SEQUENCE CORNER

Rainer Breitling · Josef-Karl Gerber

Origin of the paired domain

Received: 16 February 2000 / Accepted: 13 August 2000

Abstract Pax proteins play a diverse role in early animal development and contain the characteristic paired domain, consisting of two conserved helix-turn-helix motifs. In many Pax proteins the paired domain is fused to a second DNA binding domain of the paired-like homeobox family. By amino acid sequence alignments, secondary structure prediction, 3D-structure comparison, and phylogenetic reconstruction, we analyzed the relationship between Pax proteins and members of the Tc1 family of transposases, which possibly share a common ancestor with Pax proteins. We suggest that the DNA binding domain of an ancestral transposase (proto-Pax transposase) was fused to a homeodomain shortly after the emergence of metazoans about one billion years ago. Using the transposase sequences as an outgroup we reexamined the early evolution of the Pax proteins. Our novel evolutionary scenario features a single homeobox capturing event and an early duplication of Pax genes before the divergence of porifera, indicating a more diverse role of Pax proteins in primitive animals than previously expected.

Keywords Pax protein · Paired domain · Homeobox · Transposase · Evolution

Introduction

Pax proteins, which play key regulatory roles during animal development, are characterized by the presence of the paired domain, a highly conserved DNA binding region of about 128 amino acids (Balling et al. 1996; Dahl et al. 1997; Mansouri et al. 1999). The paired domain is composed of two helix-turn-helix subdomains, the N-terminal subdomain (also called PAI), which also contains an N-terminal β motif, and the C-terminal subdomain (also called RED). Both subdomains can bind to DNA independently (Czerny et al. 1993; Epstein et al. 1994; Vogan et al.1996; Kozmik et al. 1997). In fact, in the *Drosophila* paired protein the N-terminal subdomain is even sufficient to confer full function to the protein in vivo (Bertuccioli et al. 1996). The C-terminal subdomain is less conserved and may be involved in protein targeting (Poleev et al. 1997). In addition to the paired domain, many Pax proteins contain a further DNA binding domain of the paired class homeobox family, which has been used to subdivide them into five large subgroups comprising *Drosophila* and vertebrate genes (Pax1–9/meso, PaxD/3–7/paired/gooseberry, Pax6–4/eyeless, PaxB/2–5–8/sparkling and PaxA/neuro; e.g. Miller et al. 2000). Pax proteins have been cloned from a diversity of metazoans, including nematodes, arthropods and many vertebrates. Their highly conserved function across much of the animal kingdom, including eye development and cephalization, has prompted the recent cloning of a number of Pax homologs from more primitive organisms, namely *Hydra* (PaxA and PaxB), corals (PaxA, PaxB, PaxC and PaxD) and sponges (spongePax, a PaxB/2–5–8/sparkling-homolog) (Sun et al. 1997; Hoshiyama et al. 1998; Miller et al. 2000). More than 100 Pax protein sequences are now available in the public databases.

Recently, Galliot and Miller (2000) presented an evolutionary scenario of Pax proteins in which a PaxA-like ancestor, containing only a paired box, underwent two independent homeobox capturing events, giving rise to the PaxB/2–5–8/sparkling family group and the PaxC/1–9/3–7/4–6 family group (Fig. 1A). The first capturing event was supposed to have taken place before the divergence of sponges (about 900 million years ago), while the second occurred before the cnidarian-triploblast split (about 700 million years ago). This is represented in

Edited by D. Tautz

Additional documentary material has been deposited in electronic form and can be obtained from http://link.springer.de/link/service/journals/00427/index.htm

R. Breitling · J.-K. Gerber (✉)
GSF-National Research Center for Environment and Health,
Institute of Experimental Genetics, 85764 Neuherberg, Germany
e-mail: gerber@gsf.de
Tel.: +49-89-31873229, Fax: +49-89-31873225

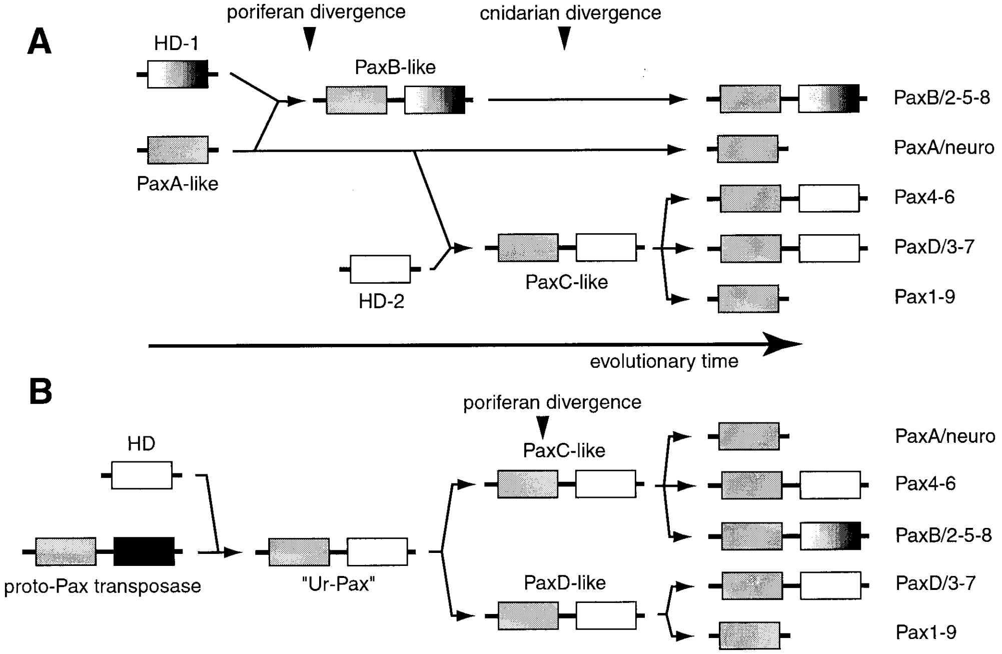


Fig. 1A, B Alternative evolutionary scenarios for the Pax protein family. **A** Scenario according to Galliot and Miller (2000). **B** Novel scenario proposed in this report after inclusion of transposases as likely suppliers of the paired domain. *Black box* Catalytic domain of transposases; *grey box* paired domain; *white box* homeodomain (*HD*); *shaded box* incomplete homeodomain

an evolutionary tree (Fig. 2A), in which PaxC is the sister group of Pax1–9/meso, Pax3–7/gooseberry/paired and Pax4–6/eyeless. A similar scenario has also been proposed by Catmull et al. (1998). This phylogeny contradicts that put forward by Hoshiyama et al. (1998) before the description of PaxC. In their evolutionary tree, Pax1–9/meso and Pax3–7/gooseberry/paired form the sister group of all other Pax proteins (Fig. 2B). These authors determined the position of the root by an analysis of the homeodomains of the Pax3–7/gooseberry/paired, PaxB/2–5–8/sparkling and Pax4–6/eyeless subfamilies.

To resolve this conflict and to elucidate the origin and early evolution of the paired box, we have undertaken a reanalysis of the available data. Because a homeodomain is absent from some of the crucial Pax proteins we decided to use a novel approach to root the evolutionary tree, by identifying several transposase proteins that are closely related to the Pax proteins and including these for the first time in the phylogenetic analysis.

Materials and methods

Sequence selection

An initial PSI-BLAST search of the non-redundant database at NCBI (http://ncbi.nlm.nih.gov) with the paired box of murine Pax9 identified more than 250 entries. Manual removal of splicing variants, truncated or mutated sequences reduced them to about 100 individual Pax proteins. This data set is avaible in electronic form.

For the subsequent phylogenetic analysis, the size of the data set was further reduced to 55 proteins by automated redundancy filtering at the 75% identity level using the Jpred2 server (http://jura.ebi.ac.uk:8888). A corresponding data set of 39 proteins was constructed for Pax-like transposases.

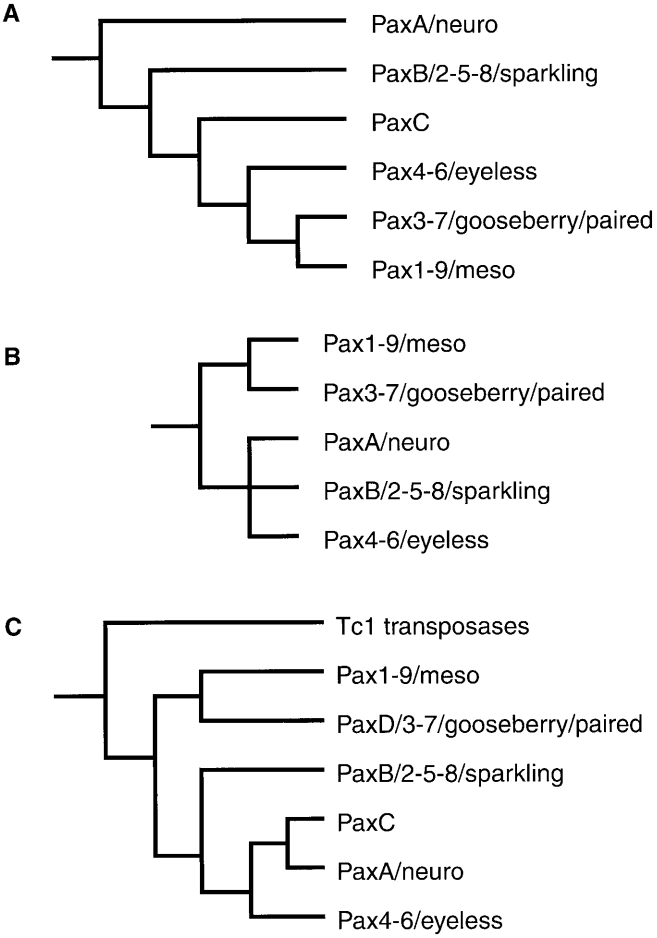
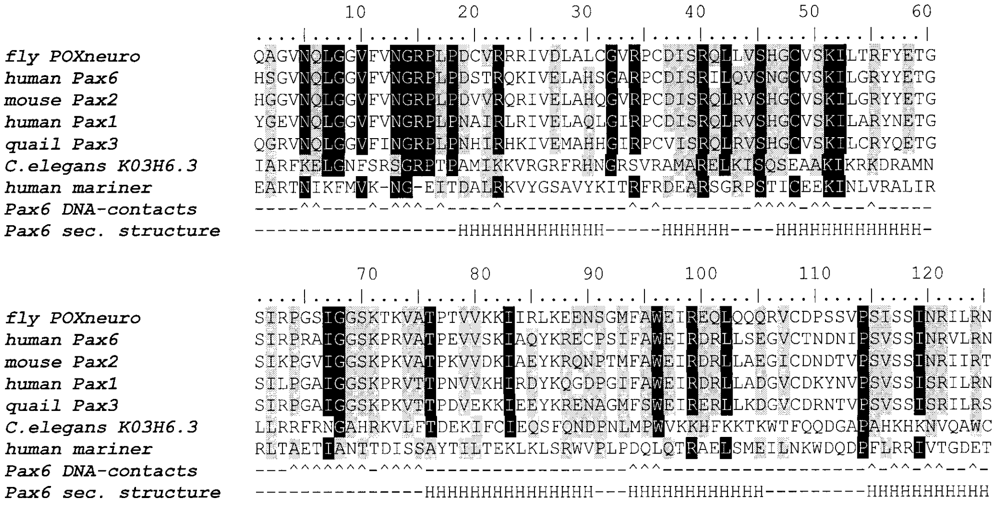


Fig. 2A–C Phylogenetic trees of the Pax family. **A** Tree according to the scenario proposed by Galliot and Miller (2000) **B** Tree derived from Hoshiyama et al. (1998). **C** Tree derived in this paper from neighbor joining and parsimony analysis and subsequent likelihood mapping, including transposases as the outgroup



```
                                10        20        30        40        50        60
                       ....|....|....|....|....|....|....|....|....|....|....|....|
fly POXneuro           QAGVNQLGGVFVNGRPLPDCVRRRIVDLALCGVRPCDISRQLLVSHGCVSKILTRFYETG
human Pax6             HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
mouse Pax2             HGGVNQLGGVFVNGRPLPDVVRQRIVELAHQGVRPCDISRQLRVSHGCVSKILGRYYETG
human Pax1             YGEVNQLGGVFVNGRPLPNAIRLRIVELAQLGIRPCDISRQLRVSHGCVSKILARYNETG
quail Pax3             QGRVNQLGGVFINGRPLPNHIRHKIVEMAHHGIRPCVISRQLRVSHGCVSKILCRYQETG
C.elegans K03H6.3      IARFKELGNFSRSGRPTPAMIKKVRGRFRHNGRSVRAMARELKISQSEAAKIKRKDRAMN
human mariner          EARTNIKFMVK-NG-EITDALRKVYGSAVYKITRFRDEARSGRPSTICEEKINLVRALIR
Pax6 DNA-contacts      ----^^----^-^^^-^----^-----------^-^--------^^^^-^^---^-----
Pax6 sec. structure    ------------------HHHHHHHHHHHHH-----HHHHHH----HHHHHHHHHHHHH-

                                70        80        90       100       110       120
                       ....|....|....|....|....|....|....|....|....|....|....|....|....|
fly POXneuro           SIRPGSIGGSKTKVATPTVVKKIIRLKEENSGMFAWEIREQLQQQRVCDPSSVPSISSINRILRN
human Pax6             SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRN
mouse Pax2             SIKPGVIGGSKPKVATPKVVDKIAEYKRQNPTMFAWEIRDRLLAEGICDNDTVPSVSSINRIIRT
human Pax1             SILPGAIGGSKPRVTTPNVVKHIRDYKQGDPGIFAWEIRDRLLADGVCDKYNVPSVSSISRILRN
quail Pax3             SIRPGAIGGSKPKVTTPDVEKKIEEYKRENAGMFSWEIRERLLKDGVCDRNTVPSVSSISRILRS
C.elegans K03H6.3      LLRRFRNGAHRKVLFTDEKIFCIEQSFQNDPNLMPWVKKHFKKTKWTFQQDGAPAHKHKNVQAWC
human mariner          RLTAETIANTTDISSAYTILTEKLKLSRWVPLPDQLQTRAELSMEILNKWDQDPFLRRIVTGDET
Pax6 DNA-contacts      ---^^^^^^^-^^^^------------------^^^------------------^-^^-^^--^-
Pax6 sec. structure    ---------------HHHHHHHHHHHHHHH---HHHHHHHHHHHH---------HHHHHHHHHH
```

Fig. 3 Multiple sequence alignment of Pax proteins and transposases. One member of every subfamily of the Pax proteins is shown, aligned to one member each of the Tc1 and mariner family of transposases. The complete alignment used for phylogenetic analysis is available in electronic format. Identical amino acids are given on *black background*, similar amino acids are *shaded in grey*. DNA contacts (∧) and alpha-helices (*H*) as derived from the 3D-structure of human Pax6 are shown *below the alignment*. Accession numbers are: human Pax1, P15863; mouse Pax2, P32114; quail Pax3, O13081; human Pax6, P26367; *Drosophila* POXneuro, P23758; *Caenorhabditis elegans* K03H6.3, T33011; human mariner transposase, Q13539

Computer analysis

The alignments of the DNA-binding domains of the Pax and transposase sequence data sets (corresponding to amino acids 5–129 of human Pax6) were combined. Proteins with partial DNA binding domains were excluded, with the exception of *Drosophila* eyegone. Sequences were aligned using ClustalW (http://www2.ebi.ac.uk/clustalw) Subsequently, positions which contains gaps in more than 50% of the sequences were removed. The resulting alignments were evaluated with the Phylip package for phylogenetic inference (Phylip version 3.5c; http://evolution.genetics.washington.edu/phylip.html). Likelihood mapping was done using the TREE-PUZZLE program (http://www.tree-puzzle.de). Results are shown for analyses using default parameter settings unless otherwise indicated.

Comparison of 3D-structures was performed at the DALI server (http://www2.ebi.ac.uk/dali/), employing an automatic pairwise three-dimensional alignment of protein structures. Secondary structures of the transposases were predicted by JPred2, which is based on a consensus evaluation of a set of prediction algorithms (http://jura.ebi.ac.uk:8888).

Results and discussion

To decide between the conflicting hypotheses of Pax evolution, we attempted to reconstruct the phylogeny and to reliably root the phylogenetic tree. As homeodomains, which were previously compared for that purpose (Hoshiyama et al. 1998), are only present in some of the Pax proteins and are conspicuously absent in the PaxA/neuro and Pax1–9 group, we restricted the analysis to the paired box itself. This was facilitated by the introduction of a novel outgroup. Comparison of the X-ray structures of the paired box of *Drosophila* paired (1PDN) and human Pax6 (6PAX) within the database of 3D-structures (http://www.rcsb.org/pdb/) revealed that the N-terminal subdomain (PAI domain) is closely related to the DNA binding domain of Tc3 transposase of *Caenorhabditis elegans* (1TC3). The DALI Z-score for the superposition of Pax6 and Tc3 is 6.5, compared to, for example, only 4.8 for the superposition with the homeodomain of engrailed. A general similarity between transposase DNA binding domains and the paired domain has been reported (Ivics et al. 1996) and their structural relationship has been observed during the analysis of the transposase structure (van Pouderoyen et al. 1997). The latter authors also show a 3D-superposition of the structures of Tc3 and *Drosophila* paired.

Initial Blast searches identified a group of transposases from *C. elegans* whose DNA binding domain seemed to be more closely related to the paired box than to most other transposases. The DNA binding domain of these *C. elegans* transposases (proteins K03H6.3, W04G5.1, F26H9.3, F49C5.8, and C27H2.1; accession numbers T33011, T26169, T21438, T22423, and T19530) shows highly significant similarity only to Bmmar1, a transposase from *Bombyx mori* [accession number AAB47739, E-score (E)=2e-27 compared to K03H6.3], and to many Pax proteins (e.g. *Hydra magnapapillata* Pax2/5/8, E=9e-05; *Phallusia mammilata* Pax6 E=3e-04; or *Paracentrotus lividus* Pax1/9 E=6e-04). The DNA binding domains of other transposases yield E-scores worse than 1e-03 (e.g. *Anopheles albimanus* transposase AAB02109, E=9e-03).

We supposed that the transposases of *C. elegans* and *B. mori* might represent molecular fossils (proto-Pax) from the time before a homeobox capturing event took place, during which the catalytic domain of the transposase was lost and the DNA-binding domain was fused to a homeobox yielding the first PAX protein. If this is indeed the case, the proto-Pax transposases should also contain the C-terminal subdomain (RED domain) of the paired box. This subdomain is less conserved among Pax proteins than the PAI domain and does not show signifi-

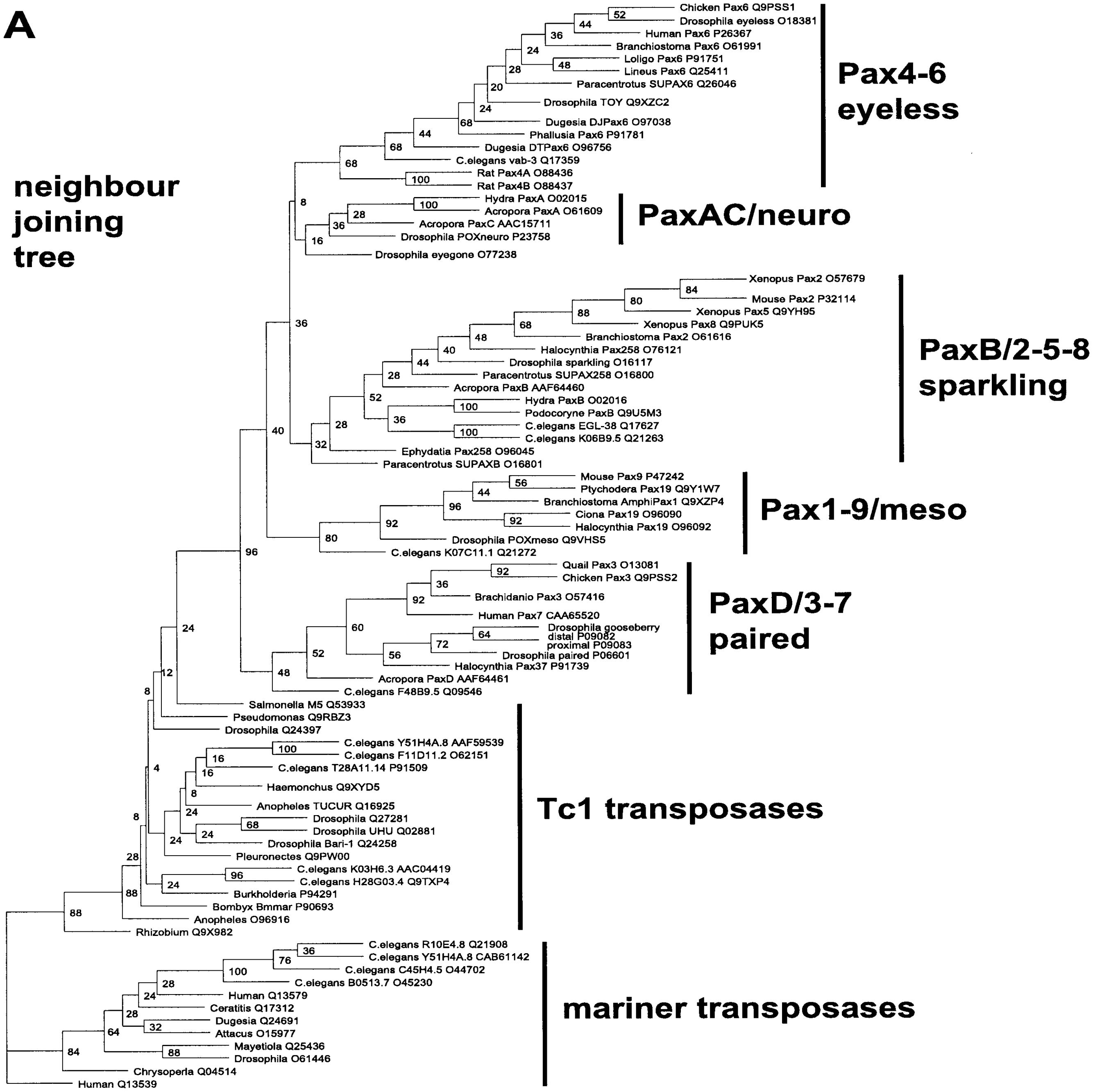


Fig. 4 Neighbor-joining (**A**) and parsimony (**B**) tree of the paired domains of Pax proteins and the DNA binding domain of transposases, arbitrarily rooted with human mariner transposase. Branch lengths and numbers indicate bootstrap support (100 pseudoreplicates for parsimony tree, 25 pseudoreplicates for neighbor joining tree)

cant homology in sequence alignments between transposases and Pax proteins. We therefore performed a secondary structure analysis of the proto-Pax transposases using a consensus method (Jpred2), which predicted that they indeed contain two helix-turn-helix motifs, homologous to both the PAI and the RED domain of Pax proteins, in agreement with earlier results (Ivics et al. 1996).

The observation that the DNA binding domain of transposases is in fact closely related to the paired box indicated that it should be possible to use them as an outgroup in the phylogenetic analysis of Pax proteins to determine the most likely evolutionary sequence. The degree of similarity of the paired domain to other helix-turn-helix domains, e.g. the homeobox, is too low for that purpose.

An alternative scenario would be that the proto-Pax transposases are not the sister group of Pax proteins, but originated within the Pax family from one of the early Pax proteins, which had before gained their DNA binding domain from some other transposase. In this case the proto-Pax transposases would be unsuitable as an out-

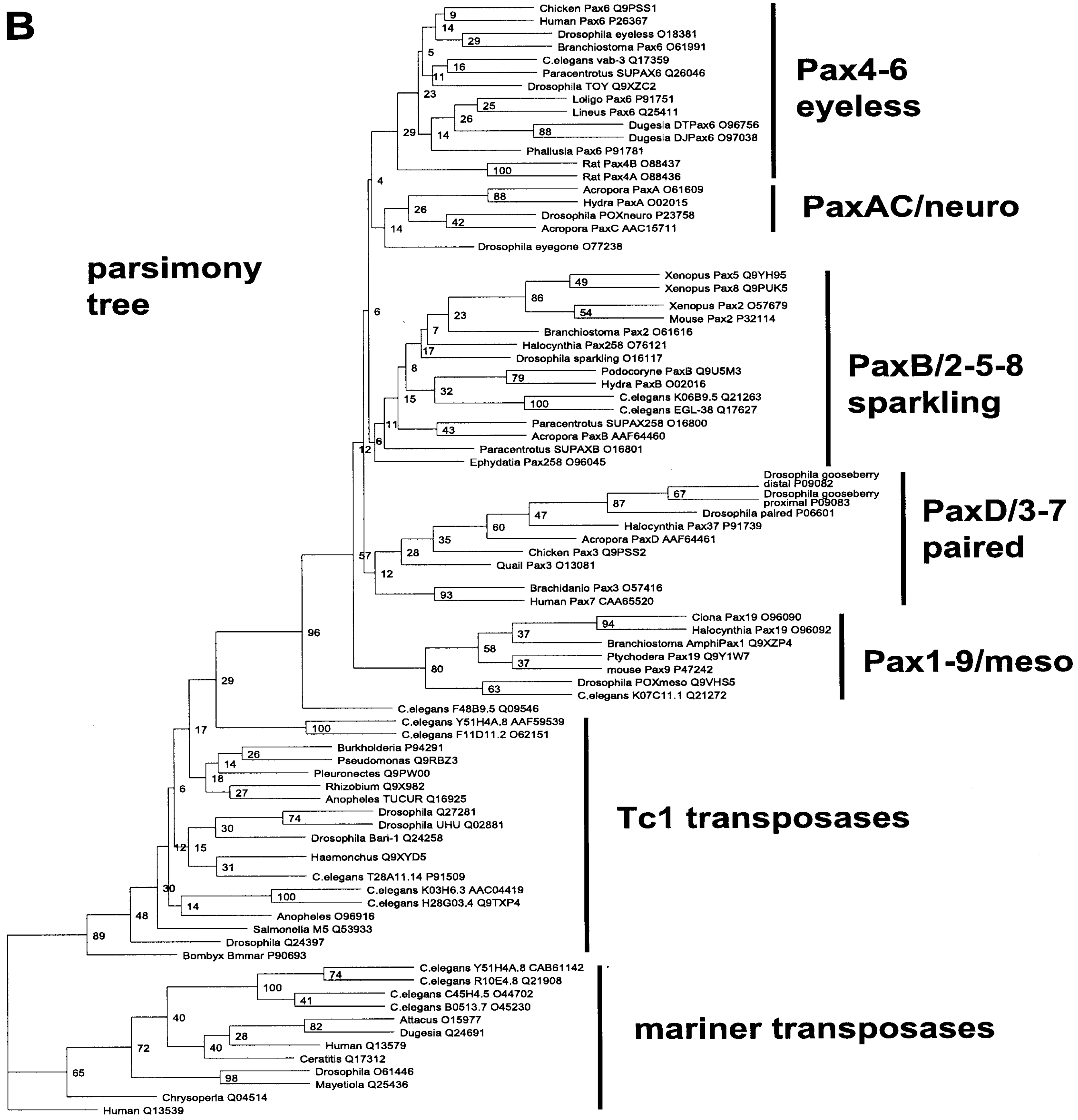


Fig. 4 Continued

group. Such a secondary reversal is, however, unlikely because it not only requires the restoration of transposase activity, but also of a functional combination between DNA recognition sequence and transposon terminal repeats.

We therefore used the transposase sequence (*C. elegans* K03H6.3, E = 2e-27) with the highest Blast score compared to Pax proteins to generate a multiple sequence alignment of Pax-like transposases using the JPred2 server. The JPred2 algorithm was also used to generate a multiple sequence alignment for Pax proteins. Both alignments were combined and realigned by using ClustalW as described. The resulting data set contained transposases of the Tc1 and mariner families, as well as a wide range of Pax proteins from all known subgroups. A resulting selection from the alignment is shown in Fig. 3, which also depicts helical regions as determined from the X-ray structure of human Pax6 as well as the residues involved in DNA-contacts. The complete alignment (available in electronic format) was then used for phylogenetic analysis.

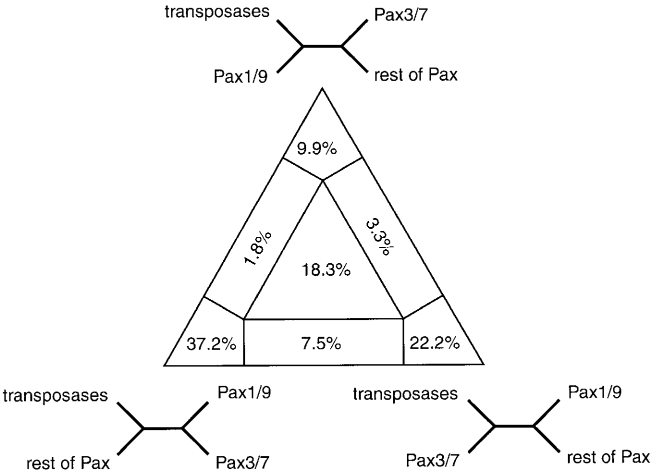


Fig. 5 Four-cluster likelihood-mapping of Pax proteins and transposases. Occupancies of the seven areas of attraction are given in percentages and the three fully resolved tree topologies are indicated at the *corners of the triangle*

Neighbor-joining and parsimony analysis reliably subdivides the Pax proteins into five large groups, which correspond to the classical subfamilies Pax1–9/meso, PaxD/3–7/gooseberry/paired, PaxB/2–5–8/sparkling, Pax4–6/eyeless and PaxA/neuro (Fig. 4A, B). The internal topology of the subfamilies agrees fairly well with the accepted evolutionary relationship of the organisms. One exception is the Pax4–6/eyeless subfamily which is extremely conserved, so that an unambiguous determination of the internal branching order was not possible. The position of *Drosophila* eyegone is also unreliable, because this protein contains only a partial paired domain (Jun et al. 1998). In both trees PaxC is significantly associated with the PaxA/neuro subfamily, although PaxC carries a homeobox, and Pax A/neuro proteins do not.

Neighbor-joining and parsimony tree reconstruction place the Pax family within the Tc1 family of transposases, while it was not possible to identify a single closest relative of the paired box. The supposed proto-Pax transposases from *C. elegans* and *B. mori*, as identified by Blast searches, are not reliably placed as a sister-group of the Pax proteins. This might be due to the general difficulty of reconstructing well-resolved phylogenetic trees of the transposase family, as described by Plasterk et al. (1999).

Also, the two methods did not agree on the position of the root of the Pax family. Either PaxD/3–7/gooseberry/paired or Pax1–9/meso were shown to diverge first from the rest of Pax proteins, but the order of these branches differed depending on the method (Fig. 4A, B). In an attempt to resolve this issue we employed the likelihood-mapping approach of Strimmer and von Haeseler (1997). The result is shown in Fig. 5. While a large number of quartets (18.3%) are found in the unresolved central area of attraction, in agreement with the partially unresolved topology of the phylogenetic tree, there is also a significant majority (37.2%) favoring a basal dichotomy of the Pax family between PaxD/3–7/gooseberry/paired plus Pax1–9/meso and the rest of the proteins, in accordance with the evolutionary scenarios suggested by Hoshiyama et al. (1998) and Balczarek et al. (1997), but not with the scenario proposed by Galliot and Miller (2000) or Miller et al. (2000). For ease of comparison a schematic phylogenetic tree, summarizing our results in comparison with previous data, is shown in Fig. 2C.

Our focus on the paired box as a descendant of a Tc1-like transposase DNA binding domain allowed us to reevaluate the early evolution of the paired domain. Our results show that the evolutionary scenario proposed by Galliot and Miller (2000; Fig. 1A) is unlikely to represent the evolution of Pax proteins correctly. This hypothesis was mainly based on the assumption that PaxA, which consists only of a paired box, resembles the probable ancestor of Pax proteins. Contrary to that idea, our scenario (Fig. 1B) is based on the assumption that the paired box is originally derived from a transposase and indicates that PaxA is probably derived by a secondary loss of the homeobox of a PaxC-like protein. Our observations also make unlikely the hypothesis that there was more than one homeodomain capturing event. Furthermore, they suggest that the first duplication of Pax proteins occurred before the divergence of the porifera. This consequently implies that sponges, which lack nerve cells and most of the organs patterned by Pax genes in higher animals, already contained (at least) two Pax genes. The function of these early Pax proteins remains a mystery.

Acknowledgements We are grateful to Prof R. Balling and J. Adamski for their continuous support and A. Lengeling for critical reading of the manuscript, and to all members of our laboratory, especially D. Laubner, Z. Marijanovic, and D. Perovic, for their encouragement and many helpful discussions. The work was supported in part by the DFG.

References

Balczarek KA, Lai ZC, Kumar S (1997) Evolution and functional diversification of the paired box (*Pax*) DNA-binding domains. Mol Biol Evol 14:829–842

Balling R, Helwig U, Nadeau J, Neubüser A, Schmahl W, Imai K (1996) Pax genes and skeletal development. Ann N Y Acad Sci 785:27–33

Bertuccioli C, Fasano L, Jun S, Wang S, Sheng G, Desplan C (1996) In vivo requirement for the paired domain and homeodomain of the paired segmentation gene product. Development 122:2673–2685

Catmull J, Hayward DC, McIntyre NE, Reece-Hoyes JS, Mastro R, Callaerts P, Ball EE, Miller DJ (1998) Pax-6 origins – implications from the structure of two coral pax genes. Dev Genes Evol 208:352–356

Czerny T, Schaffner G, Busslinger M (1993) DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site. Genes Dev 7:2048–2061

Dahl E, Koseki H, Balling R (1997) Pax genes and organogenesis. Bioessays 19:755–765

Epstein J, Cai J, Glaser T, Jepeal L, Maas R (1994) Identification of a Pax paired domain recognition sequence and evidence for DNA-dependent conformational changes. J Biol Chem 269: 8355–8361

Galliot I, Miller I (2000) Origin of anterior patterning. How old is our head? Trends Genet 16:1–5

Hoshiyama D, Suga H, Iwabe N, Koyanagi M, Nikoh N, Kuma K, Matsuda F, Honjo T, Miyata T (1998) Sponge Pax cDNA related to Pax-2/5/8 and ancient gene duplications in the Pax family. J Mol Evol 47:640–648

Ivics Z, Izsvák Z, Minter A, Hackett PB (1996) Identification of functional domains and evolution of Tc1-like transposable elements. Proc Natl Acad Sci USA 93:5008–5013

Jun S, Wallen RV, Goriely A, Kalionis B, Desplan C (1998) Lune/eye gone, a Pax-like protein, uses a partial paired domain and a homeodomain for DNA recognition. Proc Natl Acad Sci USA 95:13720–13725

Kozmik Z, Czerny T, Busslinger M (1997) Alternatively spliced insertions in the paired domain restrict the DNA sequence specificity of Pax6 and Pax8. EMBO J 16:6793–6803

Mansouri A, Goudreau G, Gruss P (1999) Pax genes and their role in organogenesis. Cancer Res 59:1707–1709

Miller DJ, Hayward DC, Reece-Hoyes JS, Scholten I, Catmull J, Gehring WJ, Callaerts P, Larsen JE, Ball EE (2000) Pax gene diversity in the basal cnidarian Acropora millepora (Cnidaria, Anthozoa): implications for the evolution of the Pax gene family. Proc Natl Acad Sci USA 97:4475–4480

Plasterk RHA, Izsvák Z, Ivics Z (1999) Resident aliens – the Tc1/mariner superfamily of transposable elements. Trends Genet 15:326–332

Poleev A, Okladnova O, Musti AM, Schneider S, Royer-Pokora B, Plachov D (1997) Determination of functional domains of the human transcription factor PAX8 responsible for its nuclear localization and transactivating potential. Eur J Biochem 247: 860–869

Strimmer K, von Haeseler A(1997) Likelihood mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci USA 94:6815–6819

Sun H, Rodin A, Zhou Y, Dickinson DP, Harper DE, Hewett-Emmett D, Li WH (1997) Evolution of paired domains: isolation and sequencing of jellyfish and hydra Pax genes related to Pax-5 and Pax-6. Proc Natl Acad Sci USA 94: 5156–5161

Van Pouderoyen G, Ketting RF, Perrakis A, Plasterk RH, Sixma TK (1997) Crystal structure of the specific DNA-binding domain of Tc3 transposase of C.elegans in complex with transposon DNA. EMBO J 16:6044–6054

Vogan KJ, Underhill DA, Gros P (1996) An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity. Mol Cell Biol 16:6677–6686