Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 31 (2004) 123–130
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

The origin of polynucleotide phosphorylase domains☆

Magdalena Leszczyniecka,[a,d,1] Rob DeSalle,[e,*,1] Dong-chul Kang,[a,b] and Paul B. Fisher[a,b,c,d,*]

[a] *Department of Pathology, Herbert Irving Comprehensive Cancer Center, College of Physicians and Surgeons, Columbia University, 630 West 168th Street, New York, NY 10032, USA*
[b] *Department of Neurosurgery, Herbert Irving Comprehensive Cancer Center, College of Physicians and Surgeons, Columbia University, 630 West 168th Street, New York, NY 10032, USA*
[c] *Department of Urology, Herbert Irving Comprehensive Cancer Center, College of Physicians and Surgeons, Columbia University, 630 West 168th Street, New York, NY 10032, USA*
[d] *Department of Biology, New York University, Washington Square, New York, NY 10003, USA*
[e] *Division of Invertebrates, American Museum of Natural History, New York, NY 10024, USA*

Received 27 February 2003; revised 7 July 2003

Abstract

In this report, we document the presence of polynucleotide phosphorylase (PNPase) in the animal eukaryotes. These proteins contain several domains, including 2 RNase PH domains (PNPase 1 and PNPase 2) which are closely related functionally and in sequence similarity to ribonuclease PH (RPH) protein. Phylogenetic analysis of the gene genealogy of these three domains suggests that PNPase was formed via a duplication event that also produced the RNase PH protein. Given the current distribution of these domains in the tree of life, these duplication events most likely occurred in the common ancestor of the three organismal superkingdoms, Archaea, Eukarya, and Bacteria. In particular, PNPase 2 and RPH are more closely related to each other than either one is to PNPase 1, suggesting a deeper differentiation of PNPase 1 in the common organismal ancestor. In addition, while PNPase 1 and PNPase 2 appear to have the same evolutionary signal as determined by the incongruence length difference (ILD) test, RPH appears to have an incongruent signal with both of the PNPase domains. This result suggests that RPH experienced different evolutionary divergence patterns than the PNPase domains, consistent with the linked nature of the two PNPase domains.

1. Introduction

Ribonucleases (RNases[2]) are enzymes that regulate stability and decay of ribonucleic acids (Deutscher and Li, 2001). They are subdivided into two categories: endo- and exo-nucleases depending on their degradative properties (Deutscher, 1993). Out of the eight distinct exonucleases identified in *Escherichia coli*, only PNPase and RNase PH catalyze phosphate-dependent degradation of RNA (Deutscher, 1993). The phosphate-dependent exoribonucleases (PDX) family of proteins contains two phosphate dependent 3′–5′ RNA exonucleases: PNPase and RNase PH (RPH; Baginsky et al., 2001). Although in vitro activity of these enzymes is similar, their degradation targets in vivo are different. While PNPase catalyzes mRNA decay, RNase PH is

☆ Supplementary data associated with this article can be found at 10.1016/S1055-7903(03)00287-2

* Corresponding authors. Fax: 1-212-769-5277.
E-mail addresses: desalle@amnh.org (R. DeSalle), pbf1@columbia.edu (P.B. Fisher).

[1] These authors contributed equally to this work.

[2] *Abbreviations:* HGT, horizontal gene transfer; PNPase, polynucleotide phosphorylase; RNase, ribonuclease; RNase PH, phosphorolytic RNA-degrading enzyme; OPS, overlapping pathway screening approach; DISH, differentiation induction subtraction hybridization; *hPNPase*$^{old-35}$, human *PNPase*; *mPNPase*$^{old-35}$, mouse, *Mus musculus PNPase*; *dPNPase*$^{old-35}$, fly, *Drosophila melanogaster PNPase*; NJ, neighbor joining; MP, maximum parsimony; TBR, tree bisection reconnection; CI, consistency index: RI, retention index; CMP-KDO synthetase, CTP:CMP-3-deoxy-manno-octulosonate cytidylyltransferase; PDX, phosphate-dependent exoribonucleases; RPH, RNase PH; ILD, incongruence length difference.

1055-7903/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2003.07.012

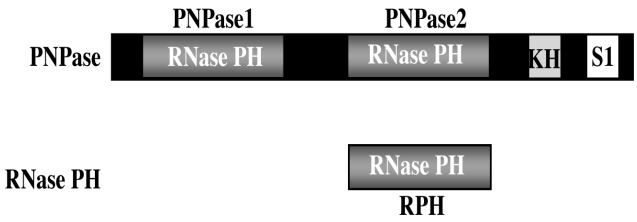

Fig. 1. The protein domain structure of the PDX (PNPase and RNase PH) proteins. The domain sizes for PNPase 1, PNPase 2, and RPH are all between 320 and 350 amino acids.

primarily involved in 3′ processing of tRNA precursors (Reuven and Deutscher, 1993; Zhang and Deutscher, 1988). The common functional properties of these proteins in combination with analysis of amino acid sequences of the active domains of these members of the PDX family indicate sufficient similarity to warrant considering them as being part of an expanded gene or protein domain family.

On a structural level, PNPase contains four distinguishable domains: two RPH domains, one KH domain, and one S1 domain (Fig. 1). Use of this structural information provides amino acid sequences from the two RPH domains that can be aligned and used to determine evolutionary relationships of these domains from a broad array of organisms. Of interest, the distribution of these proteins in the major lineages of the tree of life indicates that the RNase PH orthologue is found in all three superkingdoms—Eukarya, Archaea, and Bacteria—and is therefore an ancient and important protein domain that shares genealogy as a result of the divergence of the three superkingdoms. On the other hand, the *PNPase* gene (and therefore the two RPH domains) displays a more limited distribution amongst the three superkingdoms. In particular, PNPase is absent in all Archaea examined to date and is also missing in some single celled eukaryotes.

The current paper examines the evolutionary and phylogenetic events that led to the formation of these two proteins in the genomes of organisms in the three organismal superkingdoms. Our approach is to first construct phylogenetic trees based on matrices where sequences from the three domains are treated as taxa. This approach allows us to examine whether the domains themselves form monophyletic groups and hence indicate that each domain arose from a common ancestral domain. From this first approach we can also obtain the pattern of divergence of the three domains. Our second approach is to determine if there is any incongruence in the evolutionary or phylogenetic signal emanating from these domains. To accomplish this goal we treated the domains as character partitions and the organisms themselves as taxa in phylogenetic analysis. This approach allowed the determination of incongruence between the different domains, and lead to the inference of different evolutionary histories for the different domains. Such different evolutionary history can be interpreted in the light of horizontal gene transfer (HGT).

2. Materials and methods

2.1. Cloning of $hPNPase^{old-35}$ in human and orthologues in mouse and fly.

An overlapping pathway screening (OPS) approach, which involved hybridization of a temporally spaced subtracted HO-1 human melanoma differentiation (DISH) library (Huang et al., 1999) with a probe prepared from a senescent AG0989B progeria cell (Corriel Repository, NJ), identified a cDNA, *old-35*, that displayed elevated expression in both terminally differentiated melanoma cells and senescent progeria cells (Leszczyniecka et al., 2002). Sequence and biochemical characterization indicated that *old-35* encodes a full length human polynucleotide phosphorylase, and this gene has been named $hPNPase^{old-35}$ (GenBank Accession No. AY027528). The 5′ region of $hPNPase^{old-35}$ was cloned using an $hPNPase^{old-35}$ specific primer P1 (5′-TTTT GCTCGTTTTGATAATG-3′) from fibroblast interferon treated HO-1 human melanoma cells using a complete open reading frame approach (Kang et al., 2002). The 3′ region of $hPNPase^{old-35}$ was cloned using the 3′ RACE procedure with gene specific nested primers P2 (5′-TTTTGCTCGTTTTGATAATG-3′) and P3 (5′-CTAAT TCTCAGTGATTTTTT-3′) and dT primer, yielding an ~400-bp product. To identify a $hPNPase^{old-35}$ homologue in the mouse, $mPNPase^{old-35}$, a BLAST search was utilized to screen a mouse EST database (http://www.ncbi.nlm.nih.gov). Two primers M1 (5′-TCGGATCTTGATGG CGGCTGCAG-3′) and M2 (5′TGACTTACTTTTAAT AATAAATAT-3′) were designed in the 5′ and 3′ most distal portions of known ESTs and a full-length *Mus musculus PNPase*, $mmPNPase^{old-35}$, was amplified by RT-PCR from total RNA derived from mouse embryos (10 d.p.i.). A *Drosophila* (*D. melanogaster*) clone containing a $dmPNPase^{old-35}$ like cDNA was identified using a BLAST search and Clone No. LD03255 was purchased from Research Genetics (Carlsbad, California) and sequenced in its entirety.

2.2. Other PNPase and RNase PH sequences

PNPase and RNase PH sequences were obtained from GenBank, except for human, mouse, and fruit fly sequences, after rigorous BLAST searches of the database (Supplemental Table 1). BLAST searches were performed using Mega BLAST search at http://www.ncbi.nlm.nih.gov/BLAST/ of nr protein database (All non-redundant GenBank CDS translations + PDB

+ SwissProt + PIR + PRF). Available PNPase sequences were obtained from the Entrez database. The remaining sequences for the un-sequenced genomes were obtained by BLAST search of the microbial genomes using tblastn and blastp at http://www.ncbi.nlm.nih.gov/BLAST/. Scores for the all BLAST alignments were in the range of *e*-122 to *e*-110 between the *hPNPase*$^{old-35}$ and its bacterial homologues. The scores between *hPNPase*$^{old-35}$ and bacterial RNase PH were much lower due to the size differences between the proteins (*e*-value of *e*-07 to *e*-06). Selected members of another family of 3′–5′ RNase exonuclease RNase II and other RNA metabolism related proteins including RNase E and RNA polymerase were obtained to include as outgroups for the phylogenetic trees we generated. Since *Arabidopsis thaliana* has two *PNPase* genes, they were split in four separate protein domains that we designated as in the following Athal 11 (gene 1, RNase PH domain 1; where gene 1 indicates the gene obtained with the higher BLAST score, *e*-122 to *e*-110), Athal 12 (gene 1 RNase PH domain 2), Athal 21 (gene 2, RNase PH domain 1; where gene 2 indicates the *PNPase* gene with the lower BLAST score, *e*-07 to *e*-06), and Athal 22 (gene 2 RNase PH domain 2).

2.3. Sequence alignment and phylogenetic matrix construction

ClustalW (Thompson et al., 1994) was used to perform all alignments of the 281 protein domain sequences. We varied the alignment gap costs to do alignments for costs of 2, 10, and 50 and maintained the Gonnett cost matrix throughout all alignments. The three alignments gave 1203, 1109, and 953 alignment positions for alignment costs of 2, 10, and 50, respectively. The three alignments gave 671, 641, and 594 informative positions for alignment costs of 2, 10, and 50, respectively. The three separate alignments were then elided (Wheeler et al., 1995) to upweight regions of the alignment that are stable to the varying alignment parameters. Final alignments were arranged in two matrices. The first contained sequences only for the catalytic regions of the PNPase1, PNPase2, and RNase PH domains in a single matrix as separate taxa (Supplemental Table 2). This matrix was used to assess the monophyly of each of the three domains and to determine the relationships of the three domains to one another. The second matrix was constructed by considering the three domains as character partitions for the organisms from which these gene sequences were obtained. This matrix (Supplemental Table 3) was used to determine the presence of incongruence between the three distinct domains. Incongruence was examined using the ILD test as developed by Farris et al. (1994, 1995) and implemented in the Partition Homogeneity Test (PHT) in PAUP* (Swofford, 1998). One hundred replicates were used in all Partition Homogeneity Tests in this study.

2.4. Phylogenetic analyses

All tree building was accomplished using PAUP* (Swofford, 1998). We performed both maximum parsimony (MP) and neighbor joining (NJ) analyses in all aspects of this work. We also used a genetic identity cost matrix and a Gonnett cost matrix in separate parsimony analyses using PAUP* (Swofford, 1998). Each MP analysis used 100 replicates of random taxon additions with tree bisection reconnection (TBR) branch swapping, and searching on every tree during each replicate. Consistency index (CI), retention index (RI), and tree length were recorded for each tree search. In addition, we performed bootstrap (Felsenstein, 1985) and jackknife (Farris et al., 1996) analyses for the MP and NJ approaches. These analyses were performed using 1000 replicates of the bootstrap or jackknife re-sampling procedure. The generated bootstrap and jackknife trees are in complete agreement with the random addition trees further strengthening our contention that the search strategy we employed did an aggressive search.

3. Results and discussion

3.1. hPNPase^{old-35} sequence orthologues and paralogues

Screening of a terminally differentiating subtracted human melanoma (DISH) cDNA library (Jiang and Fisher, 1993; Huang et al., 1999) with RNA derived from senescent progeria fibroblasts, the overlapping pathway screening (OPS) approach, identified a novel cDNA, *hPNPase*$^{old-35}$ (Leszczyniecka et al., 2002). Based on sequence, length and protein domain structure, *hPNPase*$^{old-35}$ is the first *PNPase* homologue to be identified in animal eukaryotes (Fig. 1). To investigate the evolutionary history of *hPNPase*$^{old-35}$ and to identify additional *PNPases* in other organisms we performed exhaustive BLAST searches through all available genomes and EST databases. Interestingly, we found that the *PNPase* sequences were absent from the Archaea (all 16 currently completely sequenced genomes) and some single cell eukaryotes (the genomes of *Saccharomyces cereviciae*, *Schizosaccharomyces pombe*, and *Plasmodium falciparum*). We also found two *PNPase* genes in *A. thaliana* as well as in some of the other plant genomes we searched (*Pisum sativa* and *Spinachia oleracea*). To unravel the phylogenetic events that created the PDX family of proteins, we separately analyzed the sequences encoding the two RPH domains in PNPase and compared them to RNase PH proteins. They were designated as PNPase1, PNPase2, and RNase PH respectively (Fig. 1).

3.2. Phylogenetic analysis

Using multiple phylogenetic approaches we consistently found that each of the 3 PDX domains (PNPase1, PNPase2, and RNase PH) formed strong monophyletic groups amongst their own orthologues that were robustly supported by high bootstrap and jackknife values for both PNPase domains and moderately high values for the RNase PH domain (Fig. 2 and Table 1). This analysis found that PNPase2 and RNase PH are more closely related to each other than either is to PNPase1, which is surprising in that the PNPase domains (PNPase1 and PNPase2) are physically linked in all PNPases.

A comparison of *PNPase* and *RNase PH* at the cDNA level further supports the phylogenetic hypothesis. Analysis of cDNAs encoding the *RNase PH* transcript (*Homo sapiens* and all bacterial cDNAs examined) and the *PNPase* transcript revealed that they were of similar length, even though the RNase PH protein was only half the size of the PNPase protein. Additionally, the level of similarity at the nucleotide level between the *PNPase* gene and the *RNase PH* gene in all pairwise comparisons we conducted, was the same (37% raw nucleotide sequence similarity) throughout the coding and non-coding cDNA regions. The high degree of conservation of this 5′ untranslated region of *RNase PH* with respect to *PNPase 1* and *2* is remarkable and at this time is of unknown functional significance.

The phylogenetic patterns and the conservation of the 5′ untranslated region of cDNAs are compatible with a scenario where two duplication events occurred to produce the current day *PNPase* and *RNase PH* genes (Fig. 3). In the common ancestor of all life, a "PNPase1 like" domain underwent the first duplication event to produce a linked "*PNPase1* and *PNPase2* like" ancestor. The ancestral linked *PNPase1* and *PNPase2* then duplicated again to produce a "PNPase copy" that then acquired an upstream stop codon preventing the translation of the duplicated *PNPase1* domain (Fig. 3). *PNPase2* remained linked to PNPase1 to form the *PNPase* gene and *RNase PH* then was freed to move to another location in the ancestral genome. Because *RNase PH* is found in the genomes of organisms in all three superkingdoms and it is the assumed to be the product of a duplication event of PNPase 1 and 2, the most parsimonious explanation for the timing of the duplication events is that both duplications hypothesized here occurred in the common ancestor of all three superkingdom lineages. Other scenarios would require independent duplications after the three superkingdoms separated.

The topologies of species relationships for the three domains is shown in Fig. 4. In general, these topologies are congruent with our current knowledge of phylogeny in the three superkingdoms. Of particular interest are three results. First, for the RPH domain, our analysis confirms the sister relationship of Archaea and Eukarya.

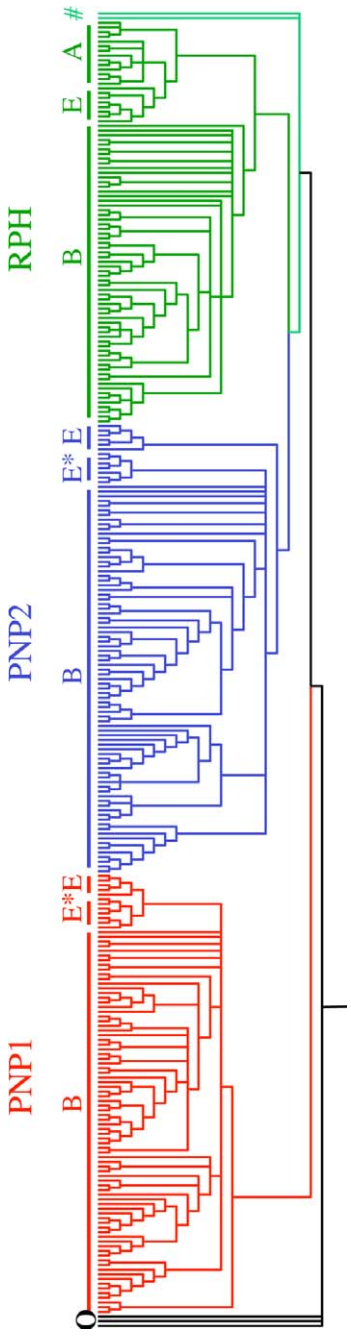

Fig. 2. Phylogenetic analysis of 281 PDX domains. Maximum parsimony jackknife consensus tree showing only those nodes present in >50% of the jackknife replicates. Specific bootstrap and jackknife values for selected higher level nodes are given in Table 1. The MP search resulted in 49 trees with a tree length = 62504, consistency index (CI) = 0.3255, and retention index (RI) = 0.6745. Branch colors are as follows: red (PNPase1), blue (PNPase2), green and aqua (RPH), and black (outgroups, RNase II, and RNAP). The bars above the branches of the tree indicate major organismal groupings where A, Archaea; B, Bacteria; and E, Eukarya. E* indicates a clade of mixed PNPases functioning in the chloroplast of plants and in *Chlamydia* and *Chlamydophila* species. Exact taxon labels are given in Fig. 4 and elaborated upon in Supplemental Table 1. (For interpretation of the references to colours in this figure legend, the reader is referred to the web version of this article.) The # sign represents two Methanobacterium thermoautotrophicum RNAP's.

Table 1
Jackknife (J) and bootstrap (B) support values under different methods of analysis (maximum parsimony on the left of the slash (/) and Neighbor Joining on the right of the slash) and different weighting schemes

Group	Test					
	GonJ	GIJ	1J	GonB	GIB	1B
PNP1	76/100	100/100	100/100	83/100	100/100	91/100
PNP2	99/100	100/100	100/100	76/100	100/100	100/100
RPH	47/99	49/100	61/99	48/88	46/100	46/93
PNP2 + RPH	49/99	64/100	63/100	45/100	69/100	48/100
PDX	</100	100/98	100/100	</68	100/100	100/86
ArchRPH	97/100	100/100	78/80	94/100	99/100	70/71
EukRPH	89/100	82/100	97/88	87/100	83/100	77/99
BactRPH	99/100	100/100	100/100	76/100	99/100	99/60
Arch+EukRPH	99/100	96/100	91/100	99/100	79/100	92/99
EukPNP1	</<	</<	</<	</<	</<	56/ <
BactPNP1	</<	</<	</<	</<	</<	</<
Plant + ChlaPNP1	95/100	88/100	96/100	90/100	91/100	89/100
EukPNP2	76/100	64/82	56/95	76/86	60/93	62/83
BactPNP2	</66	</<	</95	</97	</<	72/90
Plant + ChlaPNP2	93/100	58/100	87/100	90/100	75/99	87/100

Six analyses were performed for each of the two different methods of analysis. Three were jackknife analyses using the Gonnet (Gon) weighting matrix, the genetic identity (GI) weighting matrix and equal (1) weighting scheme, and three analyses were Bootstrap analyses using the Gonnet (Gon) weighting matrix, the genetic identity (GI) weighting matrix and equal (1) weighting scheme. Both the maximum parsimony and Neighbor Joining analyses supported monophyly of PNP1, PNP2, RPH, PNP2 + RPH and PDX under all weighting schemes. Any jackknife or bootstrap value less than 45% is indicated by a < symbol Abbreviations in GROUP column are PNP1, PNPase 1 domain; PNP2, PNPase 2 domain; RPH, RNase PH domain; PDX, phosphate dependant exonuclease; Arch, Archaea; Euk, Eukaryote; Bact, Bacteria; Chla, *Chlamydia*.

Fig. 4 and Table 1 show that for the RPH partition, this sister group relationship is well supported. Second two HGT events are a probable explanation for the distribution of these domains in the plant lineage. It has been previously shown that plants contain two PNPases (Baginsky et al., 2001). Although genes in the nucleus encode both PNPases, their presence in plants is very likely the result of two separate horizontal transfer events. Our phylogenetic analysis clearly shows that the plant chloroplast PNPase domains group with PNPase 1 and 2 from *Chlamydia* (Fig. 4 and Table 1). This result is not entirely surprising since similar relationships have been observed for these two groups of organisms for the CTP:CMP-3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase) gene (Royo et al., 2000). It is probable that the transfer of the CMP-KDO synthetase gene from a *Chlamydia* like ancestor to plants occurred at the same time as the transfer of PNPase. Similarly, the presence of the second PNPase function in the mitochondria of plants is most likely explained by an endosymbiont relationship of the plants with a proteobacterium, perhaps similar to today's *Paracoccus* (Doolittle, 1998). These results are consistent with the suggestion by Symmons et al. (2000, 2002) that these two domains are functionally distinct. The third major result is that the sequences from these three domains will be useful in determining the placement of a bacterial species into major groups, but these domains as phylogenetic tools to determine relationships of the major groups will be of limited utility. Fig. 4 demonstrates that placement of the members of major groups into those groups can be accomplished with a relatively high degree of robustness, but that bootstraps on higher level hierarchical relationships are too low to give any great confidence in inferences considering higher level relationships.

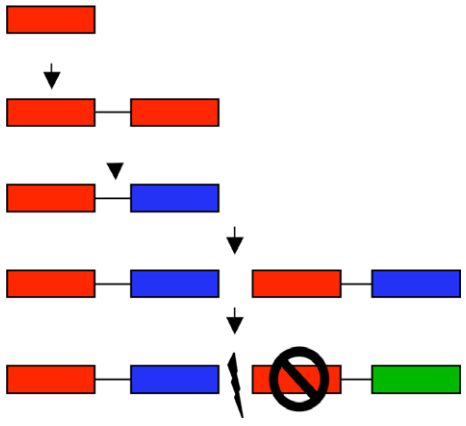

Fig. 3. A possible scenario consistent with the phylogenetic analysis and cDNA information presented in this communication where an ancestral PNPase1 gives rise to a linked, duplicated pair of PNPase1 and PNPase2 domains that are duplicated a second time to produce the PNPase and the RPH proteins. Based on the cDNA information, we hypothesize that the 5′ untranslated region has been silenced in the *RNase PH* gene to give the present distribution of these protein domains in the tree of life.

3.3. Domain signal congruence and evolutionary history of protein domains

To assess whether the various domains have experienced different evolutionary histories after their

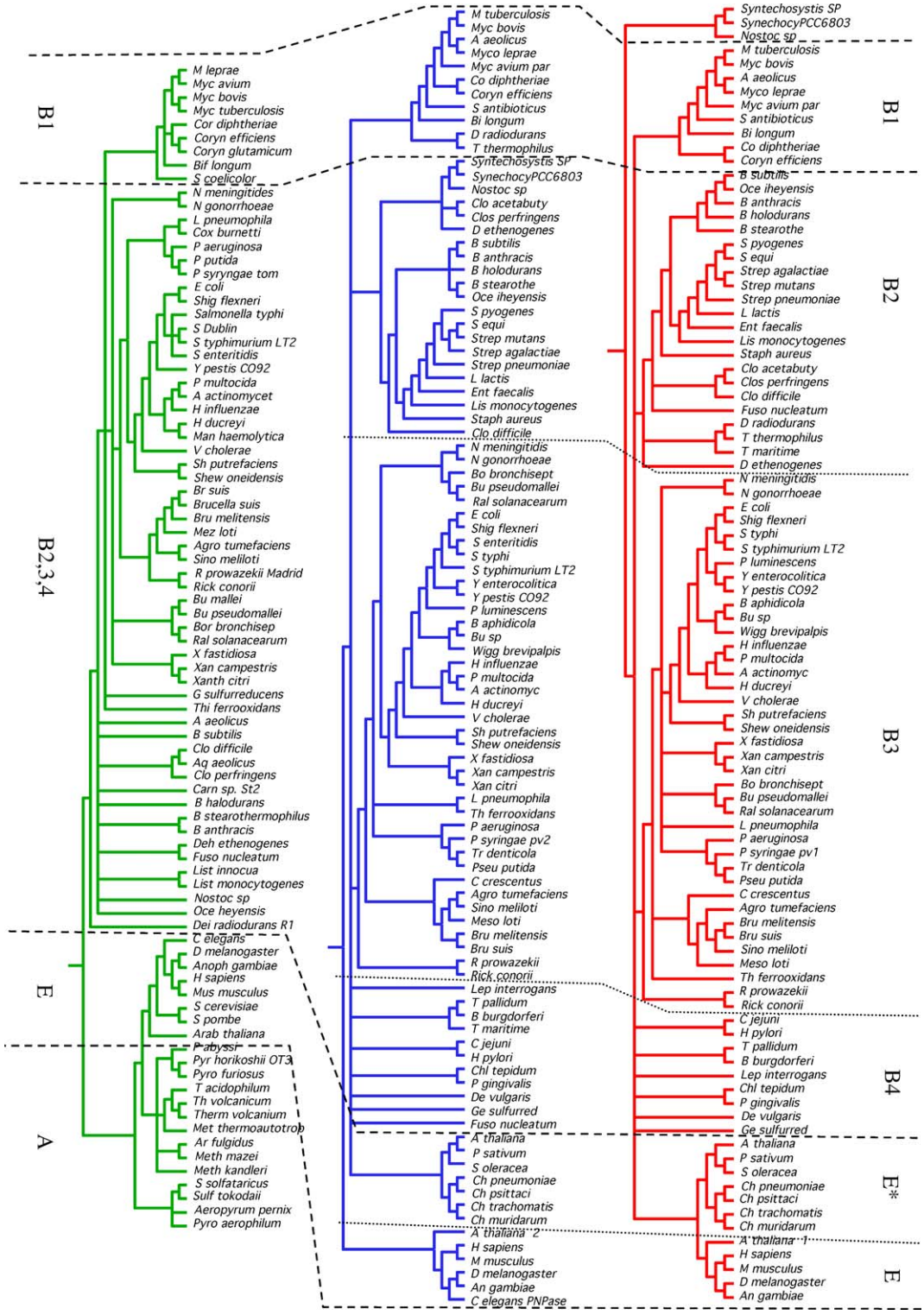

Fig. 4. Enlarged topologies for the three PDX domains shown in Fig. 2. The subtrees are pruned from the the Jakknife tree shown in Fig. 2 where only those groups supported by 45% values are shown. Red branches are the PNPase1 domains, blue branches the PNPase2 lineage, and green branches the RPH lineage. Dotted lines demarcate similar taxonomic distributions used for the three organismal domains. E indicates the Eukaryotes, E* indicates the clade containing plants and the *Chlamydia* PNPase domains, B1 is a small group of Mycobacteria and *Clostridium*, B2 is an assemblage of *Clostridium* and *Bacillus*, B3 is a large assemblage of proteobacteria found in the PNPases, B4 is a mixture of several basal bacterial lineages, and A are the Archaea that have only the RPH domain in their genomes. Full names of organisms in this tree are given in Supplemental Table 1. Bootstrap and jackknife support for some of the groups in these subtrees are given in Table 1. (For interpretation of the references to colours in this figure legend, the reader is referred to the web version of this article.)

divergence from one another in a common ancestor, we performed ILD tests (Farris et al., 1994, 1995). This analysis was complicated by the unequal distribution of the three domains in organisms over the three superkingdoms in the tree of life. For instance, while RPH is found in nearly all organisms, PNPases are absent in a large number of organisms. The matrix we used to examine the potential for incongruence amongst the three protein domains is therefore comprised of sequences for the three domains as three character partitions for those organisms that have all three domains. Some organisms such as *Arabidopsis*, that have two unlinked copies of the *PNPase* gene are included in this matrix. In addition, all Archaea and some of the single celled Eukarya that lack the *PNPase* genes are removed from the matrix. Table 2 lists the bacterial and eukaryal species that were used in the incongruence length difference (ILD) analysis. The partition homogeneity test (Farris et al., 1994, 1995; implemented in PAUP*; Swofford, 1998) results show that while the two PNPase domains are not incongruent ($p < 0.50$), the two domains by themselves (PNPase1 vs RPH, $p < 0.01$; PNPAse2 vs RPH, $p < 0.01$) or taken together (PNPase 1 plus PNPase2 vs RPH, $p < 0.01$) are incongruent with the RPH domain. These results point to an interesting case of domain family evolution where two linked domains (*PNP 1* and *2*) appear to be less closely related to each other than to the unlinked RPH domain. In addition, the linked domains give very similar or congruent patterns of organismal history (see Fig. 4) while the RPH domain gives incongruent phylogenetic signal as evidenced by the statistically significant ILD tests and the differences in pattern of divergence of species when comparing the PNPase 1 and 2 sub-phylogenies with the RPH phylogeny (Fig. 4). This result is consistent with the fact that the two PNPase domains are physically linked in the genomes of organisms as part of the gene for the same protein (PNPase) and suggests that evolutionary events such as lineage extinctions and horizontal transfer may have had a role in the evolution of the three domains.

Table 2
List of bacterial and eukaryal species used in congruence analysis

M. tuberculosis	*L. pneumophila*
A. thaliana1	*Th. ferrooxidans*
H. sapiens	*R. prowazekii*
M. musculus	*Ri. conorii*
D. melanogaster	*Ag. tumefaciens*
An. gambiae	*Si. meliloti*
C. elegans	*Mes. loti*
A. thaliana2	*Br. melitensis*
N. meningitidis	*Br. suis*
N. gonorrhoeae	*F. nucleatum*
Bo. bronchisept	*B. subtilis*
Bu. pseudomallei	*B. anthracis*
Ra. solanacearum	*B. holodurans*
E. coli	*B. stearothe*
Shi. flexneri	*O. iheyensis*
S. typhi	*Lis. monocytogenes*
S. typhimurium	*Cl. difficile*
Y. pestis. CO92	*Cl. perfringens*
H. influenzae	*Nostoc. sp*
P. multocida	*D. ethenogenes*
A. actinomyc	*D. radiodurans*
H. ducreyi	*Bi. longum*
V. cholerae	*Co. diphtheriae*
Sh. putrefaciens	*Cor. efficiens*
She. oneidensis	*A. aeolicus*
X. fastidiosa	*Myc. leprae*
Xa. campestris	*Myc. bovis*
Xa. citri	*P. syringae*
P. aeruginosa	*Pse. putida*

For Genbank accession information see Supplemental Table 1.

4. Conclusion

The PDX family of domains offers a unique opportunity to examine ancient events involved in the formation of genes in the genomes of the three superkingdoms of life. Our results address three major questions in the evolution of the three domains in the PDX family. First, we address the order of divergence of the three domains from a common ancestral domain. Second, we address the question of obvious horizontal transfer of domains within the plant lineage. Third, we address the congruence of phylogenetic patterns emanating from the sequences of the three domains. In this way we can assess whether the evolutionary history of the three domains are congruent.

Specifically, our results indicate an ancient divergence of the domains in the PDX domain family. The phylogenetic patterns obtained using the sequences from these proteins indicates that PNPase 2 and RPH are more closely related to each other than either is to PNPase 1, indicating a more ancestral divergence of PNPase 1 in the evolution of the domain family. Structural considerations of the 5′ untranslated regions of the RPH protein indicate further that this untranslated gene region is related to the PDX family of gene domains. Fig. 3 shows an evolutionary scenario for the present day formation of these domains in the PNPase protein and the RPH protein. The scenario requires two duplication events, an event to "silence" the five prime coding region of the *RPH* gene and an event to move the *RPH* gene to another location in the genome. A horizontal transfer event involving plant PDX domains can also be inferred and corroborates previous suggestions of such patterns of transfer (Baginsky et al., 2001; Doolittle, 1998; Royo et al., 2000). While caution should be used when inferring horizontal transfer events (Roeloff and Van Haastert, 2001; Salzberg et al., 2001; Stanhope et al., 2001), the patterns of relationship for the plant PDX domains are convincing evidence for a transfer event.

While the ILD test and the statistical inference made from such a test is not unflawed (Dolphin et al., 2000), it is a conservative test and in the present case easily detects incongruence between the phylogenetic signal emanating from the sequences of the PNPase protein and the RPH protein. The incongruent signal from the gene sequences of the two proteins is interpreted as the result of the two genes having different evolutionary histories after their divergence from each other.

Acknowledgments

M.L. thanks Drs. S. Small and P. Furmanski for their encouragement during the course of this research. The present study was supported in part by National Institutes of Health Grants CA35675 and CA74468, the Samuel Waxman Cancer Research Foundation and the Michael and Stella Chernow Endowment (P.B.F.). P.B.F. is the Michael and Stella Chernow Urological Cancer Research Scientist. R.D. is supported in part by the Lewis B. and Dorothy Cullman Program for Molecular Systematic Studies and the Ambrose Monell Collection for Molecular and Microbial Research.

References

Baginsky, S., Shteiman-Kotler, A., Liveanu, V., Yehudai-Resheff, S., Bellaoui, M., Settlage, R.E., Shabanowitz, J., Hunt, D.F., Schuster, G., Gruissem, W., 2001. Chloroplast PNPase exists as a homo-multimer enzyme complex that is distinct from the *Escherichia coli* degradosome. RNA 7, 1464–1475.

Deutscher, M.P., 1993. Promiscuous exoribonucleases of *Escherichia coli*. J. Bacteriol. 175, 4577–4583.

Deutscher, M.P., Li, Z., 2001. Exoribonucleases and their multiple roles in RNA metabolism.. Prog. Nucleic Acid Res. Mol. Biol. 66, 67–105.

Dolphin, K., Belshaw, R., Orme, C.D.L., Quicke, D.L.J., 2000. Noise and incongruence: interpreting results of the incongruence length difference test. Mol. Phylogenet. Evol. 17, 401–406.

Doolittle, W.F., 1998. A paradigm gets shifty. Nature 392, 15–16.

Farris, J.S., Kallersjo, M., Kluge, A.G., Bult, C., 1994. Testing significance of congruence. Cladistics 10, 315–320.

Farris, J.S., Kallersjo, M., Kluge, A.G., Bult, C., 1995. Constructing a significance test for incongruence. Syst. Biol. 44, 570–572.

Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Huang, F., Adelman, J., Jiang, H., Goldstein, N.I., Fisher, P.B., 1999. Differentiation induction subtraction hybridization (DISH): a strategy for cloning genes displaying differential expression during growth arrest and terminal differentiation. Gene 236, 125–131.

Jiang, H., Fisher, P.B., 1993. Use of a sensitive and efficient subtraction hybridization protocol for the identification of genes differentially regulated during the induction of differentiation in human melanoma cells. Mol. Cell. Different. 1, 285–299.

Kang, D.-C., Gopalkrishnan, R.V., Wu, Q., Jankowsky, E., Pyle, A.M., Fisher, P.B., 2002. mda-5: an interferon-inducible putative RNA helicase with double-stranded RNA-dependent ATPase activity and melanoma growth-suppressive properties. Proc. Natl. Acad. Sci. USA 99, 637–642.

Leszczyniecka, M., Kang, D.C, Sarkar, D., Su, Z.Z., Holmes, M., Valerie, K., Fisher, P.B., 2002. Identification and cloning of human polynucleotide phosphorylase, hPNPase old-35, in the context of terminal differentiation and cellular senescence. Proc. Natl. Acad. Sci. USA. 99, 16636–16641.

Reuven, N.B., Deutscher, M.P., 1993. Substitution of the 3′ terminal adenosine residue of transfer RNA in vivo. Proc. Natl. Acad. Sci. USA 90, 4350–4353.

Roeloff, Van Haastert, 2001. Genomics: genes lost during evolution. Nature 411, 1013–1014.

Royo, J., Gimez, E., Hueros, G., 2000. CMP-KDO synthetase: a plant gene borrowed from gram-negative eubacteria. Trends Genet. 16, 432–433.

Swofford, D.L., 1998. PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods), Version 4. Sinauer Associates, Sunderland, MA.

Salzberg, S.L., White, O., Peterson, J., Eisen, J.A., 2001. Microbial genes in the human genome: lateral transfer or gene loss? Science 292, 1903–1906.

Stanhope, M.J., Lupas, A., Italia, M.J., Koretke, K.K., Volker, C., Brown, J.R., 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature 411, 940–944.

Symmons, M.F., Williams, M.G., Luisi, B.F., Jones, G.H., Carpousis, A.J., 2002. Running rings around RNA: a superfamily of phosphate-dependent RNases. Trends Biochem. Sci. 1, 11–18.

Symmons, M.F., Jones, G.H., Luisi, B.F., 2000. A duplicated fold is the structural basis for polynucleotide phosphorylase catalytic activity, processivity, and regulation. Struct. Fold Des. 8, 1215–1226.

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

Wheeler, W.C., Gatesy, J., DeSalle, R., 1995. Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Mol. Phylogenet. Evol. 4, 1–9.

Zhang, J.R., Deutscher, M.P., 1988. Oligoribonuclease is encoded by a highly conserved gene in the 3′-5′ exonuclease superfamily. J. Bacteriol. 170, 522–527.