Two steps in the evolution of Antennapedia-class vertebrate homeobox genes

(nucleotide sequence comparisons/duplications/multigene family)

CLAUDIA KAPPEN*, KLAUS SCHUGHART*, AND FRANK H. RUDDLE*†

Departments of *Biology and †Human Genetics, Yale University, New Haven, CT 06511

Contributed by Frank H. Ruddle, April 21, 1989

ABSTRACT Antennapedia-class vertebrate homeobox genes have been classified with regard to their chromosomal locations and nucleotide sequence similarities within the 183-base-pair homeobox domain. The results of these comparisons support the view that in mammals and most likely the vertebrates, four clusters of homeobox genes exist that were created by duplications of an entire primordial gene cluster. We present evidence that this primordial cluster arose by local gene duplications of homeoboxes that were present before the divergence of arthropods and chordates. Sequence analyses indicate that the expansion of the primordial gene cluster complex was accompanied by diversification, whereas conservation predominated after the duplications of entire homeobox gene clusters.

Homeobox genes were identified initially in *Drosophila* where they encode products that regulate embryonic development (reviewed in refs. 1 and 2). Each of these genes contains the conserved 183-base-pair (bp) homeobox sequence that encodes a 61-amino acid homeodomain. More than 65 homeobox-containing genes have been isolated from several different species of vertebrates (for review, see ref. 3). Sequence characteristics permit the subdivision into an Antennapedia-class, engrailed-class, or other subclasses of homeoboxes (3). Antennapedia-class homeobox genes are located in clusters on chromosomes 2, 6, 11, and 15 of the mouse and on chromosomes 2, 7, 17, and 12 of the human (reviewed in refs. 4 and 5). We have focused on vertebrate homeobox genes of the Antennapedia-class in an attempt to determine if similarities in these homeoboxes can be used to ascertain evolutionary relationships within the Antennapedia-class gene family.

In this report, we present evidence that Antennapedia-class homeobox genes of vertebrates evolved in two distinct steps. The first involved an expansion of genes within one linkage unit. The second occurred by the duplication of the expanded linkage group. The overall result is the existence of a multigene family distributed as gene clusters on at least four chromosomes. These two steps appear to have been subjected to different selective pressures during the course of evolution, since the first phase is characterized by rapid sequence divergence, whereas the second is relatively conservative.

MATERIALS AND METHODS

Nucleotide Sequences. Nucleotide sequences were taken from the literature. These data were supplemented by the sequences of the mouse homeobox genes *Hox-3.3* (6) and *Hox-2.5* (7). Unpublished sequences of human homeobox genes were kindly provided by E. Boncinelli (International

Institute of Genetics and Biophysics, Naples), and A. Ferguson-Smith (*HOX-1.4*; our laboratory). The homeobox *Lox-1* of the lamprey was recently sequenced in our laboratory (J. Pendleton, personal communication). The complete list of nucleotide sequences of vertebrate, Antennapedia-class homeoboxes, with references is available upon request.

Comparisons Between Sequences. The alignment of nucleotides was dictated by the reading frame since all sequences encode proteins. No gaps were allowed. All sequences and groups of sequences were compared to each other pairwise to create distance matrices, using a program kindly provided by C. Stephens (Hughes Human Gene Mapping Library, New Haven, CT). Based on the list of 65 vertebrate Antennapedia-class homeobox sequences, the overall incidence of nucleotides at each particular site was investigated to determine the basic parameters for comparisons. A total number of 134 positions of 183 are occupied by different nucleotides and, conversely, 49 positions are identical in all sequences. Of the 134 variable positions, 58 can be affected by silent substitutions not altering the amino acid sequence, codon changes not taken into account. Mouse homeobox sequences or human homeobox sequences differ from each other within each species in up to 85 positions. A maximum of 34 silent differences was found between two mouse sequences. Restrictions on the introduction of nucleotides at silent sites could not be identified, since each possible nucleotide was observed at least once. Thus, when applying Poisson statistics to the above figures, and assuming that each position is equally susceptible, it emerges that each site was hit about once. Therefore, only a small fraction of the counted differences between two sequences would represent multiple hits per site so that we did not correct for these events. Replacement substitutions were determined by hand, for each pair of sequences, as the minimum number of changes between the particular codons that lead to the observed change in amino acids.

Classification of Nucleotide Sequences. A scheme for the chromosomal organization of mouse homeobox genes has been developed (4, 8). The positions of human homeoboxes on the respective clusters are taken from ref. 9 supplemented by *HOX-2.8* and *HOX-2.9*, which are located downstream from *HOX-2.7* (E. Boncinelli, personal communication). The sequence of *K8* differs from that of *HOX-2.8* at only one nucleotide and was therefore considered to represent the same homeobox. The *Xenopus* homeoboxes *MM3* and *XIHbox2* are believed to represent the same homeobox gene as are *XIHbox3* and *Xhox36*, and *XIHbox4* and *Xhox1-B* (10).

Homeoboxes for which the chromosomal locations are known were used as reference sequences for the classification of boxes that are unassigned as of this writing. Nucleotide sequences were grouped together according to similarities revealed by a distance matrix for all 65 sequences (see below). However, for some classifications, additional information was taken into account. *XIHbox3/XHox36* and *MM3/XIHbox2* were assigned to their respective mouse cognate on the basis of similarities in amino acid sequences and coding

sequences that flank the homeoboxes. Molecular data indicate that *XHox-1B* and *XHox-1A* are located close to each other (11) and, thus, can be classified into neighboring cognate groups. Molecular cloning data (12) were also used to assign the salmon homeobox pS12-B into the *Hox-2.2* cognate group. However, this particular classification must be considered tentative because pS12-B is equally similar to the mouse sequences representing the *Hox-2.3* cognate group.

Construction of Trees. The construction of trees from distance matrices was based on the assumption that the least distant sequences are most likely to be located on neighboring branches. Hypothetical ancestor sequences for branch points were reconstructed successively and lengths of branches were determined as nucleotide differences between points. The trees are presented unrooted because of the lack of a suitable outgroup. In the attempt to define a reference outside the Antennapedia class, we found that several mammalian homeobox sequences located at the base of the trees at different positions without changing the internal branch arrangements depicted here. A cladistic analysis of sets of sequences was performed by using the PAUP program (version 2.4) (13) with the branch-and-bound option, to be described in detail elsewhere (unpublished data).

RESULTS AND DISCUSSION

Classification of Homeobox Nucleotide Sequences. Only those vertebrate homeobox sequences were included that are at least 55% similar to the *Drosophila* Antennapedia ho-

meobox. Sixty-five nucleotide sequences of vertebrate homeobox genes were then classified according to similarities to murine homeoboxes. Those human sequences for which a corresponding mouse gene has not yet been described were used as a reference for comparisons. Fig. 1 shows a schematic display of the classifications obtained in this way. Mouse homeobox genes are organized in at least four clusters. These can be arranged colinearly, so that groups of cognate genes between clusters become apparent.

It is evident that such a classification as shown in Fig. 1 can only be achieved if the similarity among nucleotide sequences within a given cognate group is greater than the degree of similarity between genes within clusters. As will be shown below, such is generally the case. However, sequences from *Xenopus* and from zebrafish could not easily be classified within a cognate group exclusively on the basis of nucleotide sequences. In these cases, sequences flanking the homeobox were included in the comparisons. The preliminary classification of the two salmon homeoboxes into adjacent cognate groups was based on molecular cloning data as described in *Materials and Methods*. All assignments were confirmed by comparisons of available coding sequences flanking the homeoboxes.

The homeobox genes in mouse and human are arranged very similarly in four clusters. These relationships argue strongly for a common origin of the clusters of homeoboxes of mouse and human. Some cognate groups contain two sequences of rat or of *Xenopus* and one group contains three zebrafish homeoboxes. Thus, genes similar to those obtained to date from mouse and human may also be present in rat,

FIG. 1. Classification of nucleotide sequences of homeobox genes. The chromosomal arrangement of mouse homeobox genes in four clusters is shown schematically. Open lines and boxes represent additional information available from the human system. Shaded areas delineate cognate groups of related homeoboxes. Sequences at the bottom of each group could not be assigned unambiguously within that cognate group. The question mark in front of pS12-B indicates that this sequence could be only tentatively assigned to a group, as described in the text. Where sequences are believed to represent the same gene, all currently used names are given. Italics indicate incomplete or missing sequences. No attempt has been made to represent physical distances.

Xenopus, and zebrafish in similar chromosomal arrangements. From these data, it appears that the distribution of homeobox genes into four linkage groups may have preceded the divergence of the fishes and amphibians.

Comparisons of Nucleotide Sequences Suggest That Homeobox Clusters Arose from a Common Ancestral Complex. The similarities between homeoboxes were assessed quantitatively under the assumption that the number of nucleotide differences between two sequences reflects their relatedness. As shown in Fig. 1, most of the sequences of human homeoboxes correspond to one of the known mouse sequences. Comparisons of 16 pairs of such cognates revealed an average of 10.7 nucleotide substitutions in a total of 183 bp between mouse and human homeoboxes (Table 1). However, the differences between mouse boxes that belong to the same cognate group are 3-fold greater. The arithmetic mean of 15 comparisons gives an average of 28.7 nucleotide differences. Similar figures were obtained for human (average difference of 35), rat (35), and *Xenopus* (33) sequences. When mouse sequences within a chromosomal cluster were compared, they were found to differ at an average of 51 positions.

These data show that the cognate homeoboxes of mouse and human are more similar to each other than even the most similar sequences within either species. Interestingly, the nature of these differences underscores the high degree of similarity: only 5.9% of the observed differences between the corresponding human and murine boxes (Table 1) constitute replacement substitutions that alter the sequence of amino acids in the protein. Thus, about 94% of the differences are silent. The high level of similarity between mouse and human homeoboxes suggests that each pair of cognates is derived from a common ancestor. Additionally, the spacing between homeobox genes along both the *Hox-1* and the *Hox-2* cluster is conserved between mouse and human. These results support the interpretation that the existence of four distinct linkage groups predated the divergence of the mouse and human species.

Homeoboxes within each cognate group show similar degrees of sequence divergence in human and mouse. Most of the observed nucleotide differences between mouse cognate boxes again resemble silent substitutions, since only an average of 13% of changes result in amino acid replacements. Consequently, amino acid sequences of cognate homeoboxes on different clusters have been strongly conserved. These findings and detailed evidence to be presented elsewhere (unpublished data) indicate that the homeobox clusters have evolved similarly and that they originate from a common ancestral complex. This, in turn, implies that the four clusters arose from a primordial linkage group by duplication events that simultaneously involved whole clusters rather than individual genes. This interpretation would also be consistent with whole chromosome duplications or even genome duplications during chordate evolution (14).

Separate Phases of Conservation and Divergence in the Evolution of Homeobox Genes. The largest number of differences between two sequences was found when homeoboxes within clusters were compared. Mouse sequences of the *Hox-2* cluster differ from each other at 51 positions on the average. Analysis of the mouse *Hox-1* cluster yields similar results (average of differences, 56). A striking property of the differences between homeoboxes of the murine *Hox-2* cluster is that 52% of the nucleotide differences constitute amino acid replacements. This figure is significantly higher than in the cognate group comparisons. If nucleotide exchanges were introduced into homeobox sequences in a random fashion, 77% of all nucleotide changes would result in amino acid replacements. This figure was obtained by counting the number of possible changes in the homeobox codons, and it is in agreement with the calculated figure of 75% for all codons (15). It is conceivable that some of the theoretically possible exchanges of amino acids are detrimental. Therefore, the figure of 52% for replacement changes may be close to the maximum value for functionally tolerable changes. Accordingly, the high proportion of replacement substitutions in homeoboxes within clusters suggests that variability of amino acid sequences was tolerated during the expansion of the primordial complex. Conversely, the high proportion of silent changes after cluster duplications must be interpreted as the result of strong selection for conservation. We interpret these results to mean that two processes of selection operated successively on the homeobox genes. The expansion of the ancient linkage group was accompanied by considerable divergence of homeoboxes, whereas conservation was strongly favored following the duplication events leading to multiple clusters.

Evolution of the Primordial Homeobox Gene Cluster Complex. The high degree of conservation of homeobox gene clusters suggests they may still resemble the structure of the ancestral linkage group. *Hox-2* is the longest contiguous cluster characterized so far, and its limits have not yet been determined. It contains seven homeobox genes in the mouse and nine in the human. We analyzed the relationships between homeoboxes along the mouse *Hox-2* cluster by using a distance matrix (Fig. 2A). The relationships between individual boxes can be depicted as a tree (Fig. 2B). By serially reconstructing the branchpoint sequences, we determined the hypothetical ancestors for the furcations and the base of the tree. *Hox-2.2* and *Hox-2.3* form a branch that shares a common ancestor with *Hox-2.4* and *Hox-2.5*. The other major branch links together *Hox-2.7*, *Hox-2.6*, and *Hox-2.1*. Human *Hox-2* sequences subjected to this kind of analysis gave the same general picture (Fig. 2C), with the placement of *HOX-2.8* and *HOX-2.9* to the right of *HOX-2.7*. The two trees have small differences in branch lengths depending on the reconstruction of hypothetical ancestor sequences. Indeed, the same tree topologies resulted when distance matrices for the amino acid sequences of the murine and human *Hox-2* clusters were used

Table 1. Nucleotide differences between homeobox sequences

Comparison of sequences	Number of differences	Number of replacement changes	Proportion of replacement changes, %
Human vs. mouse (16 cognate pairs)	10.7 $\pm$ 4.5	0.63 $\pm$ 1	5.9
Mouse vs. mouse (15 cognate pairs)	28.7 $\pm$ 6.6	3.8 $\pm$ 2.2	13.2
Mouse vs. mouse (21 pairs for <i>Hox-2</i>)	51.1 $\pm$ 15.6	26.4 $\pm$ 12.9	51.7

Results of pairwise sequence comparisons are expressed as the arithmetic means $\pm$ standard deviations of the number of differences per 183 bp. The number of replacement substitutions was determined for each pair of sequences, respectively. The proportion of replacement differences was calculated as the quotient of the mean number of replacements over the mean total number of substitutions.

FIG. 2. Gene tree for homeobox sequences constructed from a distance matrix. (A) Distance matrix for nucleotide sequences of homeoboxes of the mouse *Hox-2* cluster. The total number of differences in 183 bp is given and the number of replacement substitutions is shown in parentheses. (B) Gene tree for homeoboxes of the mouse *Hox-2* cluster. The name of each box is given at the tip of the respective branch. Vertical lines are drawn proportionally to indicate branch lengths. (C) Gene tree for homeoboxes of the human *HOX-2* cluster.

(data not shown). The same branching orders were observed for the human *HOX-2* and the murine *Hox-2* trees when a cladistic analysis of the same sets of sequences was employed. In addition, when a tree for the murine *Hox-1* cluster was constructed, its topology was found to be compatible with that of the *Hox-2* tree. The congruence of tree topologies for two clusters suggests that the relationships exhibited by boxes along each individual cluster are representative of a primordial complex. Consequently, cognate homeoboxes were found linked together when a tree of all mouse homeobox sequences was constructed (unpublished data). Moreover, it is intriguing that the order of homeoboxes in the trees presented here resembles their order on the chromosome. This suggests that the spreading of genes along the primordial complex occurred as local gene duplications at the extremes and in the middle of the cluster.

The different lengths of branches in the trees may indicate that these local duplications took place at different times; the most disparate boxes would then represent the "oldest" in the complex. These oldest homeoboxes would have had a longer time to accumulate silent substitutions than "younger" boxes. Unfortunately, the number of silent differences between homeoboxes of the *Hox-2* cluster is uninformative in this respect (see Fig. 2A). Another possible explanation for unequal distances of treetips from the base of the tree assumes different degrees of conservation of the individual member genes during the original expansion of the linkage group. Some homeoboxes would have been allowed to diverge while others were more strictly conserved. In this case, a short branch length indicates stringency of selection and not difference in time. The high incidence of amino acid differences between homeoboxes along each cluster is compatible with this second hypothesis.

In summary, the picture of a tree suggests that the primordial homeobox cluster was created by local gene duplications and subsequent divergence of sequences. Most likely, this process involved successive unequal strand exchanges maintaining all homeobox genes within the same ancestral linkage group.

Vertebrate and *Drosophila* Homeobox Genes Share Similarities. However, these similarities cannot be considered highly homologous. Although the genes in *Drosophila* appear to be physically linked in the same order as the genes for mammalian homeoboxes, detailed analyses of the structure of the *Drosophila* Antp- and Ubx- complexes (refs. 16 and 17; and T. Kaufman and G. Olsen, personal communication) have revealed that the transcriptional orientations of some genes run in opposite directions. In the mammalian system, thus far, all transcription units in a cluster are oriented in the same direction (19–21). Additionally, some of the *Drosophila* genes (16, 22) contain splice sites in the homeobox that have not been found thus far in mammals. These findings suggest that a more distant line of evolution of homeobox genes took place in flies and involved independent duplications and inversions (16, 23, 24). We can, nevertheless, speculate that even before the divergence of arthropods and chordates, the homeoboxes most similar between flies and mammals were present as common ancestors within complexes that later expanded separately. The most likely candidates for ancestral genes, as extrapolated from sequence similarities and tree topologies, would include predecessors for *iab-7* (16) and *Hox-2.5*, *F90-2* (22) and *HOX-2.9*, *Dfd* (25) and *Hox-2.6*, and homeoboxes in the middle of the clusters related to *Antp/Scr/Ubx* and *ftz*.

The significance of these relationships for the function of homeobox genes in vertebrate development is unclear. Our findings could indicate that certain specifications of the body plan along the anterior-posterior axis were acquired before internal duplications within the clusters in the arthropod and chordate lineages occurred. A further increase in the number of homeobox genes was achieved in vertebrates by subsequent cluster duplications. Moreover, recent data from Fritz *et al.* (18) show duplicated versions of the *Hox-2.5*-, *Hox-2.4*-, and *Hox-2.3*-cognates in *Xenopus*, suggesting that additional duplications have occurred in the amphibian lineage as a result of tetraploidization events. It is conceivable that a larger number of specifying elements created new degrees of freedom for developmental control. If so, the innovations in the chordate body plans may have been mediated at least in part by homeobox gene cluster duplications. Homeobox transcripts are abundantly expressed in the central nervous system, neural crests, somites, and urogenital tract, organ systems that arose at the presumed time of cluster duplication. In this respect, it will be of great interest to determine the precise time of these duplications.

The homeobox gene system should be particularly useful in studying evolutionary relationships of quite diverged species

because of its demonstrated conservation over long time periods. In addition, the colinear relationships of clusters allow multiple comparisons of taxa, increasing the confidence of phylogenetic analyses. Even without knowing the time of origin and detailed sequence of events that led to the primordial complex and its duplication into multiple clusters, we can conclude that the evolution of vertebrate homeobox genes proceeded through two phases. These transitions can be envisioned as a gene multiplication process within a primordial gene cluster. During this expansion, variability on the amino acid level was favored within the homeobox domain. Subsequently, gene cluster duplication created a multigene family under strong conservational constraints, most probably by genome duplication. It will be of interest to determine how the gene expansion and cluster multiplication phases are reflected by other parts of the coding or cis-regulatory homeobox gene sequences when more data become available.

We thank those who contributed information prior to publication, especially Dr. E. Boncinelli (Naples). We are grateful to L. D. Bogarad and M. F. Utset for critical reading of the manuscript; to Drs. C. Stephens (Hughes Human Gene Mapping Library) and R. DeSalle (Department of Biology, Yale University) for help with computer programs and discussion; and to Dr. D. Irwin (Department of Biochemistry, University of California, Berkeley) for helpful comments. We also thank M. Siniscalchi for assisting in the preparation of this manuscript. This work was supported by National Institutes of Health Grant GM09966. C.K. and K.S. are recipients of postdoctoral fellowships of the Deutsche Forschungsgemeinschaft.

1. Gehring, W. J. (1988) *Science* **236**, 1245–1252.
2. Ingham, P. W. (1988) *Nature (London)* **335**, 25–34.
3. Scott, M. P., Tamkun, J. W. & Hartzell, G. W. (1989) *Biochim. Biophys. Acta Rev. Cancer*, in press.
4. Schughart, K., Kappen, C. & Ruddle, F. H. (1989) *Br. J. Cancer* **58**, 9–13.
5. Ruddle, F. H. (1989) in *The Physiology of Growth*, eds. Tanner, J. M. & Priest, M. A. (Cambridge Univ. Press, Cambridge, U.K.), in press.
6. Schughart, K., Pravtcheva, D., Newman, M. S., Hunihan, L. W., Jiang, Z. & Ruddle, F. H. (1989) *Genomics*, in press.
7. Bogarad, L. D., Utset, M. F., Awgulewitsch, A., Miki, T., Hart, C. & Ruddle, F. H. (1989) *Dev. Biol.*, in press.
8. Kappen, C., Schughart, K. & Ruddle, F. H. (1989) *Ann. N.Y. Acad. Sci.*, in press.
9. Boncinelli, E., Somma, R., Acampora, D., Pannese, M., D'Esposito, M., Faiella, A. & Simeone, A. (1988) *Hum. Reprod.* **3**, 880–886.
10. Fritz, A. & DeRobertis, E. M. (1988) *Nucleic Acids Res.* **16**, 1453–1469.
11. Harvey, R. P., Tabin, C. J. & Melton, D. A. (1986) *EMBO J.* **5**, 1237–1244.
12. Fjose, A., Molven, A. & Eiken, H. G. (1988) *Gene* **62**, 141–152.
13. Swofford, D. L. (1985) *Illinois Natural History Survey* (Champaign, IL).
14. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, Heidelberg).
15. Jukes, T. H. & King, J. L. (1979) *Nature (London)* **281**, 605–606.
16. Regulski, M., Harding, K., Kostriken, R., Karch, F., Levine, M. & McGinnis, W. (1985) *Cell* **43**, 71–80.
17. Kuroiwa, A., Kloter, U., Baumgartner, P. & Gehring, W. J. (1985) *EMBO J.* **4**, 3757–3764.
18. Fritz, A. F., Cho, K. W. Y., Wright, C. V. E., Jegalian, B. G. & DeRobertis, E. M. (1989) *Dev. Biol.* **131**, 584–588.
19. Do, M.-S. & Lonai, P. (1988) *Genomics* **3**, 195–200.
20. Graham, A., Papalopulu, N., Lorimer, J., McVey, J. H., Tuddenham, E. G. D. & Krumlauf, R. (1988) *Genes Dev.* **2**, 1424–1438.
21. Baron, A., Featherstone, M. S., Hill, R. E., Hall, A., Galliot, B. & Duboule, D. (1987) *EMBO J.* **6**, 2977–2986.
22. Hoey, T., Doyle, H. J., Harding, K., Wedeen, C. & Levine, M. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 4809–4813.
23. Ruddle, F. H., Hart, C. P., Rabin, M., Ferguson-Smith, A. C. & Pravtcheva, D. (1987) in *Human Genetics*, eds. Vogel, F. & Sperling, K. (Springer, Heidelberg), pp. 419–427.
24. Gehring, W. J. & Hiromi, Y. (1986) *Annu. Rev. Genet.* **20**, 147–173.
25. Regulski, M., McGinnis, N., Chadwick, R. & McGinnis, W. (1987) *EMBO J.* **6**, 767–777.