homolog is linked to the HOXA cluster on mouse chromosome 6.

*COL3A1 is located on chromosome 1 in mouse within a large linkage group which is conserved with human chromosome 2q31–qter.
generated multiple copies of evolutionarily related genes including the Hox, collagens, even-skipped, wingless, interleukin receptors, integrins, nerve growth factor receptors, and ErbB tyrosine kinases (Schughart, Kappen, and Ruddle 1989; Hart et al. 1992; Laudet et al. 1992; Ruddle et al. 1994; Bentley, Bradshaw, and Ruddle 1995; Wang et al. 1995; Zimonjic et al. 1995).

Both COL3A1 and COL5A2 are closely linked to the HOXD cluster in human (table 4). The map location of mouse col5a2 is unknown and may be closely linked to col3a1 as well. The pattern of the charged groups in the HOXD cluster in human (table 4). The map location of collagens and codon bias that COL5A2 most likely arose to col3a1 as well. It has also been speculated that COL5A2 arose prior to the (COL3A1, COL5A2) clade. It has been suggested based on sequence comparisons and codon bias that COL5A2 are closely linked to a duplication of COL3A1 (Weil et al. 1987). However, this is not consistent with our phylogenetic analyses, which do not support a (COL3A1, COL5A2) clade. This has also been speculated that the duplication resulting in the type V collagen genes likely occurred after the duplication events which gave rise to the Hox clusters and the type I, II, and III collagens (Fessler, Shigaki, and Fessler 1985; Eyre and Wu 1987). Our phylogenetic analyses tentatively suggest that COL5A2 is a sister taxon to the other four Hox-linked genes. This suggests either that type V collagens arose prior to type I, II, and III collagens or that the other fibrillar-type collagen genes are later derived from the COL5A2.

In summary, the evidence supporting the hypothesis that fibrillar-type collagens share phylogenetic history with the Hox clusters is the following: (1) The collagens are part of a larger conserved linkage group which includes the Hox clusters as well as many other genes. (2) In humans, the distant location of COL1A2 relative to the HOX cluster A can be traced to a translocation event after the split of lineages leading to human and mouse. (3) All Hox-linked collagens form a monophyletic gene group within all the fibrillar-type collagen genes in the mammalian genome. This indicates that the Hox-linked collagens are not translocated into this position out of an older gene family.

History of Hox-Cluster Evolution

The vertebrate genome likely has risen through a series of genome duplications (Ohno 1970, 1993) which might be responsible for the increase in Hox cluster number from one to four (Holland et al. 1994; Ruddle et al. 1994). Combining sequence information from the fibrillar-type collagen genes with that from the Hox genes allowed us to reject the hypothesis that the four-cluster state was reached in one duplication event (star topology) or a rapid succession of more events. This result implies that during early vertebrate evolution, a history of multiple duplications was responsible for creating the mammalian four-cluster state.

Maximum-likelihood, maximum parsimony, and neighbor-joining analyses and minimal-evolution tests supported grouping Hox clusters A with D and B with C. Using sea urchin and sponge fibrillar-type collagen sequences as the outgroup, the root is placed with the D cluster, leading to a (OG(D(A,B,C) topology. A midpoint rooting of the longest path in the tree (using maximum-likelihood estimated branch length) also puts the root along the sponge branch. This result suggests that there were three steps, 1 → 2 → 3 → 4 clusters, in the evolution of the four-cluster situation in higher vertebrates rather than two, i.e., 1 → 2 → 4 clusters. The problem with the historical interpretation of the three-step hypothesis is that not all three steps can be associated with the maintenance of all clusters produced by genome duplications. A genome duplication doubles the number of clusters, and three rounds of genome duplication should yield 2^3 = 8 clusters. This implies that if the cluster duplications were caused by genome duplications, four clusters have been lost. Alternatively, cluster duplications may be due to chromosome duplications, such that each event only duplicates one of the clusters. In addition, the possibility of chance rooting exists because the internal branch separating (A,D) from (B,C) is short. However, the high bootstrap support for the three-step hypothesis makes the three-step scenario highly likely.

Previous phylogenetic analysis based on homeobox nucleotide sequences and a cladistic analysis using the presence and absence of paralogous Hox genes as character states failed to provide resolution of the Hox cluster relationships (Kappen and Ruddle 1993). Recently, a study included amino acid sequences from the Droso phila abdominal B (AbdB) gene and its homologues in amphioxus and mouse, the Hox 9 genes, in a PAUP analysis and found clusters C and D groupings together, joined next by A, then B (Meyer 1996). However, the bootstrap values over 100 replications were weak ((D,C) = 57, ((C,D)A) = 46). Another study primarily focused on the relationship among HOM/HOX paralogs linked within a cluster but did address the cluster duplication using 60 amino acids from the homeodomain (Zhang and Nei 1996). NJ analysis could not distinguish among three possible topologies ((B(A(C,D)), ((A(B(C,D)), or ((A,B)(C,D)). These studies exemplify the limited scope of data sets comprised of only the 183 bp homeobox or the 61 amino acids in the homeodomain. In this report, we show that even the complete alignable coding sequences of Hox genes groups 4 and 9 do not favor one topology over another. However, analyzing the 3.6-kb collagen sequences or the combined Hox/collagen data sets did provide resolution of the cluster topology (see above).

The analysis of molecular data presented here suggests that the phylogenetic history transforming a one-cluster system, as represented by amphioxus, into the four-cluster system, as found in higher vertebrates was a complex multistage process, including at least two to three duplication events and possibly cluster loss. A comparison of the Hox genes of the recent jawless vertebrates (lamprey and hagfish) and the cartilaginous fishes (sharks, skates, and ratfish) is essential to rigorously test the possible association of Hox cluster duplications and the origin of grades of morphological organization.

Acknowledgments

We thank Kevin Bentley for his insightful contributions to the manuscript and Cliff Tabin and his co-
workers for providing us with unpublished chicken Hox sequences. The authors are grateful to Gavin Naylor for reading an earlier version of this paper, and we thank Jacques Gautier, Jon Moore, and John Maisey for discussing early chordate evolution with us. An anonymous reviewer greatly helped to improve the second revision. G.P.W. thanks Jan Tashchner for spending 3 days formatting the manuscript. Financial support by NIH grant GM09966 to F.H.R. and NSF grant IBN-9507466 to G.F.W., as well as NSF grant IBN-9630567 to F.H.R. and G.P.W. are gratefully acknowledged. J.K. was supported, in part, by a Sloan Foundation Young Investigator Award, and W.J.B. was supported by a Sloan Molecular Evolution postdoctoral fellowship.

LITERATURE CITED


HUERRE, J. C., I. HENRY, M. BERNARD, P. GALLANO, D. WEIL, K. H. GRZESCHIK, F. RAMIÈRE, and C. JUNIEN. 1986. The pro alpha 2(V) collagen gene (COL5A2) maps to 2q14-


Shozo Yokoyama, reviewing editor

Accepted April 21, 1997