PROTEINS: Structure, Function, and Genetics 20:347–355 (1994)

A P-Loop-Like Motif in a Widespread ATP Pyrophosphatase Domain: Implications for the Evolution of Sequence Motifs and Enzyme Activity

Peer Bork[1,2] and Eugene V. Koonin[3]

[1]*European Molecular Biology Laboratory, 69012 Heidelberg, and* [2]*Max-Delbrück-Center, 13189 Berlin-Buch, Germany;* [3]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894*

ABSTRACT **A conserved amino acid sequence motif was identified in four distinct groups of enzymes that catalyze the hydrolysis of the α–β phosphate bond of ATP, namely GMP synthetases, argininosuccinate synthetases, asparagine synthetases, and ATP sulfurylases. The motif is also present in *Rhodobacter capsulata* AdgA, *Escherichia coli* NtrL, and *Bacillus subtilis* OutB, for which no enzymatic activities are currently known. The observed pattern of amino acid residue conservation and predicted secondary structures suggest that this motif may be a modified version of the P-loop of nucleotide binding domains, and that it is likely to be involved in phosphate binding. We call it PP-motif, since it appears to be a part of a previously uncharacterized ATP *pyro*-*phophatase* domain. ATP sulfurylases, NtrL, and OutB consist of this domain alone. In other proteins, the pyrophosphatase domain is associated with amidotransferase domains (type I or type II), a putative citrulline-aspartate ligase domain or a nitrilase/amidase domain. Unexpectedly, statistically significant overall sequence similarity was found between ATP sulfurylase and 3′-phosphoadenosine 5′-phosphosulfate (PAPS) reductase, another protein of the sulfate activation pathway. The PP-motif is strongly modified in PAPS reductases, but they share with ATP sulfurylases another conserved motif which might be involved in sulfate binding. We propose that PAPS reductases may have evolved from ATP sulfurylases; the evolution of the new enzymatic function appears to be accompanied by a switch of the strongest functional constraint from the PP-motif to the putative sulfate-binding motif.**

Key words: **ATP hydrolysis, homology, sulfate metabolism, motif evolution**

INTRODUCTION

The concept of domains, independent structural and often functional units in proteins, has been promoted by the seminal discovery of a conserved dinucleotide-binding fold in the three-dimensional structures of otherwise very different dehydrogenases.[1] With the increasing amount of sequence data, a multidomain structure is revealed for more and more metabolic enzymes. In contrast to extracellular mosaic proteins, they mostly contain a very small number of functional units and the respective structures are unlikely to be the result of exon shuffling (for review, see ref. 2).

Distinct types of nucleoside triphosphate (NTP)-utilizing domains are among the most widespread "movable" enzyme domains. The majority of NTP-hydrolyzing enzymes cleave the β–γ bond of NTP; the best characterized and probably largest group contains the so-called Walker NTP-binding motifs A and B.[3–8] X-Ray analysis of several enzymes has confirmed that these two motifs participate in ATP hydrolysis (reviewed in refs. 5 and 8). The A motif contains the glycine-rich "P-loop" preceded by a hydrophobic β-strand and succeeded by an α-helix, and accommodates the pyrophosphate moiety of NTP.[4] The B motif contains a hydrophobic β-strand terminated by a negatively charged residue and binds NTP via Mg^{2+}. The A motif contains the strongest sequence signal and specific versions of this motif can be used as signatures for the identification of large families of NTPases (e.g., refs. 4, 6–8 and references therein).

Several other types of nucleotide-binding domains involved in the cleavage of the β–γ bond of ATP have been characterized, e.g., protein kinases,[9] the

Abbreviations: APS, adenosine-5′-phosphosulfate; Assy, argininosuccinate synthetases; Asn, asparagine synthetases; GATase, glutamine amidotransferase; NTP, nucleotide triphosphate; ORF, open reading frame; PP-motif, pyrophosphate-binding motif; PAPS, 3′-phosphoadenosine 5′-phosphosulfate.

Received May 12, 1994; revision accepted August 16, 1994.

Address reprint requests to Eugene V. Koonin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894.

superfamily of actin, hsp70, and sugar kinases[10] or carbamoyl transferases and NDP kinases.[11] Although they all contain conserved motifs with a glycine-rich loop that follows an interior β-strand, the structural context is different and they have apparently evolved independently.[12]

Enzymes that cleave the α–β bond in ATP, though less numerous, are also involved in various reactions and do not have a single phylogenetic origin. This is, for example, revealed by the two distinct topologies found among aminoacyl-tRNA synthetases.[13–15] Another indication of the variety of catalytic mechanisms among the ATP pyrophosphatases is the use of cofactors such as CoA (e.g., acyl-CoA synthetases) by some of them and the covalent binding of AMP (e.g., DNA and RNA ligases) to various enzymes (for an overview of sequenced ATP pyrophosphatases, see the enzyme database[16]).

Here we describe a putative novel pyrophosphatase domain characterized by a highly conserved motif. It is found in several groups of enzymes cleaving the α–β phosphodiester bond in ATP and appears to be a modified version of the P-loop found in β–γ NTPases. One of these pyrophosphatases, ATP sulfurylase, has a significant overall similarity to 3′-phosphoadenosine 5′-phosphosulfate (PAPS) reductases, sulfate-binding proteins which do not have ATPase activity. The P-loop-like motif is drastically modified in PAPS reductases; sulfate-binding proteins including ATP sulfurylase are, however, characterized by another, distinct conserved motif. Thus, evolution of a new protein function apparently can be accompanied by a shift of the principal selection pressure from one conserved motif to another.

METHODS

Amino acid sequences were extracted from protein databases or obtained by translation of nucleotide databases supported at the National Center for Biotechnology Information (NIH) and at EMBL.

Our strategy for delineating protein families and locating conserved motifs combines iterative database screening using methods for detection of pairwise similarity and motif search with construction and inspection of multiple alignments.[17]

Initial database searches are performed using the programs of the BLAST series[18,19] after masking compositionally biased segments using the SEG program.[19,20] Database searches for conserved motifs are carried out using the programs PROPAT[21] and MoST.[22] Briefly, under the MoST procedure, multiple alignment blocks are constructed by parsing consistent segments from ungapped pairwise alignments produced by a BLAST search. These blocks are converted into position-dependent weight matrices using a mixture of nine Dirichlet distributions[23] to approximate the probabilities of different amino acid residues in the alignment columns. It has been shown[22] that the use of the Dirichlet distributions results in significantly more sensitive weight matrices than simple averaging of the amino acid weights implemented in the PROFILE method.[24] Using these weight matrices, scores are computed for all segments of corresponding length in the amino acid sequence database, and the distribution of scores obtained is compared with the theoretically expected distribution. This procedure allows a precise evaluation of the statistical significance of the alignment between a candidate segment from the database with the weight matrices. The ratio of the expected to the observed number of sequence segments with a given score is used as the cut-off in database searches. The segments with scores exceeding the cut-off are added to the original alignment block and the process is iterated automatically until convergence (for details see ref. 22).

All potentially related proteins identified by motif searches are subjected to new BLAST searches as a control and in order to identify related proteins that might contain modified motifs. Multiple alignments are constructed using the programs ClustalW (a new version of ClustalV[25]; D. Higgins, J. Thompson, and T. Gibson, personal communication) and MACAW.[26] The latter program allows determination of the optimal boundaries of conserved blocks and evaluation of their statistical significance, i.e., the probability of finding these blocks by chance. Protein secondary structure elements are predicted using the PHD program that has an average accuracy of over 71% in a three-state prediction.[27]

RESULTS AND DISCUSSION

A Widespread Motif Probably Involved in Phosphate Binding by α–β Bond-Cleaving Enzymes

This analysis was initiated by searching the amino acid sequence databases for similarity to the *E. coli* GMP synthetase (GuaA) using the BLASTP program. The N-terminal part of GuaA was masked, given its known similarity to other amidotransferases.[28,29] The C-terminal portion matched not only several GMP synthetases from other organisms, but also asparagine synthetases and AdgA proteins from *Rhodobacter capsulatus* and yeast (Fig. 1a). Although these alignments were not highly significant statistically, they were consistent, i.e., they all matched the same region of the GMP synthetase. Thus, a single block could be constructed that was converted into a position-dependent weight matrix and used for iterative database search (see Methods).

The search resulted in the detection of a unique motif that is conserved in four diverse groups of enzymes, namely (1) GMP synthetases, (2) asparagine synthetases, (3) argininosuccinate synthetases, and (4) ATP sulfurylases. It is also present in several proteins without known enzymatic activity, such as AdgA, NrtL, OutB, and an uncharacterized open

```
a)
                  bbbbbb    aaaaaaaaaa
Guaa_Dicdi  268 KKVLVLVSGGVDSTVCAALISKAIGPEN    P32073
Guaa/Yeast  224 AEVIGAVSGGVDSTVASKLMTEAIGDRF    X70397
Guaa_Ecoli  228 DKVILGLSGGVDSSVTAMLLHRAIGKNL    P29727
Guaa/Mycle    ? GHAICGLSGGVDSAVAAALVQRAIGDRL    U00015
Guaa_Bacsu  220 KQVLCGLSGGVDSSVVAVLIHKAIGDQL    P29727
AdgA/Rhoca  285 SRVVLGLSGGIDSALVAVIAADALGAGN    X59399
AdgA/Yeast  354 TGFFLPLSGGIDSCATAMIVHSMCRLVT    U10556
Outb_Bacsu   40 KGFVLGISGGQDSTLAGRLAQLAVESIR    P08164
Ntrl_Ecoli   40 KSLVLGISGGQDSTLAGKLCQMAINELR    P18843
P486/Bacsu   30 ATIIVGVSGGPDSMALLHALHTLCGRSA    D26185
Assy_Human    5 GSVVLAYSGGLDTSCILVWLKEQGYDVI    P00966
Assy_Yeast    4 GKVCLAYSGGLDTSVILAWLLDQATEVV    P22768
Assy_Metba    3 KKVALAYSGGLDTSVCIPILKEKYGYDE    P13257
Assy_Metva    4 KIAVLAYSGGLDTSCCLKLLEDKYNYKV    P13256
Assy_Strco   18 ERVGIAFSGGLDTSVAVAWMRDKGAVPC    P24532
Assy_Ecoli   12 QRIGIAFSGGLDTSAALLWMRQKGAVPY    P22767
Asns_Mesau  250 RRIVCLLSGGLDSSLVASSLLKQLKEAQ    P17714
Asnl_Pea    226 VPFGVLLSGGLDSSLVASVTARYLAGTK    P19251
Asnb_Ecoli  228 VPYGVLLSGGLDSSIISAITKKYAARRV    P22106
Cysd_Ecoli   28 SNPVMLYSIGKDSSVMLHLARKAFYPGT    P21156
Nodp_Rhime   25 SNPVVLYSIGKDSSVLLHLAMKAFYPAK    P13441
Nodp_Azobr   27 TKPVLLYSIGKDSGVLLHLARKAFHPSP    P28603
                  hhhhhS G Dothhhhhh  t h

Cysh_Ecoli   45 GEYVLSSSFGIQAAVSLHLVNQ.IRPDI    P17854
CysH/Thiro   42 PQHVLSSSFGTQSAVMLHLVSR.QMPEI    Z23169
Cysh/Synsp   33 SGLVLSTSFGIQSAVMLHLATQ.VQPDI    M84476
Metg_Yeast   42 PHLFQTTAFGLTGLVTIDMLSK.LSEKY    P18408
Y334/Bp186   11 TINIVSVSGGKDSLAQWILAVENDVPRT    X53318
Y290_Lambd    1 MINVVSFSGGRTSAYLLWLMEQKRRAGK    P03766
                   hho t G tthh h hh t

b)
                        aaaaaaaaaaaaaa
Guaa_Dicdi  424 RDSGRVVEPLKDYHKDEVRELGKSLGLSDSLVWRQPFPGP
Guaa/Yeast  367 NMKLKLIEPLRELFKDEVRHLGELLGIPHDLVWRHPFPGP
Guaa_Bacsu  356 DMQFELIEPLNTLFKDEVRALGTELGIPDEIVWRQPFPGP
Guaa/Mycle    ? NLRFKLVEPLRLLFKDEVRAVGRQLDLPEEIVARQPFPGP
Guaa_Ecoli  367 EMKMGLVEPLKELFKDEVRKIGLELGLPYDMLYRHPFPGP
Adga/Rhoca  412 GDMAGGYNPLKDLYKTRVFETCRWRNATHRPWMQAPAGEI
Adga/Yeast  537 DCSSADINPIGGISKTDLKRFIAYASKQYNMPILNDFLNA
Outb_Bacsu  173 GDGGADLLPLTGLTKRQGRTLLKELGAPERLYLKEPTADL
Ntrl_Ecoli  174 GDGGTDINPLYRLNKRQGKQLLAALACPEHLYKKAPTADL
                t   t h PL tL Ktth thht LthPtthhhttPhtt
```

Fig. 1. Alignment of motifs common to the described family of ATP pyrophosphatases. The PP-motif was delineated by scanning the amino acid sequence database with a position-dependent weight matrix derived from the block of aligned segments parsed from the BLAST output for the *E. coli* GuaA sequence. A cut-off ratio of 0.02 (expected number of selected segments divided by the observed number) was used. The second conserved motif was detected using the MACAW program. The first column contains the SWISSPROT codes[45] if available (underscores in the names). Protein designations: Asn, asparagine synthetase; Assy, argininosuccinate synthetase; Metg, PAPS reductase encoded by the yeast Met16 gene; P486, an uncharacterized ORF product from *Bacillus subtilis;* Y334 and Y290, putative bacteriophage proteins. The second column shows the position of the PP-motif in the respective sequence. Top line: predicted secondary structures (a, α-helix; b, β-strand); bottom line: consensus of the alignment (capitals, invariant amino acids; h, hydrophobic positions; o, S or T; t, turn-like or polar positions). Bold typing shows amino acid residues that are conserved in at least 60% of the sequences. The last column contains the database accession numbers of the sequences. (**a**) PP-motif and its modified version in sulfate-binding proteins; (**b**) a second conserved motif in a subset of the sequences.

reading frame (ORF) from *Bacillus subtilis* (hereafter P486 because of its length; Fig. 1). The characterized enzymes containing this conserved motif belong to anabolic pathways and have ATP pyrophosphatase activity, i.e., they hydrolyze the α–β bond in ATP. The energy of the phosphodiester bond in ATP is, however, transferred to a distinct bond in each reaction, e.g., a C–N bond in the case of aspar-

TABLE I. Summary of Enzymes Containing the PP-Motif

Protein	EC number	Reaction
NtrL/OutB*		ATP = AMP + PP
GMP synthetase (GuaA)	EC 6.3.5.2	XMP + L-Gln + H_2O + ATP = AMP + PP + L-Glu + GMP
AdgA*		Y-NH_2 + H_2O + ATP = AMP + PP + Y=O + NH_3
Argininosuccinate synthetase (Assy)	EC 6.3.4.5	L-citrullin + L-Asp + ATP = AMP + PP + L-argininosuccinate
Asparagine synthetase (Asn)	EC 6.3.5.4	L-Gln + L-Asp + ATP = AMP + PP + L-Glu + L-Asn
ATP sulfurylase (CysD/NodP)†	EC 2.7.7.4	SO_4^{2-} + ATP = AMP-SO_4 + PP
PAPS reductase (CysH/Met16)	EC 2.8.2.-	Thioredoxin + PAPS = thioredoxin-SO_3^- + PAP

*Proposal based on sequence similarities. NtrL/OutB are likely to be associated with another subunit carrying the energy-requiring function; for AdgA a more precise prediction is impossible due to the only remote similarities of the N-terminus to nitrilases and amidases.[39]

†The reaction involves hydrolysis of the α–β bond of ATP as the first step, whereas the second step is the sulfate transfer to AMP.[32]

agine and argininosuccinate synthetases, and a P–S bond in the case of ATP sulfurylases; the respective overall reaction schemes are also different (Table I).

GMP synthetase is involved in the biosynthesis of purine nucleotides and transfers the NH_4^+ group to XMP as a last step in the pathway (ref. 29 and references therein). Asparagine synthetase (glutamine hydrolyzing) that contains the conserved motif catalyzes an alternative way of asparagine biosynthesis by cleaving the amide bond of glutamine and transferring the NH_2 group to aspartate (Table I). Argininosuccinate synthetase is a major enzyme of the urea cycle, the most important ammonia detoxification mechanism. It attacks the amido group of aspartate and catalyzes its condensation with citrullin (Table I). ATP sulfurylase catalyses the first step in the inorganic sulfate activation pathway (Table I) yielding adenosine-5′-phosphosulfate (APS), one of the forms of activated sulfate. APS serves as a sulfate donor in various pathways such as cysteine biosynthesis, but is also involved in ATP synthesis in litotrophic bacteria.[30] The conserved motif is located in the small subunit, which is encoded by the *nodP* in nitrogen-fixing bacteria[30] and by *cysD* gene in *E. coli*,[31] and cleaves the α–β bond of ATP, resulting in AMP as an intermediate.[32]

For the other proteins containing the conserved motif (hereafter PP-motif for *pyrophosphatase* motif), no molecular details are known so far. Mutations of the *adgA* gene (for *a*mmonia-*d*ependent growth) in *Rhodobacter* species cause the unability of the bacteria to use a number of amino acids as a growth source.[33] These mutations can be complemented by the *E. coli* protein NtrL (*ntr*-like).[34] NtrL, in turn, is very similar to the *B. subtilis* OutB protein (for *out*growth spore factor), mutations of which are temperature-sensitive and show no outgrowth of spores after germination at the nonpermissive temperature. The gene also appears to be essential for vegetative growth.[35]

The probability of finding the PP-motif by chance in all these proteins (Fig. 1) is below 10^{-19} as computed using the MACAW program.[26] Secondary structure predictions of the PP-motif suggest a β–α–β organization (Fig. 1) that resembles several phosphate-binding motifs including the classical NAD/FAD-binding motif,[1] the Walker type NTP-binding motif A and its modified versions,[3–8] as well as phosphate-binding motifs in phosphoribosyltransferases,[36,37] thymidine phosphorylases,[37] and the so-called "firefly luciferase family."[38] Similarly to the PP-motif, these binding sites consist of a glycine-rich loop preceded by an interior hydrophobic β-strand and succeeded by an α-helix. Furthermore, all these motifs occur at the N-terminus of the respective domains. Since the PP-motif is the only conserved region among different groups of α–β bond-hydrolyzing ATPases, we predict that it is involved in the binding of the phosphate moiety of ATP. The conservation of this motif in GuaA, AdgA, NtrL, and OutB and its similarity to the Walker A motif has previously been noticed by Willison.[33]

A Movable ATP Pyrophosphatase Domain?

A conserved sequence motif usually corresponds to a functional site within a larger domain with a conserved topology. The existence of a defined ATP pyrophosphatase domain including the PP-motif is supported by the presence of a second motif that is conserved among GMP synthetases, NtrL, OutB, and AdgA (Fig. 1b). This motif is located about 130 amino acids from the PP-motif towards the C-terminus (Fig. 2). It has a probability of occurrence in all these proteins by chance of about 10^{-18}, i.e., it is also significant.

The proposal that the PP-motif resides in a distinct domain is consistent with the location of this motif and its surrounding regions within the proteins (Fig. 3). The motif is either located near the N-terminus (OutB, NtrL, ATP sulfurylase, argininosuccinate synthetase) or is immediately downstream from a functionally well-characterized domain. A glutamine amidotransferase (GATase) type I domain and a GATase type II domain precede the

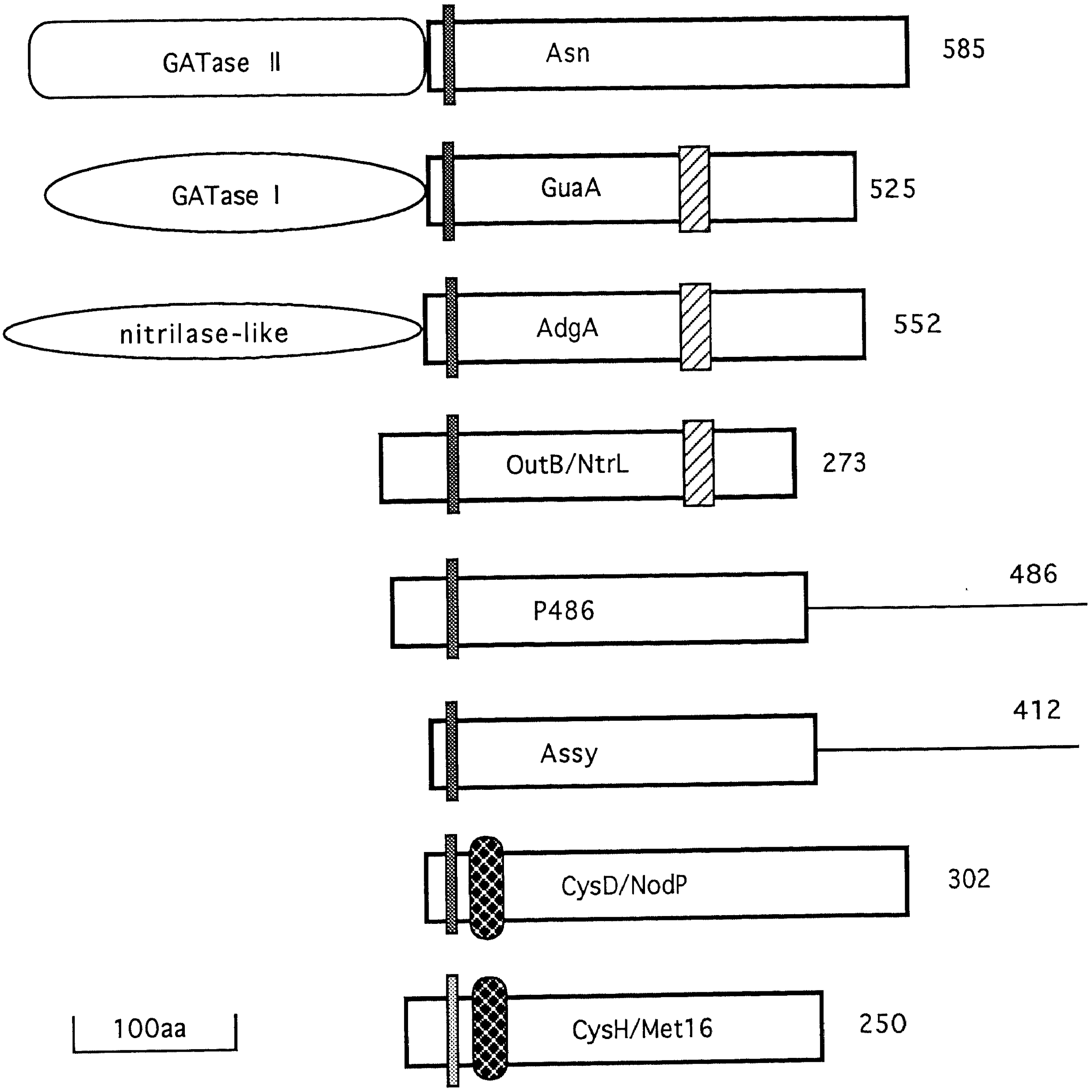

Fig. 2. Modular architecture of enzymes containing the PP-motif. Distinct enzymatic activities are denoted by different symbols. The location of the motifs in the ATP pyrophosphatase domain (rectangle) is indicated by small boxes (PP-motif, dotted; second motif, lined; DT motif, hatched). The modified PP-motif in sulfate-binding proteins is indicated by fewer dots. The approximate length (in amino acids) of the proteins is given. Note that single domain proteins exist which are related to the GATase and nitrilase domains shown here.

PP-motif in asparagine synthetases and GMP synthetases, respectively (Fig. 2). Thus, the same type of ATP pyrophosphatase domain is combined with two apparently unrelated types of GATases. AdgA has a large, functionally uncharacterized N-terminal segment that precedes the PP-motif. Comparative sequence analysis of this region revealed that it is a well-defined, C–N bond hydrolyzing domain related to nitrilases and amidases.[39] Both the two GATase domains and the N-terminal domain of AdgA act as individual proteins in other pathways and they can catalyze similar reactions without the need for the energy provided by the ATP pyrophosphatase domain. In argininosuccinate synthetases the domains appear to have the opposite order, with the putative ligase activity located downstream from the ATP pyrophosphatase domain (Fig. 2). Finally, the uncharacterized putative protein P486 from *B. subtilis* contains the PP-motif in the typical, N-terminal position, and judging from its size, may have a unique domain in its C-terminal portion (Fig. 2).

Thus in the majority of the proteins discussed here, the ATP pyrophosphatase domain apparently contains about 250–300 amino acids, with the PP-motif near its N-terminus, and is covalently linked

```
                   bbbbb      aaaaaaaaaa       bbbbb    aaaaaaaaaaaaaaa       aaaaaaa          aaaaaaa    aaaa
Cysd_Ecoli   28 SNPVMLYSIGKDSSVMLHLARKAFYPGTLPFPLLHVDTGWKFREMYEFRDRTAKAY -24- AKHTDIMKTEG - 0- LKQALNKYGFDAAFGGA
Nodp_Rhime   27 SNPVVLYSIGKDSSVLLHLAMKAFYPAKPPFPFLHVDTKWKFREMIEFRDRMAREL -24- NVHTHVMKTMG - 0- LRQALEKYGFDAALAGA
Nodp_Azobr   25 TKPVLLYSIGKDSGVLLHLARKAFHPSPVPFPLLHVDTGWKFREMIAFRDATVRRL -24- ALHTRVMKTEA - 0- LRQALDRHGFDAAIGGA
Cysh_Ecoli   45 GEYVLSSSFGIQAAVSLHLVNQ.IRPDI...PVILTDTGYLFPETYRFIDELTDKL -31- EKYNDINKVEP - 0- MNRALKELNAQTWFAGL
Cysh_Salty   44 GEYVLSSSFGIQAAVSLHLVNQ.IRPDI...PVILTDTGYLFPETYQFIDELTDKL -31- EKYNEINKVEP - 0- MNRALKELKAQTWFAGL
Cysh/Thiro   42 PQHVLSSSFGTQSAVMLHLVSR.QMPEI...PVILVDTGYLFPETYRLVDALTDRF -31- ERYNRLNKIDP - 0- MERALRDLDAGTWFAGL
Cysh/Synsp   33 SGLVLSTSFGIQSAVMLHLATQ.VQPDI...PVIWIDTGYLPTETYRFAAELTERL -32- NRYDQMRKVEP - 0- MNRALQELGATAWLSGV
Metg_Yeast   42 PHLFQTTAFGLTGLVTIDMLSK.LSEKYYMPELLFIDTLHHFPQTLTLKNEIEKKY -36- DKYDYLAKVEP - 0- AHRAYKELHISAVFTGR
Y334/Bp186   11 TINIVSVSGGKDSLAQWILAVENDVPR....TTVFADTGHEHSQTMEYLDYLESRL -44- AKALEILKPTG -33- VLPALEKYDEVILWQGV
Y290_Lambd    1 MINVVSFSGGRTSAYLLWLMEQKRRAGK.DVHYVFMDTGCEHPMTYRFVREVVKFW -43- KKYGTPYVGGA -13- CYDDHFGRGNYTTWIGI
                   Vh  ShG ts V h Lh t   P     thhhhDTGh  t hhth t h t h      t h  h K  t      h tAL tht  hhh Gh

                       aaaaaaaa         bbbbbb     aaaaaaaaa                bbbbbb               aaaaaaaaaaa
Cysd_Ecoli      RDEEKSRAKERIYSFR -25- GETIRVFPLSNWTEQDIWQYIWLENIDIVPLYL -9- DGMLMMIDDN -0- RIDLQPGEVIKKRMVRFRTLGCWP
Nodp_Rhime      RDEEKSRAKERIFSIR -25- GETMRVFPLSNWTEFDIWQYILREEIPIVPLYF -9- EGMLIMVDDD -0- RMPIQPEEEVTEQLVRFRTLGCYP
Nodp_Azobr      RDEEKSRAKERVFSIR -25- GESVRVFPLSNWTELDVWRYVAAQSIPVVPLYF -9- SGALIMVDDG -0- RLPLNPGETPEMRRVRFRTLGCYP
Cysh_Ecoli      REQSGSRANLPVLAIQ - 0- RGVFKVLPIIDWDNRTIYQYLQKHGLKYHPLWD -0- EGYLSVGDTH -3- KWEPGMAEEET.RFFGLKRECGLH
Cysh_Salty      REQSGSRAHLPVLAIQ - 0- RGVFKVLPIIDWDNRTVYQYLQKHGLKYHPLWD -0- QGYLSVGDTH -3- KWEPGMAEEET.RFFGLKRECGLH
Cysh/Thiro      RQQANSRAELPVLRRQ - 0- DGRIKFHPIIDWHRPRRARYLRRHDLPDHPLRD -0- QGYVSIGDVH -3- PLLPGMLEEET.RFFGIKRECGLH
Cysh/Synsp      RQQTAHRQSMEIVELK - 0- RDRYAIRPILGWHSRDVYQYLTAHDLPYHPLFD -0- QGYVTVGDWH -3- PLQADDSDERTTRFRGLKQECGLH
Metg_Yeast      KSQGSARSQLSIIEID - 2- NGILKINPLINWTFEQVKQYIDANNVPYNELLD -0- LGYRSIGDYH -3- PVKEG.EDERAGRWKGKARPSVEF
Y334/Bp186      AQESPARAALPMWEED - 4- PGLHVYRPILNWTHEDVFALAKRHGIKPNPLYQ -0- QGCSRVGCMP -9- AEIFARWPEEIARVAEWERLVAAC
Y290_Lambd      ADEPKRLKPKPGIRYL -37- RDEEGLQRVFNEVITGSH.VRDGHRETPKEIMY -0- RGRMSLDGIA -4- ENDYQALYQDMVRAKRFDTGSCSE
                ttttttRtt thht        tt   h Ph tW   t h Yh  tt t  PLh       Gh  h D t     t           Rhhthtt  h
```

Fig. 3. Alignment of ATP sulfurylases with PAPS reductases and related proteins. The nomenclature is as in Figure 1. The expected accuracy of prediction for the three N-terminal secondary structure elements surrounding the PP- and DT-motifs is higher than 82% as computed using the PHD program.[27] A database search using position-dependent matrixes derived from the "DT" motif did not reveal any additional related sequences.

to other enzymatic domains that catalyze the reactions, which may be coupled to ATP hydrolysis (Fig. 2). NtrL and OutB appear to be stand-alone versions of the ATP pyrophosphatase domain. We propose that they may be subunits of larger, noncovalent heterooligomeric enzyme complexes, in which the other subunits catalyze reaction(s) requiring the energy of the ATP hydrolysis. ATP sulfurylase, which does not contain additional domains either, utilizes ATP as the substrate for sulfurylation. In *E. coli,* this enzyme (CysD) is found as a tetramer composed of two heterodimers; the other, larger subunit (CysN) carries a regulatory GTPase activity.[32]

A Bridge Between Two Enzyme Classes

Database searches with the ATP sulfurylase sequences (CysD and NodP) revealed statistically significant similarity (probability of occurrence by chance below 10^{-7}, computed using BLAST) with 3′-phosphoadenosine-5′-phosphosulfate (PAPS) reductases (CysH in prokaryotes and Met16 in yeast). PAPS reductases catalyse the reduction of activated sulfate to sulfite, for example, in the context of cysteine and methionine biosynthesis. During the reaction, the sulfate is transferred to thioredoxin which probably serves as a thiol carrier.[40] As PAPS is formed by adenosine-5′-phosphosulfate (APS) kinase, PAPS reductase belongs to the same pathway with ATP sulfurylase (Table I).

Database searches with the sequences of ATP sulfurylases and PAPS reductases also detected moderate similarity with two uncharacterized bacteriophage proteins. This relationship is strongly supported by the multiple alignment of all these sequences (Fig. 3). Surprisingly, the most conserved feature of this rather diverse protein group is not the PP-motif, but another characteristic region (Fig. 3) around the invariant DT dipeptide (hereafter, DT-motif). The DT-motif follows an internal β-strand (predicted with an expected accuracy higher than 82%[27]; Fig. 3) and is located in a position analogous to the Mg^{2+}-binding motif (Walker B box) in classical ATP-binding sites (e.g., ref. 5). Whereas the DT-motif seems to be the hallmark of the sulfate-transferring proteins (Fig. 3), the PP-motif is dramatically modified in PAPS reductases and in the bacteriophage λ protein (Fig. 1). It was not detected by MoST and PROPAT motif searches based on alignments of the ATP pyrophosphatases described above, but was only identified after the multiple alignment of the sulfate-transferring proteins (Fig. 3).

These observations suggest the following evolutionary scenario. One of the widespread ATP pyrophosphatase domains (ATP sulfurylase) could have acquired the ability to bind APS at the PP-site as the result of the emergence of the DT-motif that may be spatially juxtaposed with the PP-motif. PAPS reductases could have then evolved from ATP sulfurylases using the ability to bind APS for developing a new enzymatic activity. This was accompanied by the loss of the ATP pyrophosphatase reaction. The PP-motif in PAPS reductases may now function as the PAPS-binding site whereas the DT-motif is likely to be involved in sulfate binding and transfer. The latter appears to be the principal functional site, with the strongest functional constraints restricting its diversification.

The two bacteriophage proteins belonging to the family are too divergent for a precise prediction of function, but the overall similarity and the conservation of the DT-motif suggest that bacteriophages 186 and λ encode enzymes involved in sulfate transfer. Thus, 12 years after the complete sequence of the bacteriophage λ DNA was determined[41] and with the bacteriophage biology known in great detail,[42] new protein functions are still predicted as the result of sequence comparison.

Motif Evolution and Evolution of Enzymatic Activity

The pattern of motif conservation in the ATP pyrophosphatases and PAPS reductases exposes several general aspects of enzyme evolution.

1. *Convergence of different protein folds toward a similar function.* There are several other families of ATP pyrophosphatases which do not contain the PP-motif and it is likely that they are not homologous to the family described here. Some of these enzyme families appear to contain other versions of the P-loop[6,36–38] that despite the predicted structural analogy, do not overlap with the PP-motif described here in any database searches. While it is difficult to rule out divergence at a very early stage of evolution, functional convergence due to certain structural requirements appears to be a plausible explanation for the evolution of at least some of the distinct versions of the P-loop.[4,12] Even within the functionally very similar aminoacyl-tRNA synthetases which also cleave the α–β bond in ATP, the two structurally distinct families clearly indicate functional convergence.[13–15]

In the same vein, neither yeast MET3 ATP sulfurylase nor AsnA asparagine synthetase from *E. coli* have sequence similarity to the proteins described here. This suggests that identical or similar activity has independently evolved in several evolutionarily unrelated enzymes families. Such functional convergence appears to be a frequent event rather than an exception in enzyme evolution.[12,37,43]

2. *Evolutionary mobility of functional units.* Once a functional unit is established, gene duplication, subsequent mutational modifications, and gene fusions provide a basis for the spread of such a domain. Thus, homologous domains are found in very different settings and, in particular, energy-providing

ATPase domains are coupled with the catalysis of different reactions that have to overcome an energy barrier. So far, the ATP pyrophosphatase domain described here has been found to be coupled to at least five chemically distinct reactions (Fig. 3; Table I).

3. *Sequence conservation only in functional sites.* Significant sequence conservation in the PP-motif contrasts the lack of similarity in other portions of the ATP pyrophosphatase sequences (apart from the second motif in GuaA, NtrL, OutB, and AdgA) although they may have an overall structural similarity. Whereas the common topology does not necessarily require conservation at the sequence level, functional sites such as the PP-motif are under much stronger constraints and are often the only discernable common features.

4. *Different functions of the same fold and motif evolution.* The sequence conservation between ATP sulfurylase and PAPS reductase, two different enzymes in the sulfate activation pathway, highlights the relationship between evolution of sequence motifs and evolution of the enzymatic activity. In the PAPS reductases, the PP-motif is significantly modified, perhaps because of the switch to a related but distinct substrate (PAPS instead of ATP/APS). The major functional constraints are apparently transferred to another site (DT motif) which may have a common function in both types of enzymes, e.g., sulfate-binding. One may speculate that proteins may exist which contain the DT-motif but not the PP-motif as they might have developed another binding site for a sulfate donor.

5. *Pathway extension by duplication and modification.* The probable evolution of PAPS reductase from ATP sulfurylase is an example of how metabolic pathways could have extended. Gene duplication and subsequent acquisition of the ability to bind different substrates by fine tuning sites that are already "approved" by evolution may be the starting point in the evolution of many new protein functions.

Sequence comparison studies are frequently impeded by the lack of intermediates and by rapid evolution of protein sequences upon functional changes. Here, we may have found a scenario in which the switch from a specific enzymatic activity to a new one can be traced. Support for our hypotheses could be provided by the determination of the three-dimensional structures of the respective enzymes; X-ray analysis of the GMP synthetase is already underway.[44]

ACKNOWLEDGMENTS

We thank Rebecca Wade and Matti Saraste for helpful comments and critical reading of the manuscript.

REFERENCES

1. Rossmann, M.G., Moras, D., Olsen, K.W. Chemical and biological evolution of a nucleotide-binding protein. Nature (London) 250:194–199, 1974.
2. Doolittle, R.F., Bork, P. Evolutionarily mobile modules in proteins. Sci. Am. 269(4):50–56, 1993.
3. Walker, J.E., Saraste, M., Runswick, M.J., Gay, N.J. Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases, and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. 1:945–951, 1982.
4. Saraste, M., Sibbald, P.R., Wittinghofer, A. The 'P-loop'—a common motif in ATP- and GTP-binding proteins. TIBS 15:430–434, 1990.
5. Schulz, G.E. Binding of nucleotides by proteins. Curr. Opin. Struct. Biol. 2:61–67, 1992.
6. Koonin, E.V. A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif. J. Mol. Biol. 229:1165–1174, 1993.
7. Koonin, E.V. A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. Nucleic Acids Res. 21:2541–2547, 1993.
8. Traut, W. The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide-binding sites. Eur. J. Biochem. 222:9–19, 1994.
9. Hanks, S.K., Quinn, A.M. Protein kinase catalytic domain database: Identification of conserved features of primary structure and classification of family members. Methods Enzymol. 200:38–62, 1991.
10. Bork, P., Sander, C., Valencia, V. An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsc70 heat shock proteins. Proc. Natl. Acad. Sci. U.S.A. 89:7290–7294, 1992.
11. Dumas, C., Lascu, I., Morera, S., Glaser, Fourme, R., Wallet, V., Lacombe, M-L., Veron, M., Janin, J. X-ray structure of nucleoside diphosphate kinase. EMBO J. 11:3203–3208, 19xx.
12. Doolittle, R.F. Convergent evolution: The need to be explicit. Trends Biochem. Sci. 19:15–18, 1994.
13. Cusack, S., Berthet-Colominas, C., Hartlein, M., Nassar, N., Leberman, R. A second class of synthetase structures revealed by X-ray analysis of E. coli seryl-tRNA synthetase at 2.5 Å. Nature (London) 347:249–255, 1990.
14. Nagel, G.M., Doolittle, R.F. Evolution and relatedness in two aminoacyl-tRNA synthetase families. Proc. Natl. Acad. Sci. U.S.A. 88:8121–8125, 1991.
15. Delarue, M., Moras, D. The aminoacyl-tRNA synthetase family: Modules at work. BioEssays 15:675–687, 1993.
16. Bairoch, A. The enzyme data bank. Nucleic Acids Res. 21:3155–3156, 1993.
17. Koonin, E.V., Bork, P., Sander, C. Yeast chromosome III: New gene functions. EMBO J. 13:493–503, 1994.
18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D. Basic local alignment search tool. J. Mol. Biol. 215:403–410, 1990.
19. Altschul, S.F., Boguski, M.S., Gish, W., Wootton, J.C. Issues in searching molecular sequence databases. Nature Genet. 6:119–129, 1994.
20. Wootton, J.C., Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. 17:149–163, 1993.
21. Rohde, K., Bork, P. A fast, sensitive pattern-matching approach for protein sequences. CABIOS 9:183–189, 1993.
22. Tatusov, R.L., Altschul, S.F., Koonin, E.V. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. Proc. Natl. Acad. Sci. U.S.A., in press, 1994.
23. Brown, M., Hughey, R., Krogh, A., Mian, S.I., Sjolander, K., Haussler, D. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In "Proc. First Int. Conf. Intelligent Systems Mol. Biol." Hunter, L., Searls, D., Shavlik, J. (eds). Menlo Park, CA: AAAI Press, 1993:47–55.
24. Gribskov, M., McLachlan, A.D., Eisenberg, D. Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. 84:4355–4358, 1987.
25. Higgins, D., Bleasby, A.J., Fuchs, R. ClustalV: Improved

software for multiple alignment. CABIOS 8:189–191, 1992.

26. Schuler, G.D., Altschul, S.F., Lipman, D. A workbench for multiple sequence alignment construction and analysis. Proteins 9:180–190, 1991.
27. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70%. J. Mol. Biol. 232:584–599, 1993.
28. Zalkin, H., Argos, P., Narayana, S.V.L., Tiedemann, A.A., Smith, J.M. Identification of a *trp*G-related glutamine amide transfer domain in *E. coli* GMP synthetase. J. Biol. Chem. 260:3350–3354, 1985.
29. Mäntsälä, P., Zalkin, H. Cloning and sequence of Bacillus subtilis purA and guaA, involved in the conversation of IMP to AMP and GMP. J. Bacteriol. 174:1883–1890, 1992.
30. Schwedock, J., Long, S.R. ATP sulfurylase activity of the *nodP* and *nodQ* gene products of *Rhizobium meliloti*. Nature (London) 348:644–647, 1990.
31. Leyh, T.S., Vogt, T.F., Suo, Y. The DNA sequence of the sulfate activation locus from *E. coli* K-12. J. Biol. Chem. 267:10405–10410, 1992.
32. Liu, C., Martin, E., Leyh, T.S. GTPase activation of ATP sulfurylase: The mechanism. Biochemistry 33:2042–2047, 1994.
33. Willison, J.C. Biochemical genetics revisited: The use of mutants to study carbon and nitrogen metabolism in the photosynthetic bacteria. FEMS Microb. Rev. 104:1–38, 1993.
34. Albertini, A.M., Galizzi, Z. The *B. subtilis* outB gene is highly homologous to an *E. coli ntr*-like gene. J. Bacteriol. 172:5482–5485, 1990.
35. Albertini, A.M., Caramori, Henner, D., Ferrari, E., Galizzi, A. Nucleotide sequence of the *B. subtilis outB* gene and regulation of its expression. J. Bacteriol 169:1480–1484, 1987.
36. de Boer, J.G., Glickman, B.W. Mutational analysis of the structure and function of the adenine phosphoribosyltransferase enzyme of chinese hamster. J. Mol. Biol. 221:163–174, 1991.
37. Mushegian, A.R., Koonin, E.V. Unexpected sequence similarity between nucleosidases and phosphoribosyltransferases of different specificity. Protein Sci. 3:1081–1088, 1994.
38. Toh, H. Sequence analysis of firefly luciferase family reveals a conservative sequence motif. Protein Seq. Data Anal. 4:111–117, 1991.
39. Bork, P., Koonin, E.V. A new family of carbon-nitrogen hydrolases. Prot. Sci., 3:1344–1346, 1994.
40. Thomas, D., Barbey, R., Surdind-Kerjan, Y. Gene-enzyme relationship in the sulfate assimilation pathway of Saccharomyces cerevisiae. J. Biol. Chem. 265:15518–15524, 1990.
41. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., Petersen, G.B. Nucleotide sequence of bacteriophage λ DNA. J. Mol. Biol. 162:729–773, 1982.
42. Hendrix, R., Roberts, J., Stahl, F., Weisburg, R. (eds.). "Lambda II." New York: Cold Spring Harbor, 1983.
43. Bork, P., Sander, C., Valencia, A. Convergent evolution of similar enzymatic function on different protein folds: The hexokinase, ribokinase, and galactokinase families of sugar kinases. Protein Sci. 2:31–40, 1993.
44. Tesmer, J.J.G., Stemmler, T.L., Penner-Hahn, J.E., Davisson, V.J., Smith, J.L. Preliminary X-ray analysis of *Escherichia coli* GMP synthetase: Determination of anomalous scattering factors for a cysteinyl mercury derivative. Proteins 18:394–403, 1994.
45. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank, recent developments. Nucleic Acids Res. 21:3093–3096, 1993.