Protein Science (1994), *3*:1081-1088. Cambridge University Press. Printed in the USA.

Unexpected sequence similarity between nucleosidases and phosphoribosyltransferases of different specificity

ARCADY R. MUSHEGIAN[1] AND EUGENE V. KOONIN[2]

[1] Department of Plant Pathology, University of Kentucky, Lexington, Kentucky 40546-0091
[2] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

(RECEIVED February 17, 1994; ACCEPTED April 29, 1994)

Abstract

Amino acid sequences of enzymes that catalyze hydrolysis or phosphorolysis of the *N*-glycosidic bond in nucleosides and nucleotides (nucleosidases and phosphoribosyltransferases) were explored using computer methods for database similarity search and multiple alignment. Two new families, each including bacterial and eukaryotic enzymes, were identified. Family I consists of *Escherichia coli* AMP hydrolase (Amn), uridine phosphorylase (Udp), purine phosphorylase (DeoD), uncharacterized proteins from *E. coli* and *Bacteroides uniformis*, and, unexpectedly, a group of plant stress-inducible proteins. It is hypothesized that these plant proteins have evolved from nucleosidases and may possess nucleosidase activity. The proteins in this new family contain 3 conserved motifs, one of which was found also in eukaryotic purine nucleosidases, where it corresponds to the nucleoside-binding site. Family II is comprised of bacterial and eukaryotic thymidine phosphorylases and anthranilate phosphoribosyltransferases, the relationship between which has not been suspected previously. Based on the known tertiary structure of *E. coli* thymidine phosphorylase, structural interpretation was given to the sequence conservation in this family. The highest conservation is observed in the N-terminal α-helical domain, whose exact function is not known. Parts of the conserved active site of thymidine phosphorylases and anthranilate phosphoribosyltransferases were delineated. A motif in the putative phosphate-binding site is conserved in family II and in other phosphoribosyltransferases. Our analysis suggests that certain enzymes of very similar specificity, e.g., uridine and thymidine phosphorylases, could have evolved independently. In contrast, enzymes catalyzing such different reactions as AMP hydrolysis and uridine phosphorolysis or thymidine phosphorolysis and phosphoribosyl anthranilate synthesis are likely to have evolved from common ancestors.

Keywords: nucleosidases; phosphoribosyltransferases; sequence similarity

Enzymes that catalyze hydrolysis or phosphorolysis of the *N*-glycosidic bond in nucleotides, nucleosides, and related compounds are central to salvage pathways of nucleotide metabolism and are also important in de novo synthesis of nucleotides and certain amino acids (Table 1; reviewed by Lin, 1987; Neuhard & Nygaard, 1987). Nucleosidases are involved in the regulation of the intracellular concentration of nucleotides and allow utilization of ribose and deoxyribose as a source of carbon and energy. Phosphoribosyltransferases provide, via 5-phosphoribosyl-1-pyrophosphate (PRPP), a crucial link between nucleotide and amino acid metabolism. The substrates of these enzymes are very diverse, but the reacting groups always involve phosphate and ribose or deoxyribose (Table 1). The interest to the nucleosidases has been enhanced by the recent observation that human platelet-derived endothelial cell growth factor is identical to thymidine phosphorylase (Barton et al., 1992; Ishizawa & Yamada, 1992). Phosphoribosyltransferases also have attracted considerable attention because deficiency in these enzymes leads to various metabolic disorders in humans, e.g., Lesch-Nyhan disease (Stout & Caskey, 1985).

A single organism, i.e., *Escherichia coli*, which has been studied in the most detail, encodes numerous nucleosidases and phosphoribosyltransferases, including enzymes that catalyze essentially identical reactions but differ in their specificity toward the nucleotide base (e.g., thymidine phosphorylase, uridine phosphorylase, and purine phosphorylase; see Table 1). A number of nucleotide sequences of genes encoding these enzymes from all types of organisms have been reported (Table 1). The 3-dimensional (3D) structure has been determined for human purine nucleoside phosphorylase (PNP; Ealick et al., 1990) and *E. coli* thymidine phosphorylase (DeoA; Walter et al., 1990), leading to detailed characterization of the respective active centers. Except for the obvious similarity between human TYPH

Reprint requests to: Eugene V. Koonin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894; e-mail: koonin@ncbi.nlm.nih.gov.

Table 1. *Nucleosidases, phosphoribosyltransferases, and related enzymes*[a]

Protein/ gene[b]	Organism(s)	Enzymatic activity	Reaction	M_r/quaternary structure[c]	Metabolic pathway	References
Udp/*udp*	*E. coli*	Uridine phosphorylase	Uridine + P_i = ribose-1-P + uracil	8 × 22	Pyrimidine salvage	Neuhard & Nygaard, 1987
DeoD/*deoD*	*E. coli*	Purine nucleoside phosphorylase	Purine (deoxy)-ribonucleoside + P_i = (deoxy)-ribose-1-P + purine	6 × 23.7	Purine salvage	Neuhard & Nygaard, 1987
Amn/*amn*	*E. coli*	AMP glycosylase	AMP + H_2O = adenine + ribose-5-P	6 × 52	Purine salvage	Leung & Schramm, 1980; Leung et al., 1989; Neuhard & Nygaard, 1987
PNP	Mammals, *B. subtilis*, *M. leprae*	Purine nucleoside phosphorylase	Purine (deoxy)-ribonucleoside + P_i = (deoxy)-ribose-1-P + purine	3(2) × 32	Purine salvage	Neuhard & Nygaard, 1987; Lin, 1987; Ealick et al., 1990
DeoA/*deoA*	*E. coli*	Thymidine phosphorylase	Thymidine + P_i = ribose-1-P + thymine	2 × 45	Thymidine salvage	Neuhard & Nygaard, 1987; Lin, 1987; Walter et al., 1990
TYPH	Human	Thymidine phosphorylase (=PD-ECGF)	Thymidine + P_i = ribose-1-P + thymine	2 × 45	Thymidine salvage	Yoshimura et al., 1990; Bartonet al., 1992
TrpD/*trpD* (Eubacteria)	Eubacteria, archaea, yeast, plants	Anthranilate phosphoribosyl-transferase	Anthranilate + PRPP = *N*-phosphoribosyl-anthranilate + PP_i	2(?) × 36	Second step in tryptophan biosynthesis	Crawford, 1989; Kim et al., 1993
TrpG/*trpG*	*E. coli* and several other bacterial species	Anthranilate phosphoribosyl-transferase (with N-terminal glutamine aminotransferase domain)	Anthranilate + PRPP = *N*-phosphoribosyl-anthranilate + PP_i; chorismate + L-glutamine = anthranilate + pyruvate + L-glutamine	2 × 58.3	First and second step in tryptophan biosynthesis	Pittard, 1987; Crawford, 1989
Apt/*apt* (*E. coli*)	Eubacteria, eukaryotes	Adenine phosphoribosyltransferase	Adenine + PRPP = AMP + PP_i	2 × 20	Purine salvage	Neuhard & Nygaard, 1987
Gpt/*gpt* (*E. coli*)	Eubacteria, eukaryotes	Guanine phosphoribosyltransferase	Guanine + PRPP = GMP + PP_i; xanthine + PRPP = XMP + PP_i; hypoxanthine + PRPP = HXMP + PP_i	3 × 16.9	Purine salvage	Neuhard & Nygaard, 1987
Hpt/*hpt* (*L. lactis*)	Eubacteria, eukaryotes	Hypoxanthine phosphoribosyl-transferase	Hypoxanthine + PRPP = IMP + PP_i	? × 20	Purine salvage	Neuhard & Nygaard, 1987
Upp/*upp* (*E. coli*)	Eubacteria, eukaryotes	Uracil phosphoribosyltransferase	Uracil + PRPP = UMP + PP_i	3 × 23.5	Pyrimidine salvage	Neuhard & Nygaard, 1987
PyrE/*pyrE* (*E. coli*)	Eubacteria, eukaryotes	Orotate phosphoribosyltransferase	Orotate + PRPP = OMP + PP_i	2 × 23.4	Fifth step in pyrimidine biosynthesis	Neuhard & Nygaard, 1987

(*continued*)

Table 1. *Continued*

Protein/gene[b]	Organism(s)	Enzymatic activity	Reaction	M_r/quaternary structure[c]	Metabolic pathway	References
PurF/*purF* (*E. coli*)	Eubacteria, eukaryotes	Glutamine phosphoribosyltransferase	L-glutamine + PRPP = L-glutamate + PP_i + 5-phosphoribosylamine	4(3) × 53	First step in de novo purine biosynthesis	Neuhard & Nygaard, 1987
HisG/*hisG* (*E. coli*)	Eubacteria, eukaryotes	ATP phosphoribosyltransferase	ATP + PRPP = 1-(5-phospho-D-ribosyl)-ATP + PP_i	6 × 33	First step in histidine biosynthesis	Winkler, 1987
PncB/*pncB* (*E. coli*)	*E. coli*	Nicotinate phosphoribosyltransferase	Quinolinate + PRPP = NaMN + PP_i + CO_2	? × 44	NAD biosynthesis	White, 1982; Tritz, 1987
NadC/*nadC*	*S. typhimurium*	Quinolinate phosphoribosyltransferase	Nicotinate + PRPP + ATP = NaMN + PP_i + ADP + P_i	2 × 31	NAD biosynthesis	White, 1982; Tritz, 1987
PrsA/*prsA* (*E. coli*)	Eubacteria, eukaryotes	Ribose-phosphate pyrophosphokinase (PRPP synthetase)	ATP + ribose-5-phosphate = AMP + PRPP	5(?) × 31	PRPP synthesis for de novo and salvage pathways of nucleotide metabolism	Neuhard & Nygaard, 1987

[a] Only enzymes for which sequence information is available were included.
[b] Where sequences are available for several organisms, the protein/gene name is for the organism(s) indicated in parentheses.
[c] The first number is the number of identical subunits and the second number is the M_r.

(endothelial growth factor) and *E. coli* thymidine phosphorylase encoded by the *deoA* gene (Barton et al., 1992), and the somewhat lower similarity between human PNP and partial nucleosidase sequence from *Bacillus subtilis* (Wu et al., 1992), no significant relationships have been derived from analysis of amino acid sequences of nucleosidases. On the other hand, comparative analysis of the amino acid sequences of phosphoribosyltransferases has revealed a conserved motif that has been implicated in PRPP binding (Busetta, 1988; de Boer & Glickman, 1991).

In this work, using computer methods for sequence analysis, we delineate 2 new families of nucleosidases and phosphoribosyltransferases, each including unexpected relationships between enzymes that catalyze very different reactions.

Results and discussion

Comparison of the amino acid sequences of nucleosidases with the nonredundant sequence database (National Center for Biotechnology Information, NIH) using the BLASTP (Altschul et al., 1990) program revealed highly significant similarity between *E. coli* purine phosphorylase (DeoD) and uridine phosphorylase (Udp), with the probability of matching by chance (P) about 10^{-8}. More unexpectedly, significant similarity ($P = 3.6 \times 10^{-5}$) was observed between human thymidine phosphorylase (TYPH) and *E. coli* anthranilate phosphoribosyltransferase (TrpG). Further analysis by iterative database search using BLASTP, TBLASTN (screening of a nucleotide sequence database translated in 6 reading frames for similarity to an amino acid sequence), and multiple alignment using the MACAW program (Schuler et al., 1991) showed that these pairs of relatively strongly similar proteins belonged to 2 distinct families of enzymes that have not been described previously.

Family I: Bacterial nucleosidases and plant vegetative storage proteins

Family I included DeoD, Udp, *E. coli* AMP glycosidase (Amn), uncharacterized proteins from *E. coli*, *Bacteroides uniformis*, and *Treponema pallidum*, and, unexpectedly, bark storage proteins and wound-induced proteins (BSP-win4 family) from poplars (Fig. 1). Other than the aforementioned relationship between DeoD and Udp, these proteins showed only limited similarity to each other, with P values between 0.1 and 0.01. Nevertheless, analysis of BLAST outputs for mutually consistent pairwise alignments and multiple alignment using MACAW (Schuler et al., 1991) showed that the similarity concentrated in 3 distinct, conserved motifs (Fig. 1). The probability of obtaining each of these blocks by chance alone, as computed using MACAW, was below 10^{-5} (only 1 of the closely related BSP-win4 sequences was used for these calculations; see Methods and Schuler et al. [1991] for details of significance evaluation by MACAW), suggesting that the observed relationship is functionally and evolutionarily relevant. Therefore, it is likely that the uncharacterized proteins belonging to this family, including the plant stress-induced proteins, possess nucleosidase activity.

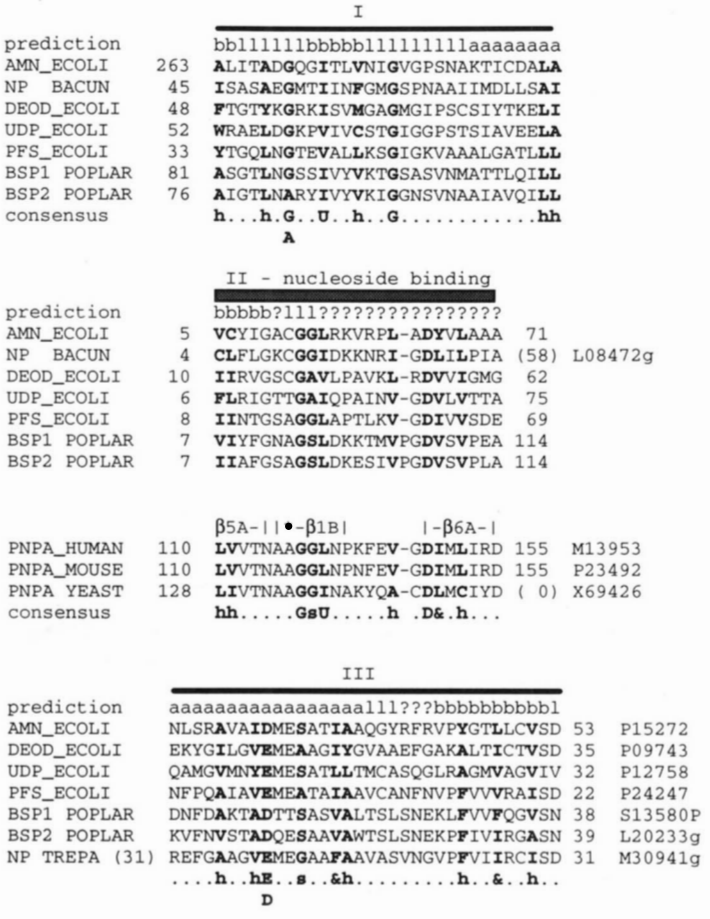

Fig. 1. Conserved sequence motifs in the nucleosidase family I. The alignment of family I proteins was constructed using the MACAW program, and the boundaries of the 3 conserved blocks were determined so as to achieve maximum statistical significance. The distances between the blocks as well as the distances from protein ends are shown by numbers. For the putative ribose-binding motif (II), the alignment additionally includes mammalian purine nucleoside phosphorylases (PNPs) and the putative PNP from yeast. The protein sequence from *Mycobacterium leprae* that is related to eukaryotic PNPs (GenBank U00022; E.V. Koonin, unpubl. obs.) is not shown; the related sequence from *B. subtilis* (PIR A42708) is incomplete and includes only the C-terminal part of the protein downstream from motif II. The consensus includes amino acid residues or pairs of similar residues that are conserved in all of the aligned sequences. U designates a bulky aliphatic residue (I, L, V, or M); & designates a bulky hydrophobic residue (I, L, V, M, F, Y, or W); h designates any hydrophobic residue (I, L, V, M, F, Y, W, C, or A); s designates a small residue (G, A, or S); and dot designates any residue. The secondary structure for human PNP is from the published 3D structure; A and B designate 2 distinct β-sheets found in this protein (Ealick et al., 1990). The secondary structure for family I proteins is the consensus of predictions for individual proteins (a designates an α-helix, b designates a β-sheet, and l designates a loop). The asterisk designates the alanine residue in human PNP that interacts with the nucleoside ribose. Each sequence is accompanied by its accession number in SWISS-PROT, PIR (P), or GenBank (g). The partial sequences of putative nucleosidases from yeast, *Bacteroides uniformis* (BACUN), and *Treponema pallidum* (TREPA) have not been described previously and were identified in this study by database searches using the TBLASTN program.

The proteins of the nucleosidase family I also contained a region of low similarity to eukaryotic purine nucleoside phosphorylases (PNPs). This similarity has been detected by BLAST searches but failed to attain statistical significance. It was remarkable, however, that this region coincided with the conserved motif II (Fig. 1). The regular expression [&C][&C]x_2[GN]x_2[GAS]Ux_5[UA]x_{1-2}D&x[UC] (alternative residues are shown in brackets; U designates a bulky aliphatic residue, namely I, L, V, or M; & designates a bulky hydrophobic residue, namely I, L, V, M, F, Y, or W; and x designates any residue) was specific for family I and eukaryotic PNPs, with only 1 false positive (insect general odorant-binding protein 1 precursor; SWISS-PROT P31418) selected during the database screening.

Inspection of the 3D structure of human PNP (Ealick et al., 1990) showed that the conserved motif comprised part of the active center. The central feature of the PNP structure is a distorted β-barrel that consists of 2 β-sheets, one of them 8-stranded (sheet A) and the other one 5-stranded (sheet B). Motif II includes 3 β-strands, two of which belong to sheet A and one to sheet B (Fig. 1). The turn between β5A and β1B (or in other words, at the interface between the 2 β-sheets) and β1B strand itself are directly involved in nucleoside binding. In particular, the backbone amido group of alanine 116 (Fig. 1) forms a hydrogen bond with the ribose 3′ hydroxyl (Ealick et al., 1990). Although this residue is not conserved in all of the proteins of family I, this and the adjacent positions contain mostly small residues, namely glycines and serines (Fig. 1), that are likely to perform the same function. Therefore, we believe that the structure of the ribose-binding part of the nucleoside-binding site is partially conserved between family I and eukaryotic PNPs. Detailed comparisons failed to reveal any other sequence conservation between these groups of proteins.

The functions of the 2 other conserved motifs in the (putative) nucleosidases of family I remain unknown. It is plausible that motif I, which appears to contain a hydrophobic β-strand separated by a flexible loop from a downstream α-helix (Fig. 1), may be the phosphate-binding site (Saraste et al., 1990; Schulz, 1992).

Delineation of family I resulted in prediction of nucleosidase activity for several uncharacterized proteins. In particular, pfs protein appears to be a new, not identified previously nucleosidase in *E. coli*; it is distinct from inosine phosphorylase (XapA) whose sequence is not yet available because the respective genes are located in different regions of the chromosome (Brun et al., 1990). Interestingly, the *pfs* gene is adjacent to the *dgt* gene encoding dGTP phosphohydrolase, another enzyme of nucleotide metabolism (Wurgler & Richardson, 1990). Conceivably, dGTP phosphohydrolase and the new nucleosidase may be functionally linked. BSP-win4 family proteins in poplars are encoded by a family of physically linked genes and are induced by various stress signals, namely, elevated nitrogen level, short photoperiod, or mechanical wounding; BSP proteins serve as transient nitrogen deposit in bark (Parsons et al., 1989; Clausen & Apel, 1991; Coleman et al., 1992; Davis et al., 1993). Nucleosidase ac-

tivity appears to be compatible with a stress-related function because it would increase the intracellular pools of free bases and pentoses that could be reused for biosynthetic processes. Alternatively, it is possible that the storage function of BSPs is unrelated to their proposed nucleosidase activity; an obvious parallel is the recruitment of various enzymes as animal lens crystallins (Piatigorsky, 1993).

Family II: Thymidine phosphorylases and anthranilate phosphoribosyltransferases

Family II consisted of bacterial and mammalian thymidine phosphorylases and numerous anthranilate phosphoribosyltransferases from bacteria, archaea, and yeast (Fig. 2). Anthranilate phosphoribosyltransferase is a ubiquitous enzyme that catalyzes

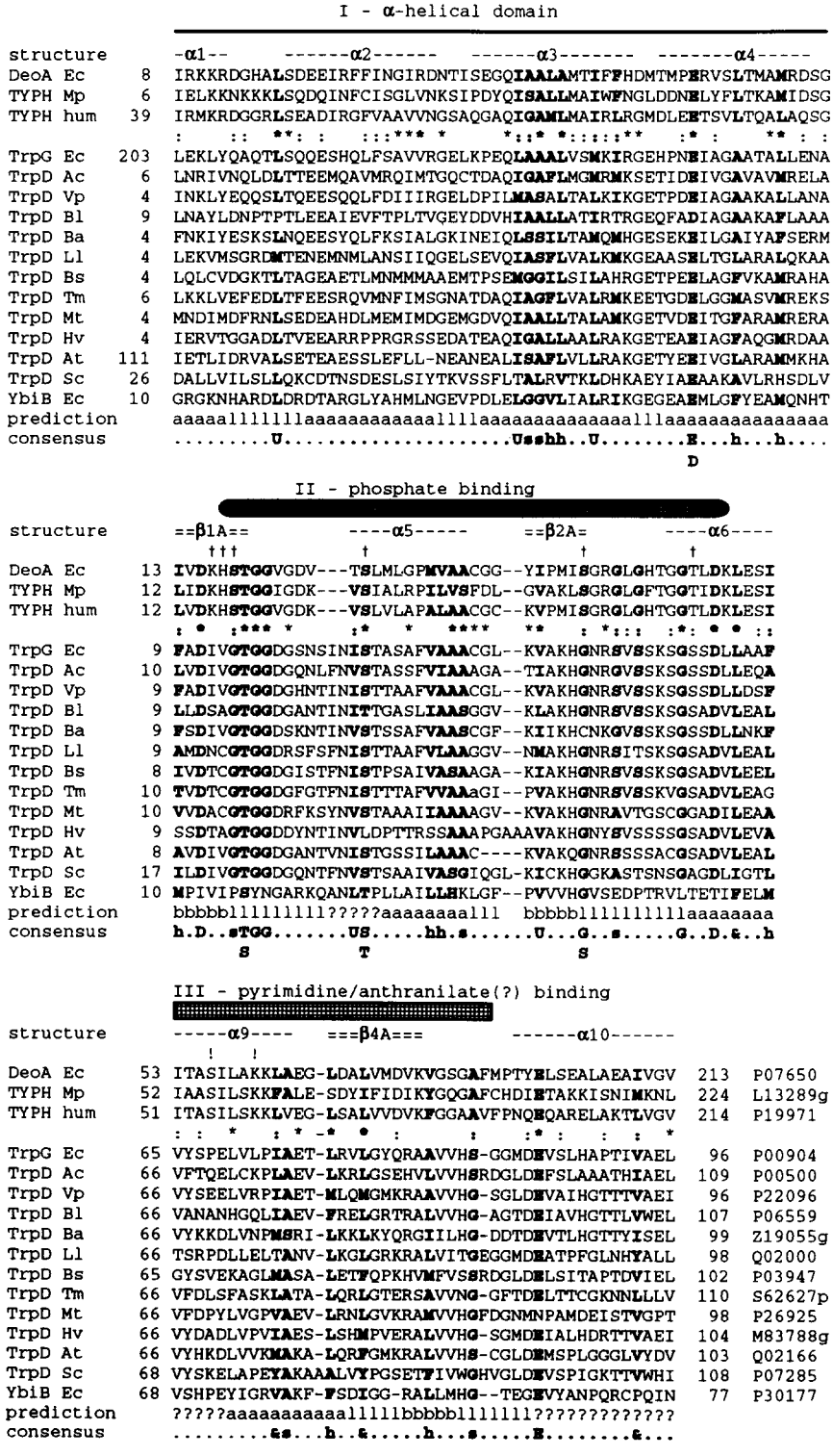

Fig. 2. Conserved sequence motifs in family II: thymidine phosphorylases and anthranilate phosphoribosyltransferases. The consensus shows the conserved amino acid residues with 1 possible exception; residues that conform to the consensus are highlighted by bold. Asterisks show identical residues, and colons show similar residues in the sequences of human TYPH and *E. coli* TrpG. The secondary structure for DeoA is from the published 3D structure (Walter et al., 1990); the secondary structure for anthranilate phosphoribosyltransferases is the consensus of predictions. Daggers denote the amino acid residue that directly contacts the phosphate in DeoA; exclamation marks show residues that contact the pyrimidine base. For the other details and designations see legend to Figure 1. Ec, *E. coli*; Mp, *Mycoplasma pyrum*; Ac, *Acinetobacter calcoaceticus*; Vp, *Vibrio parahaemolyticus*; Bl, *Brevibacterium lactofermentum*; Ba, *Buchnera aphidicola*; Ll, *Lactococcus lactis*; Bs, *Bacillus subtilis*; Tm, *Thermotoga maritima*; Mt, *Methanobacterium thermoautotrophicum*; Hv, *Haloferax volcanii*; At, *Arabidopsis thaliana*; Sc, *Saccharomyces cerevisiae*.

the second step in tryptophan biosynthesis (Table 1; reviewed by Crawford, 1989). Apart from the already mentioned highly significant similarity between human TYPH and *E. coli* TrpG, BLAST searches detected moderate similarity between various anthranilate phosphoribosyltransferases and thymidine phosphorylases, with *P* values between 0.1 and 0.001. The region of highest conservation in most of these analyses included a block of about 60 amino acid residues that in the majority of the family II proteins is located near the N-terminus (with the exception for 2 domain enzymes of tryptophan biosynthesis, e.g., *E. coli* TrpG that contain the N-terminal anthranilate synthetase domain). Subsequent multiple alignment analysis revealed 2 additional conserved blocks (Fig. 2). For the set of 6 sequences including the *E. coli* and human thymidine phosphorylases, together with 4 diverse sequences of anthranilate phosphoribosyltransferases, the random matching probabilities for each of the 2 N-terminal blocks were below 10^{-19}, as computed using MACAW.

The alignment of family II is readily interpretable in terms of the known 3D structure of *E. coli* thymidine phosphorylase, DeoA (Walter et al., 1990). The protein molecule consists of the small α-helical domain and the large α/β domain. The α-helical domain is comprised of amino acid residues 1–65 and 163–193. Strikingly, it is this domain that is most highly conserved between thymidine phosphorylases and TrpD (blocks I and III in Fig. 2 containing α1–4 and α9, respectively). Secondary structure prediction indicated that the respective regions of TrpD have α-helical conformation and, for the N-terminal part of the domain, suggested an almost perfect alignment of the 4 helices in TYPH and TrpD (Fig. 2). The N-terminal portion of the α-helical domain, and in particular α1 and α3, is involved mostly in subunit interaction in the DeoA dimer. On the other hand, sequence conservation in α4, including the nearly invariant glutamic acid (Fig. 2), may suggest a more specific, not yet described function. The distal part of the α-helical domain containing α9 is implicated in the pyrimidine base binding (Walter et al., 1990). Because the geometry of the anthranilate ring is similar to that of a pyrimidine, it seems likely that in TrpD this region contributes to the anthranilate-binding site.

The middle conserved alignment block II in Figure 2 belongs to the large α/β domain of DeoA and contains the phosphate-binding site (Walter et al., 1990). This region includes 2 flexible loops between β1A and α5, and between β2A and α6, both of which contribute to phosphate binding. The conservation in the first of these loops, which bears general resemblance to the phosphate-binding loops in other classes of enzymes, e.g., ATPases and kinases (Saraste et al., 1990; Schulz, 1992; Koonin, 1993), is particularly striking (Fig. 2; this loop is dramatically altered in *E. coli* YbiB, suggesting that despite the highly significant similarity to TrpD, this protein may lack the phosphoribosyltransferase activity). This segment belongs to one of the regions of highest conservation in the complete alignment of TrpD (Kim et al., 1993). Previous sequence comparisons and structural modeling indicated that this motif appears to be conserved in almost all phosphoribosyltransferases and is located in a β-strand–loop–α-helix unit within the putative structural core (Argos et al., 1983; Busetta, 1988; de Boer & Glickman, 1991; Fig. 3). It has been speculated by these authors that the conserved motif is part of the PRPP-binding site. The conservation in thymidine phosphorylases that do not interact with PRPP rather indicates that it is a specific form of the phosphate-

```
sec. struct.     bbbbbllllaaaaaa
DeoA Ec      79  NGPIVDKHSTGGVGDV TYPH
TYPH Mp      76  KKILIDKHSTGGIGDK
TYPH hum    109  RQQLVDKHSTGGVGDK

TrpG Ec     270  DYLFADIVGTGGDGSN AnthrPRT
TrpD Bl      76  GAGLLDSAGTGGDGAN
TrpD Ll      71  LTNAMDNCGTGGDRSF
TrpD Bs      70  LPDIVDTCGTGGDGIS
TrpD Tm     320  SPRTVDTCGTGGDGFG
TrpD Mt      74  SMRVVDACGTGGDRFK
TrpD Hv      72  AARSSDTAGTGGDDYN
TrpD At     176  LVDAVDIVGTGGDGAN
TrpD Sc     101  GPVILDIVGTGGDGQN

HIS1_ECO    163  ADAICDLVSTGATLEA ATP-PRT   P10366
HIS1_LAC    149  ADAIVDIVETGNTLSA           Q02129
HIS1_YEA    164  GDAIVDLVESGETMRA           P00498

APT_ECOL    119  VLVVDDLLATGGTIEA APRT      P07672
YSCAPRT_1   123  VVVVDDVLATGGTAYA           L14434
LEIADPH_1   139  VVLIDDVLATGGTALS           L25411
APT_ARAT    122  AIIIDDLIATGGTLAA           P31166
APT_HUMA    120  VVVVDDLLATGGTMNA           P07741

S19720       94  VLVVEDIIDTGHTISK UPRT      S19720
BCPYRQP_1    97  VILVDDVLFTGRTVRA           X76083
PYR5_BOV    116  CLIIEDVVSSGSSVWE           P31754
UPP_ECOL    123  ALIVDPMLATGGSVIA           P25532
STRFTFA_1   124  IFVVDPMLATGGSAIL           L07793
FUR1_YEA    165  VFLLDPMLATGGSAIM           P18562

XGPT_ECO     81  FIVIDDLVDTGGTAVA HPRT      P00501
HPRT_VIB     91  VLIVEDIIDTGNTLNK           P18134
HPRT_PLA    137  VLIVEDIIDTGKTLVK           P20035
HPRT_SCH    187  VLVVEDIIDTGKTITK           P09383
TRBHGPRT    104  VLIVEDIVDTALTLNY           L07486
LEIHGPT_    118  ILIVEDIVDSAITLQY           L25412
HPRT_HUM    127  VLIVEDIIDTGKTMQT           P00492

PYRE_ECO    117  VMLVDDVITAGTAIRE OPRT      P00495
PYRE_BAC    119  TVVIEDLISTGGSVLE           P25972
NGRUMPA_    122  CILIEDVITSGASIVE           L08073
CET07C4_    123  LILIEDVVTTGGSILD           Z29443
PYRE_YEA    125  ILIIDDVMTAGTAINE           P13298
PYRX_YEA    128  VLIIDDVMTAGTRINE           P30402
ATPYR3FA    111  CLIIEDLVTSGASVLE           X71842
PYR5_DIC    114  VLVVEDLVTSGASVLE           P09556
PYR5_DRO    117  CLIVEDVVTSGSSILD           Q01637
PYRE_CRY    123  IVIIDDVLTSGKAIRE           P18132
PYR5_HUM    116  CLIIEDVVTSGSSVLE           P11172

PUR1_ECO    360  VLLVDDSIVRGTTSEQ AmPRT     P00496
PUR1_BAC    349  VVMVDDSIVRGTTSRR           P00497
PUR1_YEA    366  VLIVDDSIVRGTTSKE           P04046
JC1414      382  IVLVDDSIVRGNTISP           JC1414

KPRS_ECO    213  CVLVDDMIDTGGTLCK RPPK      P08330
KPRS_YEA    318  AIILDDMIDRPGSFIS           P32895
LEIPRPP_    263  CIIVDDMIDTGGTLVK           M76553
KPR1_HUM    213  AILVDDMADTCGTICH           P09329
SCDNAPRS    216  CLLIDDMADTCGTLVK           X75075

RP4TRANO    135  YIVVDDTLTMGGTIAS           L10330
YORF_HAE    190  VALVDDVITTGSTLNE           P31773
YPYB_BAC     98  VILVDDVLYTGRTVRA           P25982
consensus        hhhhDDhh.TG.Th..
                   E    S  S
```

Fig. 3. Conservation of the phosphate-binding motif in thymidine phosphorylases and various groups of phosphoribosyltransferases. Closely related sequences are omitted. The consensus shows amino acid residues conserved in the majority of the enzyme groups. The motif could not be identified in nicotinate phosphoribosyltransferase and quinolinate phosphoribosyltransferase. Secondary-structure prediction is based on the experimentally determined 3D structure of DeoA, our prediction for anthranilate phosphoribosyltransferases, and the published model of the conserved phosphoribosyltransferase core (Busetta, 1988; de Boer & Glickman, 1991). The designations are as in Figures 1 and 2. Phosphoribosyltransferases: AnthrPRT, anthranilate; ATP-PRT, ATP; APRT, adenine; UPRT, uracil; HPRT, hypoxanthine (guanine); OPRT, orotate; AmPRT, glutamine; RPPK, ribose pyrophosphate kinases (PRPP synthetases).

binding loop, the function of which is not limited to PRPP binding. It has to be emphasized that the conservation of the phosphate-binding site between family II and other phosphoribosyltransferases is very limited, barely detectable by automatic methods of sequence analysis, and becomes apparent only upon comparison of multiple alignments for different families (Fig. 3).

Concluding remarks: Implications for enzyme evolution

The present analysis of the relationships between the amino acid sequences of nucleosidases and phosphoribosyltransferases showed that in these enzymes sequence similarity does not necessarily reflect similarity between the catalyzed reactions. Enzymes that catalyze the same reaction and may be even capable of utilizing identical substrates may show no appreciable sequence similarity to each other, like uridine phosphorylase (Udp) and thymidine phosphorylase (DeoA), or very weak similarity at the level of a degenerate conserved motif, like anthranilate phosphoribosyltransferases and other groups of phosphoribosyltransferases. Functional convergence may account for the analogous specificity in at least some of these cases (Doolittle, 1994).

In contrast, enzymes that catalyze very different reactions may share significant sequence similarity and probably have evolved by divergence from a common ancestor. Examples of such unexpected similarities were found in both enzyme families described here. In family I, there is a clear distinction between the reactions catalyzed by Amn and such proteins as Udp and DeoD, with the former being a hydrolase and the latter being phosphotransferases (Table 1); thus, this family unites enzymes that formally even belong to different classes. Obviously, however, the chemical moieties involved in both reactions are the same, i.e., phosphate, ribose, or deoxyribose, and a purine or pyrimidine base. This identity of the chemical constituents may account for the conservation of the enzyme structure. Essentially the same pertains to thymidine phosphorylases and anthranilate phosphoribosyltransferase in family II that catalyze superficially very different reactions and are involved in different biochemical pathways (salvage pathway of nucleotide metabolism and amino acid biosynthesis, respectively). Again, however, chemically, the ingredients are very similar, namely phosphate, ribose, or deoxyribose, and a nitrogen-containing compound with a single aromatic ring (thymidine or anthranilate, respectively).

Significant sequence conservation between enzymes that catalyze different reactions may be widespread. Relevant examples include haloalkane dehalogenase and epoxide hydrolase (Janssen et al., 1989; Ollis et al., 1992); creatinase, methionyl amino peptidase, and prolidases (Murzin, 1993); asparagine synthetase and aspartyl tRNA synthetase (Furukawa et al., 1992); and putative active centers of numerous phosphoesterases that hydrolyze a wide variety of substrates (Koonin, 1994).

Materials and methods

Sequences and databases

Amino acid and nucleotide sequences were retrieved from the nonredundant sequence database (NRDB) that is constructed by merging nonidentical entries from SWISS-PROT, PIR, and translated versions of GenBank and EMBL sequence databases at the National Center for Biotechnology Information (NIH).

Computer-assisted sequence analysis

Amino acid sequences were compared with the NRDB using programs based on the BLAST algorithm (Altschul et al., 1990). The BLASTP program was used to screen the amino acid sequence database and the TBLASTN program was used to screen the conceptual translation of the nucleotide sequence database in 6 reading frames. The BLAST algorithm provides the significance estimate for each ungapped pairwise alignment produced in the database search using the statistical theory of asymptotic extremal distribution for high-scoring segments (Karlin & Altschul, 1990, 1993; Altschul et al., 1994). For all database searches, the amino acid substitution matrix BLOSUM62 (Henikoff & Henikoff, 1992, 1993) was employed. Compositionally biased segments in the query sequences that may produce spurious "hits" in database searches were detected and masked using the SEG program (Altschul et al., 1994; Wootton & Federhen, 1993).

Multiple amino acid sequence alignments were constructed using the MACAW program (Schuler et al., 1991). MACAW produces alignments consisting of ungapped blocks separated by unaligned spacers of variable length by parsing compatible high-scoring segments detected by pairwise comparison of all sequences in a set. The statistical significance of each block is calculated separately using a generalization of the theory for pairwise alignments (Karlin & Altschul, 1990; Schuler et al., 1991). The boundaries of the blocks are adjusted so as to maximize the significance. The probability of obtaining a block by chance increases with the growth of the "search space." Usually, the search space is set to be equal to the product of the lengths of the compared sequences. Therefore, MACAW actually provides significance estimates given that the set of sequences under comparison has been initially delineated using some independent criteria. In addition, MACAW may overestimate the significance if a subset of closely related sequences is included in the analysis. To alleviate this problem, only 1 representative of each of such subsets was included in the calculations.

Screening of the NRDB for amino acid patterns (regular expressions) was performed using the program PAST (R.L. Tatusov, unpubl.).

Protein secondary structure was predicted using the recently developed neural network method (Rost & Sander, 1993).

Acknowledgment

We are grateful to Dr. Roman L. Tatusov for providing unpublished computer programs.

References

Altschul SF, Boguski MS, Gish W, Wootton JC. 1994. Issues in searching molecular sequence databases. *Nature Genet 6*:119–129.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Argos P, Hanei M, Wilson JM, Kelley WN. 1983. A possible nucleotide-binding domain in the tertiary fold of phosphoribosyltransferases. *J Biol Chem 258*:6450–6457.

Barton GJ, Ponting CP, Spraggon G, Finnis C, Sleep D. 1992. Human platelet-derived endothelial cell growth factor is homologous to *Escherichia coli* thymidine phosphorylase. *Protein Sci 1*:688–690.

Brun YV, Breton R, Lanouette P, Lapointe J. 1990. Precise mapping and

comparison of two evolutionarily related regions of the *Escherichia coli* K-12 chromosome. Evolution of *valU* and *lysT* from an ancestral tRNA operon. *J Mol Biol 214*:825–843.

Busetta B. 1988. The use of folding patterns in the search of protein structure similarities: A three-dimensional model of phosphoribosyltransferases. *Biochim Biophys Acta 957*:21–33.

Clausen S, Apel K. 1991. Seasonal changes in the concentration of the major storage protein and its mRNA in xylem ray cells of poplar trees. *Plant Mol Biol 17*:669–678.

Coleman GD, Chen THF, Fuchigami L. 1992. Complementary DNA cloning of poplar bark storage protein and control of its expression by photoperiod. *Plant Physiol 98*:687–693.

Crawford IP. 1989. Evolution of a biosynthetic pathway: The tryptophan paradigm. *Annu Rev Microbiol 43*:567–600.

Davis JM, Egelkrout EE, Coleman GD, Chen THH, Haissig BE, Riemenschneider DA, Gordon MP. 1993. A family of wound-induced genes in *Populus* share common features with genes encoding vegetative storage proteins. *Plant Mol Biol 23*:135–143.

de Boer JG, Glickman BW. 1991. Mutational analysis of the structure and function of the adenine phosphoribosyltransferase enzyme of Chinese hamster. *J Mol Biol 221*:163–174.

Doolittle RF. 1994. Convergent evolution: The need to be explicit. *Trends Biochem Sci 19*:15–18.

Ealick SE, Rule SA, Carter DC, Greenhough TJ, Babu YS, Cook WJ, Habash J, Heliwell JR, Stoeckler JI, Parks RE, Chen S, Bugg CE. 1990. Three-dimensional structure of human erythrocytic purine nucleoside phosphorylase at 3.2 Å resolution. *J Biol Chem 265*:1812–1820.

Furukawa T, Yoshimura A, Sumizawa T, Haraguchi M, Akiyama SI, Fukui K, Hinchman SK, Henikoff S, Schuster SM. 1992. A relationship between asparagine synthetase A and aspartyl tRNA synthetase. *J Biol Chem 167*:144–149.

Henikoff S, Henikoff J. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA 89*:10915–10919.

Henikoff S, Henikoff J. 1993. Performance evaluation of amino acid substitution matrices. *Proteins Struct Funct Genet 17*:49–61.

Ishizawa M, Yamada Y. 1992. Angiogenic factor. *Nature (Lond) 356*:668.

Janssen DB, Pries F, van der Ploeg J, Kazemier B, Terpstra P, Witholt B. 1989. Cloning of 1,2-dichloroethane degradation genes of *Xanthobacter autotrophicus* GJ10 and expression and sequencing of the *dhlA* gene. *J Bacteriol 171*:6791–6799.

Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA 87*:2264–2268.

Karlin S, Altschul SF. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci USA 90*: 5873–5877.

Kim CW, Markiewicz P, Lee JJ, Schierle CF, Miller JH. 1993. Studies of the hyperthermophile *Thermotoga maritima* by random sequencing of cDNA and genomic libraries. *J Mol Biol 231*:960–981.

Koonin EV. 1993. A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif. *J Mol Biol 229*:1165–1174.

Koonin EV. 1994. Conserved sequence pattern in a wide variety of phosphoesterases. *Protein Sci 3*:356–368.

Leung HB, Kvalnes-Krick KL, Meyer SL, de Riel JK, Schramm VL. 1989. Structure and regulation of the AMP nucleosidase gene (*amn*) from *Escherichia coli*. *Biochemistry 28*:8726–8733.

Leung HB, Schramm VL. 1980. Adenylate degradation in *Escherichia coli*. The role of AMP nucleosidase and properties of the purified enzyme. *J Biol Chem 255*:10867–10874.

Lin ECC. 1987. Dissimilatory pathways for sugars, polyols, and carboxylates. In: Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE, eds. *Escherichia coli and Salmonella typhimurium. Cellular and molecular biology*. Washington, D.C.: American Society for Microbiology. pp 244–284.

Murzin AG. 1993. Can homologous proteins evolve different enzymatic activities? *Trends Biochem Sci 18*:403–405.

Neuhard J, Nygaard P. 1987. Purines and pyrimidines. In: Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE, eds. *Escherichia coli and Salmonella typhimurium. Cellular and molecular biology*. Washington, D.C.: American Society for Microbiology. pp 445–473.

Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschueren KHG, Goldman A. 1992. The alpha/beta hydrolase fold. *Protein Eng 5*:197–211.

Parsons TJ, Bradshaw HD, Gordon MP. 1989. Systemic accumulation of specific mRNAs in response to wounding in poplar trees. *Proc Natl Acad Sci USA 86*:9851–9855.

Piatigorsky J. 1993. Puzzle of crystallin diversity in eye lenses. *Dev Dyn 196*:267–272.

Pittard AJ. 1987. Biosynthesis of aromatic amino acids. In: Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE, eds. *Escherichia coli and Salmonella typhimurium. Cellular and molecular biology*. Washington, D.C.: American Society for Microbiology. pp 368–394.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Saraste M, Sibbald PR, Wittinghofer A. 1990. The P-loop – A common motif in ATP- and GTP-binding proteins. *Trends Biochem Sci 15*:430–434.

Schuler GD, Altschul SF, Lipman DJ. 1991. A workbench for multiple alignment construction and analysis. *Proteins Struct Funct Genet 9*:180–190.

Schulz GE. 1992. Binding of nucleotides by proteins. *Curr Opin Struct Biol 2*:61–67.

Stout JT, Caskey CT. 1985. HPRT: Gene structure, expression and mutation. *Annu Rev Genet 19*:127–148.

Tritz GJ. 1987. NAD biosynthesis and recycling. In: Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE, eds. *Escherichia coli and Salmonella typhimurium. Cellular and molecular biology*. Washington, D.C.: American Society for Microbiology. pp 557–566.

Walter MR, Cook WJ, Cole LB, Short SA, Koszalka GW, Krenitsky TA, Ealick SE. 1990. Three-dimensional structure of thymidine phosphorylase from *Escherichia coli* at 2.8 Å resolution. *J Biol Chem 265*:14016–14022.

White HB. 1982. Biosynthetic and salvage pathways of pyridine nucleotide coenzymes. In: Everse J, Anderson BM, You KS, eds. *Pyridine nucleotide coenzymes*. New York: Academic Press. pp 1–17.

Winkler ME. 1987. Biosynthesis of histidine. In: Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE, eds. *Escherichia coli and Salmonella typhimurium. Cellular and molecular biology*. Washington, D.C.: American Society for Microbiology. pp 395–411.

Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem 17*:149–163.

Wu JJ, Schuch R, Piggot PJ. 1992. Characterization of a *Bacillus subtilis* sporulation operon that includes genes for an RNA polymerase sigma factor and for a putative DD-carboxypeptidase. *J Bacteriol 174*:4885–4892.

Wurgler SM, Richardson CC. 1990. Structure and regulation of the gene for dGTP triphosphohydrolase from *Escherichia coli*. *Proc Natl Acad Sci USA 87*:2740–2744.

Yoshimura A, Kuwazuru Y, Furukawa T, Yoshida H, Yamada K, Akiyama S. 1990. Purification and tissue distribution of human thymidine phosphorylase: High expression in lymphocytes, reticulocytes and tumors. *Biochim Biophys Acta 1034*:107–113.