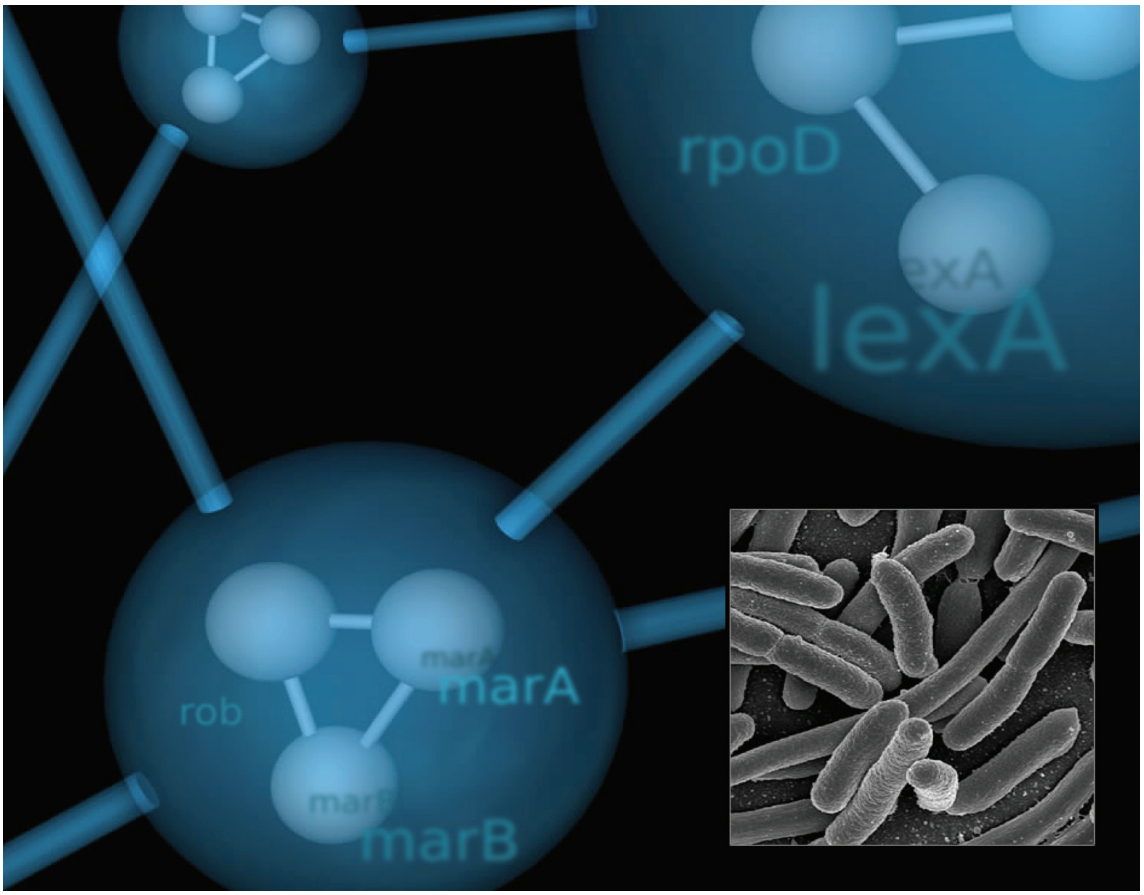
Molecular BioSystems

This article was published as part of the

Computational and Systems Biology themed issue

Please take a look at the full table of contents to access the other papers in this issue.



Amidoligases with ATP-grasp, glutamine synthetase-like and acetyltransferase-like domains: synthesis of novel metabolites and peptide modifications of proteins†‡

Lakshminarayan M. Iyer,[a] Saraswathi Abhiman,[a] A. Maxwell Burroughs[b] and L. Aravind*[a]

Received 28th August 2009, Accepted 28th August 2009
First published as an Advance Article on the web 13th October 2009
DOI: 10.1039/b917682a

Recent studies have shown that the ubiquitin system had its origins in ancient cofactor/amino acid biosynthesis pathways. Preliminary studies also indicated that conjugation systems for other peptide tags on proteins, such as pupylation, have evolutionary links to cofactor/amino acid biosynthesis pathways. Following up on these observations, we systematically investigated the non-ribosomal amidoligases of the ATP-grasp, glutamine synthetase-like and acetyltransferase folds by classifying the known members and identifying novel versions. We then established their contextual connections using information from domain architectures and conserved gene neighborhoods. This showed remarkable, previously uncharacterized functional links between diverse peptide ligases, several peptidases of unrelated folds and enzymes involved in synthesis of modified amino acids. Using the network of contextual connections we were able to predict numerous novel pathways for peptide synthesis and modification, amine-utilization, secondary metabolite synthesis and potential peptide-tagging systems. One potential peptide-tagging system, which is widely distributed in bacteria, involves an ATP-grasp domain and a glutamine synthetase-like ligase, both of which are circularly permuted, an NTN-hydrolase fold peptidase and a novel alpha helical domain. Our analysis also elucidates key steps in the biosynthesis of antibiotics such as friulimicin, butirosin and bacilysin and cell surface structures such as capsular polymers and teichuronopeptides. We also report the discovery of several novel ribosomally synthesized bacterial peptide metabolites that are cyclized *via* amide and lactone linkages formed by ATP-grasp enzymes. We present an evolutionary scenario for the multiple convergent origins of peptide ligases in various folds and clarify the bacterial origin of eukaryotic peptide-tagging enzymes of the TTL family.

Introduction

Conjugation of peptide or polypeptide tags to target proteins (peptide tagging) is a pervasive feature in both eukaryotes and bacteria. In eukaryotes Ub and Ub-like (Ubl) proteins, which are conjugated to target proteins, are the best known peptide tags.[1,2] Shorter tags in the form of single amino acids, such as leucine/phenylalanine and arginine are also added to the N-termini of proteins as a part of the N-end rule pathway.[3] A comparable single amino acid tag in eukaryotes is the addition of tyrosine to targets such as tubulin. Tubulin and several other proteins are also targets of longer peptide chains, primarily comprised of poly-glutamate or poly-glycine.[4,5] In bacteria the most conserved tag is an oligopeptide which is co-translationally added to the C-termini of proteins *via* the tmRNA system.[6] Several bacteria further possess equivalents of the N-end rule tagging *via* leucine/phenylalanine and arginine, and polyglutamate tags such as those added to the ribosomal protein S6.[3,7] More recently, we described homologs of the eukaryotic ubiquitin system in certain bacteria, which appear to be functionally linked to E1 and E2 homologs, suggesting the possibility of Ub-like conjugation in these organisms.[8] A similar system involving just the E1 protein is also used in bacterial and possibly archaeal cysteine synthesis pathways in which the newly synthesized cysteine is tagged as the terminal residue of an Ubl.[9–11] Furthermore, another small protein Pup, which is unrelated to Ub, also appears to be conjugated to target proteins in actinobacteria and certain other bacterial lineages (pupylation).[12,13] Eukaryotic ubiquitination, N-end rule modification, tmRNA-based tagging and pupylation appear to have a key role in destabilizing tagged proteins—*i.e.* targeting them for degradation *via* ATP-dependent unfolding and proteolytic systems (*e.g.* the proteasome or the ClpA/B system).[14,15] This convergent evolution of peptide-tagging on multiple occasions is probably due to the strong selective pressure exerted by the presence of

[a] *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, MD 20894, Bethesda, USA. E-mail: aravind@mail.nih.gov; Tel: +1-301-594-2445*
[b] *Omics Science Center (OSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama-shi, 230-0045 Kanagawa, Japan*
† This article is part of a *Molecular BioSystems* themed issue on Computational and Systems Biology.
‡ Electronic supplementary information (ESI) available: html file. See DOI: 10.1039/b917682a. Supplementary material can also be accessed at: ftp://ftp.ncbi.nih.gov/pub/aravind/peptide_ligases/supplementary_material_novel_peptide_synthesis.html

powerful generalized proteolytic systems in cells—the addition of a specific tag ensures that only a tagged protein and not just any protein in the cell is targeted for destruction by such systems. It is evident that in both bacteria and eukaryotes these tags have been further used as regulatory modifications that modify the properties of targeted proteins, especially in terms of their interactions with other biomolecules as a part of signaling pathways.[8,16]

These modifications usually occur *via* peptide or isopeptide linkages. Peptide linkages involve the N- and C- termini of target proteins (*e.g.* N-end, bacterial S6 glutamylation and tmRNA-based tagging and certain ubiquitination events), whereas isopeptide linkages occur *via* the epsilon amino group of lysine or the gamma carboxylate of glutamate (*e.g.* ubiquitination, polyglutamylation and pupylation). In a small number of instances the linkage appears to involve a thioester bond with a target cysteine.[17] Despite the diversity in these modifications there are some general themes that unify the enzymes catalyzing them. The tmRNA functions as both a tRNA and an mRNA in the formation of a conventional peptide bond *via* the ribosome.[6] Likewise the leucyl/phenylalanyl or argininyl transfer in the N-end rule utilizes a tRNA charged with either of these amino acids as a substrate to link the amino acid in a peptide linkage to the target NH_2 group.[3] However, this peptide bond formation is catalyzed by transferases that are related to the NH_2-group acetyltransferases (GNAT), but differ from them in using an aminoacylated tRNA in place of acetyl-coA.[18,19] In the case of Ub/Ubl conjugation the core pathway is comprised of two enzymes, E1 and E2. The first step in this process, catalyzed by E1, involves charging of the carboxyl group *via* adenylation in an ATP-dependent reaction.[9,10,20] This is followed by the transfer of the Ub/Ubl *via* successive thiocarboxylate linkages to the target NH_2 group to form a peptide or isopeptide linkage. In contrast to the E1-catalyzed reaction, pupylation, polyglutamylation, polyglycinylation and tyrosinylation involve a single enzyme that catalyzes the condensation of the COOH and NH_2 groups using the free energy of ATP.[4,12,21] These peptide bond formations proceed through the charging of the carboxylate *via* formation of an acyl phosphate, which is then attacked by the nitrogen of the amino group to form a carbonyl linkage. Two distinct folds of enzymes catalyze this reaction. In the case of pupylation, the ligase belongs to the carboxylate-amino group ligase (COOH–NH_2 ligases) or glutamine synthetase fold,[12] whereas polyglutamylation, polyglycination and tyrosinylation are catalyzed by members of the ATP-grasp fold.[4,5]

A number of studies have shown that the E1 enzymes of the Ub-conjugation pathway first arose in the context of ancient biosynthetic pathways of the cofactors thiamine and molybdopterin.[22–26] Subsequently, in course of prokaryotic diversification they appear to have been recruited to a range of biosynthetic pathways including those for cysteine and numerous small molecule secondary metabolites such as siderophores, modified peptides, and acylated amino acid derivatives.[8–10,27,28] A preliminary investigation of the COOH–NH_2 ligase fold shows that members of this fold are enzymes involved in the two ancient glutamine biosynthesis pathways—stand-alone (glutamine synthetase) and tRNA linked (GatABC).[29–31] Peptide ligases such as the glutamate-cysteine ligases, which catalyze the first step in the synthesis of the peptide cofactor glutathione, gamma-glutamylputrescine synthetase, which conjugates putrescine to the gamma carboxylate of glutamate,[32] and the Pup-ligase appear to be evolutionary derivatives of glutamine synthetase.[12] The ATP-grasp is also an ancient fold of which two distinct forms, namely the nucleic acid ligase and the classical version, had already emerged well before the last universal common ancestor (LUCA). The latter version had further diversified into several representatives by the time of LUCA, which were involved in basic metabolic pathways such as glycolysis, TCA, nitrogen assimilation and purine biosynthesis. Several distinct ATP-grasp enzymes catalyze peptide ligation reactions. In archaea, MptN and CofF catalyze the ligation of multiple glutamate residues to precursors of the coenzymes tetrahydrosarcinapterin (a folate-like pterin derivative) and F420 (a flavin-like molecule), respectively.[33,34] Bacterial ATP-grasp peptide ligases include the enzymes which catalyze the second step in glutathione synthesis (glutathione synthetase), further modifications of glutathione (glutathionyl spermidinesynthetase),[35] and synthesis of the storage polypeptide cyanophycin,[36] and peptide antibiotics, such as bacilysin[37] and butirosin.[38] Certain peptide bonds of the non-ribosomally synthesized peptides in peptidoglycan (*e.g.* the D-Ala dipeptide) and amide linkages in siderophores such as vibrioferrin are also formed by ATP-grasp enzymes.[39,40] Like E1 enzymes, ATP-grasp enzymes also catalyze modification of ribosomally synthesized peptides such as marinostatin and microviridin by forming cyclic lactone and amide linkages in them.[41,42] Interestingly, the GNAT fold peptide-ligases of the N-end rule system are related to the similar amino-acyl tRNA-utilizing enzymes involved in linking the L-amino acids in the side chain peptides of peptidoglycan from Gram-positive bacteria (the Fem/MurM family).[43]

These observations point to pervasive evolutionary links between various peptide/amide-bond-forming enzymes in ancient basic metabolic pathways, in cofactor and secondary metabolite biosynthesis, and peptide-tagging systems across a mechanistically and structurally diverse set of protein folds that catalyze such reactions (Fig. 1). Our preliminary investigations of these enzymes indicated that there is a wealth of poorly characterized prokaryotic non-ribosomal peptide/amide bond-forming systems, including possible pathways for novel metabolites and potential peptide-tagging systems.[10,12] Given the biological importance of biosynthetic systems for co-factors, antibiotics, other secondary metabolites, and peptide-tagging systems, we sought to systematically identify the ligase systems that might be involved in such processes. In earlier works we had identified such systems that utilize homologs of the E1- and E2-like enzymes of the Ub-system.[8,10] In this work we focus primarily on peptide-bond forming enzymes which form acyl-phosphate intermediates, namely ATP-grasp and COOH–NH_2 ligase fold enzymes, and to a certain extent the amino acyl tRNA-utilizing members of the GNAT fold. Using the abundance of prokaryotic genomic data we identified conserved gene neighborhoods or predicted operons and domain architectures of these proteins. We then used the contextual information derived from these associations to make functional inferences regarding the biosynthetic pathways

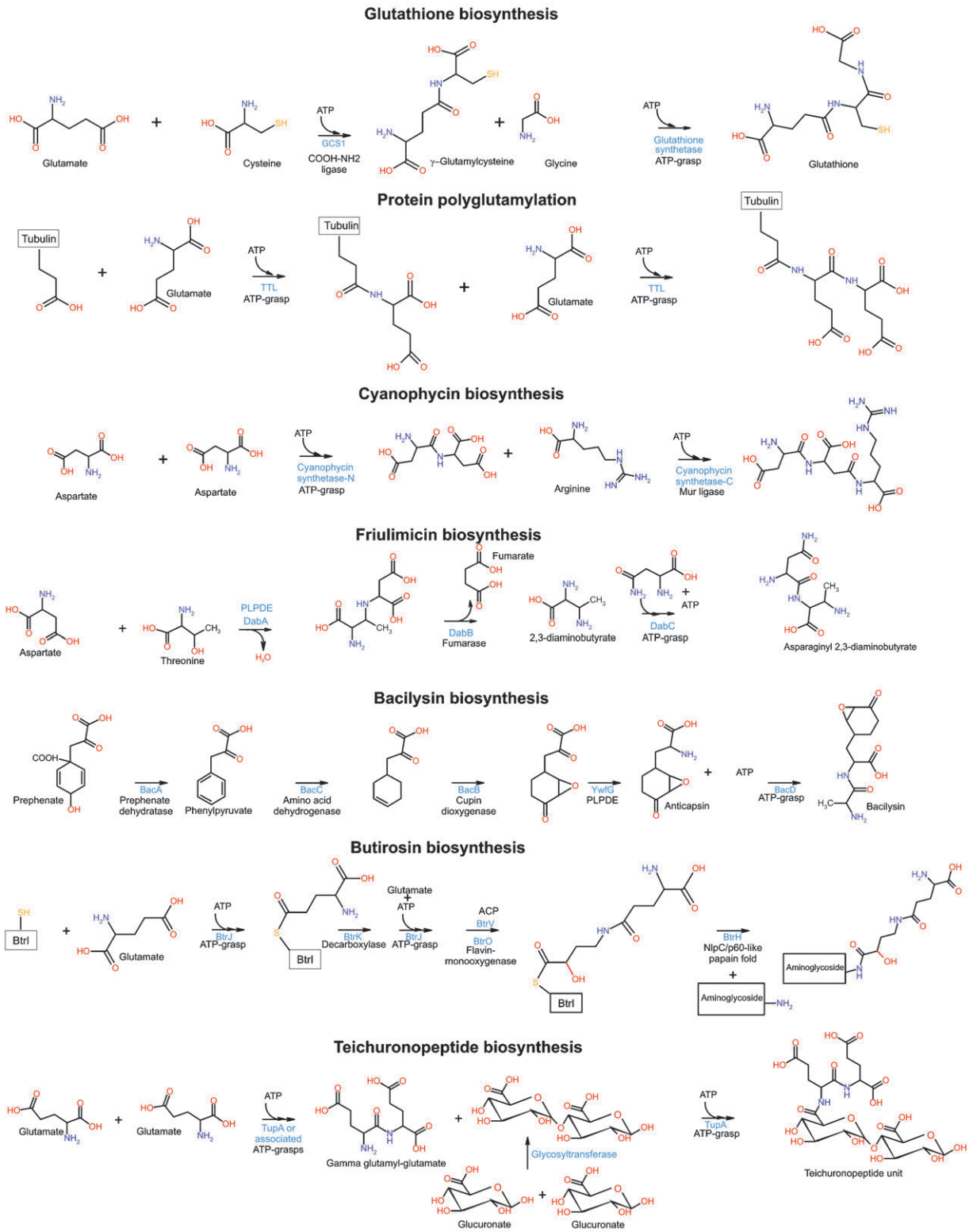


Fig. 1 Biosynthetic pathways with amidoligases. Examples of previously known and predicted biosynthetic pathways containing amidoligases are shown. Previously characterized reactions are glutathione biosynthesis, protein glutamylation by the TTL family and cyanophycin biosynthesis. Also shown are various predicted reactions elucidated in part or entirely in this study: 2,3-diaminobutryrate synthesis in the friulimicin biosynthesis pathway; anticapsin biosynthesis in the bacilysin biosynthesis pathway; the transpeptidase reaction catalyzed by the BtrH-like peptidase; the biosynthesis of teichuronopeptide.

in which they operate. As a consequence we were able to identify several novel peptide and related secondary metabolite biosynthetic pathways and also systems that might add peptide tags to proteins.

Results and discussion

Identification and classification of ATP-grasp, COOH–NH_2 ligase and amino-acyl tRNA dependent GNAT domains

In order to systematically identify novel members of the classical ATP-grasp and COOH–NH_2 ligases, we generated structure-based sequence alignments of known structures of these folds using the DALILITE program. We then used these sequence alignments as templates to prepare initial multiple alignments by adding previously characterized members of these folds (Fig. 1, ESI). These alignments were then used to initiate sequence profile searches with the PSI-BLAST program and the recently released improved hidden Markov search package HMMER3 (beta 2 version). We also set up independent sequence profile searches using the Fem peptide ligases and L/F transferases to identify novel aminoacyl tRNA-dependent GNAT fold peptide ligases. This family of peptide-ligases is characterized by an ancestral duplication and contains either two complete or partial versions of the GNAT domain[18,43] (see SCOP database; http://scop.mrc-lmb.cam.ac.uk/scop/). The newly identified domains recovered in all the above searches were confirmed by reciprocal searches and by checking the concordance of their predicted secondary structures to the structural alignment-derived template (See Material and Methods). In order to generate a natural classification of these domains we then clustered all recovered members using the BLASTCLUST program and further refined these clusters based on the information from shared conserved sequence features and domain architectures. For higher order classification we relied on shared structure and sequence synapomorphies (shared derived characters). Basic characterization and classification of enzymes of these folds have been previously reported by others and us (see SCOP database; http://scop.mrc-lmb.cam.ac.uk/scop/).[12,30,33,44,45]

The primary evolutionary split in the ATP-grasp fold separated the nucleic acid ligases from the remainder of the fold. The two differ in the way the six-stranded core domain of the ATP-grasp, which supplies two acidic residues to the active site, is combined with the smaller RAGNYA domain which supplies two basic residues (usually lysines) to the active site (Fig. 2).[44] Following this, the pyruvate phosphate dikinase and the succinyl CoA synthetase lineages successively branched off (see ESI).[46] All remaining ATP-grasp domains are unified by the fusion of an N-terminal pre-ATP-grasp domain (see SCOP database; http://scop.mrc-lmb.cam.ac.uk/scop/). Those which catalyzed reactions comparable to peptide ligation in purine biosynthesis such as phosphoribosylamine–glycine ligase (PurD), phosphoribosylglycinamide formyltransferase (PurT) and phosphoribosylaminoimidazole carboxylase (PurK), form a distinct lineage within this group. The majority of the classical peptide ligases, except those involved in bacilysin, friulimicin and butirosin biosynthesis, form another large monophyletic clade within this group (Fig. 2; see ESI). Amongst the classical peptide ligases we identified a previously unrecognized, large monophyletic clade of peptide ligases comprised of eukaryotic glutathione synthetases, glutathionyl spermidine synthetase and several novel families characterized in this study (see below), which are unified by a circularly permuted version of the ATP-grasp domain (Fig. 2; see ESI).

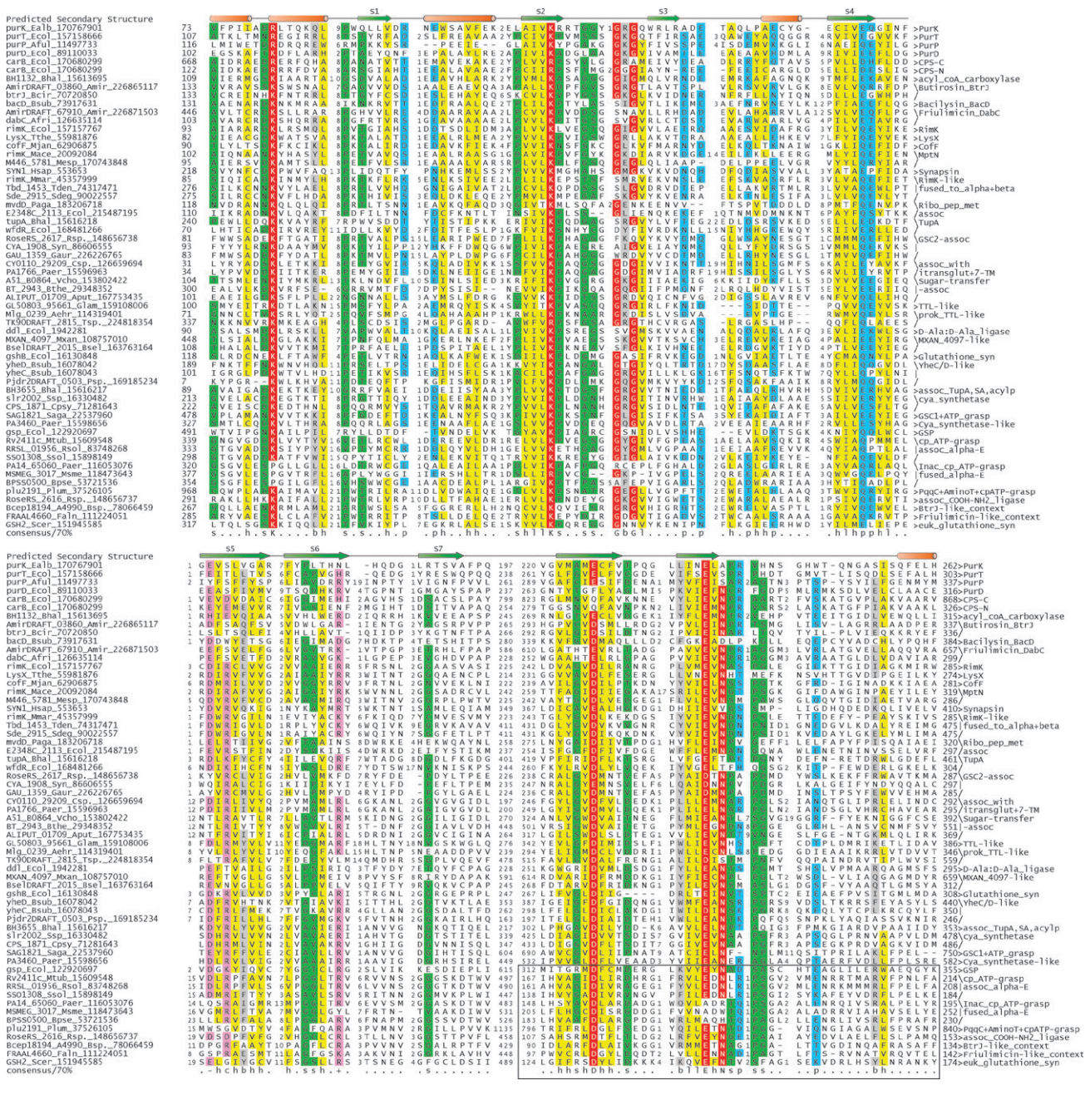
The evolution of the COOH–NH_2 ligase fold follows a relatively simple pattern with an early pre-LUCA split separating the classical glutamine synthetases from the GatB-type enzymes that synthesize glutamine in association with glutamate charged tRNAs. The classical glutamine synthetase appears to have been the precursor of an extensive radiation in bacteria that resulted in several lineages such as the GCS1, GCS2 and the Pup-ligase (PafA) families, two novel families of COOH–NH_2 ligases characterized here, and a few smaller distinct groups (ESI). Further, within the glutamine synthetase and the GCS2 families there were multiple duplications in bacteria spawning paralogous subfamilies. We observed that the N-terminal tRNA-binding domain of the divergent seryl tRNA synthetases from methanogenic archaea is an inactive version of this fold that probably evolved through rapid divergence from a classical glutamine synthetase-like precursor. The arginine and creatine kinases form a clade with the GatB-type enzymes unified by the presence of additional strands inserted into the core fold (ESI). The GNAT fold peptide ligases appear to have radiated in bacteria giving rise to at least five distinct families, the F/L transferase, R-transferase, Fem/MurM, and two new families identified here and also a few other minor lineages (Table 1). In this study, after the initial identification and classification steps, we specifically concentrated on defining members of these folds that are predicted to catalyze peptide bond or related amide linkage formations (see Table 1 and ESI for details). For example, in the case of the classical ATP-grasp fold we did not consider lineages such as the pyruvate phosphate dikinase in detail as they are not known to contain representatives performing peptide condensation reactions. Similarly, in the case of the COOH–NH_2 ligase fold we did not investigate in detail the "kinase-only" versions such as the arginine kinase and creatinine kinase (see SCOP database, http://scop.mrc-lmb.cam.ac.uk/scop/).

In addition to previously identified and experimentally characterized families our searches recovered several novel members of these folds which to the best of our knowledge have either not been previously identified or have been poorly characterized (Table 1). Some examples of such ATP-grasp domains include a large family typified by *Mycobacterium tuberculosis* Rv2411c (gi: 15609548), an inactive family prototyped by Rv2567 (gi: 15609704) and another family typified by *Pseudomonas aeruginosa* PA3460 (gi: 15598656), which is related to the cyanophycin synthetase family. Examples of novel COOH–NH_2 ligases recovered in this study include a peculiar circularly permuted version (*e.g.* Rv2566, gi: 15609703), two versions found in phages such as phiEco32 (gi: 167583639 and 167583641) and a version fused to the eukaryotic chromatin-associated YEATS domain from the chromist alga *Thalassiosira pseudonana* (gi: 223997528). In the case of the Fem/MurM-like GNAT protein an example of a previously uncharacterized distinctive version recovered in the above searches is a version fused to the 2nd lysyl-tRNA synthetase paralog of actinomycetes (*e.g.* 15608778 from *M. tuberculosis*) and a version associated with a cyanophycin-synthetase-like ATP-grasp domain (Table 1, see below).

Domain architecture and gene-neighborhood syntaxes observed in known peptide or amide bond forming enzymes

The evolutionary classification of these enzymes by itself might not be sufficient to predict their functions. For example, in the classical glutamine synthetase family of the COOH–NH_2 ligase fold, in addition to enzymes catalyzing glutamine synthesis, there are paralogous subfamilies such as PuuA and FluG which do not seem to function in amino acid biosynthesis. Instead PuuA catalyzes the condensation reaction of L-glutamate and putrescine.[32] Likewise, in the ATP-grasp fold, the LysX and the RimK-like proteins are closely related within a large monophyletic assemblage of amide-bond forming enzymes (Fig. 2, ESI). However, the former catalyzes the formation of an amide linkage *via* the condensation of alpha-amino adipate with a fatty acid (in lysine synthesis); whereas the latter catalyzes the formation of oligo-glutamate peptide tags on proteins or pterin and F420[33,34] or the synthesis of cyclic peptides *via* amide and lactone linkages.[41,42] Similarly, the ATP-grasp protein BtrJ involved in butirosin biosynthesis catalyzes both the formation of a thioester of glutamate gamma-carboxylate with a cysteine from an acyl carrier protein as well as a peptide bond in the diglutamate moiety found in the precursor of this antibiotic.[38] Thus, it became clear that additional information, such as contextual links, would be required to clarify the exact catalytic role or pathway in which these enzymes might participate.

Enzymes catalyzing successive steps in a biochemical pathway and physically interacting partners related to a



particular reaction (*e.g.* Ubls and E1-like enzymes) tend to co-occur in conserved gene neighborhoods in prokaryotes or in certain cases fuse to form multidomain proteins.[8] These associations provide contextual information that is extremely useful in predicting functional specificities of proteins.[47–49] Hence, we first tried to identify common syntactical features of domain architectures and conserved gene neighborhoods amongst previously characterized representatives of these peptide/amide-bond-forming enzymes. Then we used these syntaxes to predict potential functional pathways in which the poorly characterized forms might participate. We observed two common themes in terms of domain architectures (Fig. 3): (1) Fusion of the peptide/amide-forming catalytic domain to a peptidase domain. This is exemplified by the glutathionyl spermidine synthetase, where the ATP-grasp domain which condenses spermidine to glutathione is linked to a C-terminal domain of the NlpC/p60 superfamily (papain-like fold).[50] (2) Fusion of two distinct amide/peptide-bond-forming domains. An example of this is the fusion of glutamate–cysteine ligase-1 (GCS1) of the COOH–NH_2 ligase fold to an ATP-grasp domain that catalyzes the subsequent ligation of glycine in glutathione synthesis.[51] Similarly, in the cyanophycin synthetase an N-terminal ATP-grasp domain catalyzes the formation of the poly-aspartate and a C-terminal Mur family peptide ligase (P-loop kinase fold)[52] catalyzes condensation of the amino group of arginine to the beta carboxylate of the aspartates.[36,53] A duplication of two ATP-grasp domains is also observed in the carbamoyl phosphate synthetase large subunit (CPS) that also catalyzes two ATP-dependent reactions, one of which is similar to amide/peptide bond formation[54] (Fig. 3).

These associations were reinforced and further refined by links furnished by conserved gene-neighborhoods (Fig. 4). We observed that operons encoding well-characterized peptide/amide-bond-forming enzymes showed frequent linkages to genes for diverse peptidases. For example, the cyanophycin synthetase gene was found to be frequently linked to cyanophycinase,[53] a peptidase of the flavodoxin fold. Occasionally a second potential peptidase/amidase of the TIM Barrel amidohydrolase fold[55] was also found linked to the cyanophycin synthetase gene. Similarly, in the case of the D-Ala–D-Ala ligase, gene neighborhood linkages are found to the VanY peptidase (Hedgehog C-terminal-like fold) and a peptidase of the α/β hydrolase fold in mutually exclusive contexts (Fig. 4).[56] We observed that the RimK subfamily members were most frequently linked to genes of the pepsin-like peptidase fold (Fig. 4). The related subfamilies involved in cofactor glutamylation, MptN and CofF, respectively show linkages to metallopeptidases of the phosphorylase fold (also known as the peptidyl-tRNA hydrolase fold or M20-like peptidases; see SCOP database; http://scop.mrc-lmb.cam.ac.uk/scop/) and "M50"-like metallopeptidases. The GCS1 peptide ligase of the glutathione pathway shows conserved operonic linkages to the insulinase(LuxS)-like metallopeptidases or the gamma-glutamyltranspeptidase (Fig. 4). In several proteobacteria and bacteroidetes we discovered a novel linkage between the glutathione synthetase and two peptidase domains that are often fused in one polypeptide or are neighbors—one of the phosphorylase fold ("M20-like") and another which is a modified version of the zincin-like metallopeptidase fold (Fig. 4, ESI). The Pup-ligase likewise shows a strong operonic linkage to peptidases in the form of the proteasomal subunits of the NTN-hydrolase fold.[12] Even in the case of amide bond formations catalyzed by the COOH–NH_2 ligase domain in tRNA-linked glutamine synthesis and the ATP-grasp domain in carbamoyl phosphate biosynthesis we observed operonic or physical associations to distinct amidases—respectively, those of the acyl-amidohydrolase (GatA subunit) and flavodoxin folds (CPS small subunit) (ESI). As with domain architectures, two distinct peptide/amide-bond forming enzymes also tended

Fig. 2 Multiple sequence alignment of the ATP-grasp domain. Proteins are labeled with the gene names, species abbreviations and Genbank index (gi) numbers separated by underscores. Sequences are arranged by family types, which is shown to the right of the alignment. In a subset of sequences (labeled as GCS2_assoc), an insert in the [ED]hN motif at the C-terminus is shown in smaller font. The consensus secondary structure is shown above the alignment and was derived from a combination of crystal structures and alignment-based structure prediction. Strands are depicted as green arrows and helices as cylinders. The circularly permuted portion unifying the circularly permuted clade is enclosed within a box and aligned to the C-terminus of the unpermuted forms. Columns in the alignment are colored based on their conservation at 70% consensus. The coloring scheme and consensus abbreviations (shown at the bottom of the alignment) is c: charged (DEHKR), −: acidic (DE) and +: basic (HKR) residues in magenta; b big residues (KMILEWRYFQ) in grey, p: polar residues (CDEHKNQRST) in blue; h hydrophobic (ACFILMVWY) and l: aliphatic (LIV) residues in yellow, and s: small (ACDGNPSTV) and u: tiny (GAS) residues in green. Absolutely conserved residues are in red. Species abbreviations are as follows: Abau: *Acinetobacter baumannii*; Aehr: *Alkalilimnicola ehrlichii*; Afri: *Actinoplanes friuliensis*; Aful: *Archaeoglobus fulgidus*; Amir: *Actinosynnema mirum*; Aput: *Alistipes putredinis*; Asp.: *Acidovorax* sp.; Atum: *Agrobacterium tumefaciens*; Bcen: *Burkholderia cenocepacia*; Bcir: *Bacillus circulans*; Bhal: *Bacillus halodurans*; Bpet: *Bordetella petrii*; Bpse: *Burkholderia pseudomallei*; Bsel: *Bacillus selenitireducens*; Bsp.: *Burkholderia* sp.; Bsub: *Bacillus subtilis*; Bthe: *Bacteroides thetaiotaomicron*; Cagg: *Chloroflexus aggregans*; Cpha: *Chlorobium phaeobacteroides*; Cpsy: *Colwellia psychrerythraea*; Csp.: *Cyanothece* sp.; Daut: *Desulfobacterium autotrophicum*; Drad: *Deinococcus radiodurans*; Dthi: *Desulfonatronospira thiodismutans*; Ealb: *Escherichia albertii*; Ecol: *Escherichia coli*; Faln: *Frankia alni*; Gaur: *Gemmatimonas aurantiaca*; Gdia: *Gluconacetobacter diazotrophicus*; Glam: *Giardia lamblia*; Hsap: *Homo sapiens*; Lbif: *Leptospira biflexa*; Lsp.: *Leptospirillum* sp.; Mace: *Methanosarcina acetivorans*; Mesp: *Methylobacterium* sp.; Mjan: *Methanocaldococcus jannaschii*; Mmag: *Magnetospirillum magnetotacticum*; Mmar: *Methanococcus maripaludis*; Msed: *Metallosphaera sedula*; Msme: *Mycobacterium smegmatis*; Mtub: *Mycobacterium tuberculosis*; Mxan: *Myxococcus xanthus*; Paer: *Pseudomonas aeruginosa*; Plum: *Photorhabdus luminescens*; Plut: *Pelodictyon luteolum*; Pnec: *Polynucleobacter necessarius*; Psp.: *Paenibacillus* sp.; Rbal: *Rhodopirellula baltica*; Rcen: *Rhodospirillum centenum*; Rsol: *Ralstonia solanacearum*; Rsp.: *Roseiflexus* sp.; Saga: *Streptococcus agalactiae*; Scer: *Saccharomyces cerevisiae*; Sdeg: *Saccharophagus degradans*; Ssol: *Sulfolobus solfataricus*; Ssp: *Synechocystis* sp.; Susi: *Solibacter usitatus*; Syn: *Synechococcus* sp.; Tden: *Thiobacillus denitrificans*; Telo: *Thermosynechococcus elongatus*; Tsp.: *Thioalkalivibrio* sp.; Tthe: *Thermus thermophilus*; Vbac: *Verrucomicrobiae bacterium*; Vcho: *Vibrio cholera*; Vspi: *Verrucomicrobium spinosum*; Xaut: *Xanthobacter autotrophicus*.

Table 1 Amidoligase systems and their functional partners

Family/sub-family ATP-grasp amidoligases	Pathway	Functional partners, operonic associations and comments
Glutathione synthetase	Glutathione biosynthesis	GCS1, M20 family peptidase, zincin; The glutathione synthetase is occasionally fused to GCS1; In some instances, the M20 peptidase and zincin are fused in a single polypeptide
ATP-grasp related to cyanophycin synthetase fused to GCS1	Glutathione biosynthesis mainly in Gram-positive bacteria	This ATP grasp is related to cyanophycin synthetase ATP-grasp domain rather than the classical glutathione synthetase; fused to an N-terminal GCS1 type COOH–NH_2 ligase domain
Glutathionyl-spermidine synthetase-like circularly permuted ATP-grasp	Glutathionyl-spermidine biosynthesis	Conserved aspartate containing protein, membrane protein with four conserved cysteines. In proteobacteria, firmicutes and kinetoplastids the ATP-grasp domain is fused to a NlpC/p60-like papain fold thiol peptidase
LysX	Lysine biosynthesis	LysW, ArgB, acetylornithine deacetylase and ArgD-like aminotransferase
CofF	Cofactor F420 glutamylation	M50-family peptidase, Mur family ligase
MptN	Tetrahydromethanopterin glutamylation	Tetrahydromethanopterin biosynthesis genes in bacteria such as methylene tetrahydromethanopterin cyclohydrolase, formaldehyde activating enzyme, tetrahydromethanopterin formyltransferase. Operons also contain a second ATP-grasp
RimK	peptide tag, cofactor glutamylation[a]	Pepsin-like peptidase, M20-like peptidase, succinyl glutamate desuccinylase (phosphohydrolase fold), 5-formyltetrahydrofolate cycloligase
RimK-like ATP-grasp fused to RLAN (RimK-like ATPase N-terminal) domain; *T. denitrificans* Tbd_1454, (gi: 74317472)	Predicted peptide/peptide tag biosynthesis[a]	GNAT fused to a papain-like peptidase, M20-like peptidase, GCS2, 4Fe–4S Ferredoxin, metal–sulfur cluster protein, and ribosomal proteins; in one instance the GNAT is fused to the ATP-grasp protein
7-TM associated RimK-like ATP-grasp; *P.aeruginosa* PA1766 (gi: 15596963)	Predicted peptide/peptide tag biosynthesis[a]	7-TM protein with an extracellular N-terminal inactive transglutaminase domain, pepsin-like peptidase; the ATP-grasp domain is predicted to modify the 7TM protein or a cofactor that interacts with it
Mycobacterium Rv2411c-like (gi: 15609548) circularly permuted ATP-grasp	Predicted peptide/peptide-tag biosynthesis[a]	alpha-E, transglutaminase, Anbu-like peptidase (NTN-hydrolase), inactive circularly permuted ATP-grasp fused N-terminal to alpha-E, transglutaminase fused N-terminal to circularly permuted glutamine synthetase, GAT-I amidotransferase (flavodoxin fold)
Roseiflexus RoseRS_2616-like (gi: 148656737) circularly permuted ATP-grasp	Predicted peptide biosynthesis[a]	*Roseiflexus* RoseRS_2615-like GCS2, alpha/beta hydrolase fold peptidase, *Phytopthora*-type transglutaminase, GAT-II (amidohydrolase), M20 peptidase (phosphorylase fold), Fem/MurM ligase, non-permuted ATP-grasp
Polyglutamylase/TTL/polyglycinase	Add peptide-tags to target proteins	Modification of several eukaryotic proteins; some predicted in modifying chromatin proteins along with SET domain protein methylases
Bacterial TTL-like (*Thioalkovibrio*TK90DRAFT_2815 (gi: 224818354)	Predicted amino acid/peptide tags in bacteria[a]	Versions of the domain in bacteria are fused to a 2-oxoglutarate Fe(II) dependent dioxygenase domain
D-Ala:D-Ala ligase	Peptidoglycan biosynthesis	VanY (Hedgehog C-terminal-like fold), alpha/beta hydrolase, alanine racemase, FtsA, FtsQ, FtsZ, Mur family ligases, MurG
Myxococcus MXAN_4097-like (gi: 108757010)	Predicted peptide biosynthesis in cell wall metabolism[a]	D-Aminopeptidase, gamma-glutamyltranspeptidase (NTN-hydrolase fold), M20-like peptidase; paralogous ATP-grasp domains are sometimes fused in a single polypeptide; the ATP-grasp is also occasionally fused to M20 peptidase and D-aminopeptidase
Bacillus YheC/D-like (gi: 16078043, 16078042)	Modification of spore coat components[a]	YheA, SspB, a membrane protein, multiple ATP-grasp paralogs in same operon; sporadic fusion to Mur ligase and PP2A domains involved in poly-gamma glutamate synthesis
Phage phiEco32-ATP-grasp gi: 167583635)	Predicted cell wall modification[a]	Glutamine: D-hexose-6-phosphate amidotransferase (NTN-hydrolase), two distinct COOH–NH_2 ligases
TupA involved in teichuronopeptide biosynthesis *Bacillus halodurans* (gi: 15616218)	Formation of poly-gamma glutamate poly-glucuronate copolymers[a]	Fused or in operonic associations with two distinct ATP-grasp domains, acylphosphatase, PP2A-like phosphatases, alpha/alpha toroid, and genes involved in teichoic acid biosynthesis and sugar uptake
TupA-like ATP-grasp in cell surface polymer and capsule synthesis	Synthesis of cell surface sugar–L/D-amino acid polymers[a]	Family 1 and Family 2 glycosyltransferases, other ATP-grasps, and genes involved in the biosynthesis of cell surface polysaccharides such as O-antigen, teichoic acid and capsule; up to six ATP-grasp genes can be combined in an operon
RimK-like ATP-grasp involved in capsule synthesis	Synthesis of cell surface sugar–L/D-amino acid polymers[a]	Family 1 glycosyltransferase, sugar-lipid carrier phosphotransferase, TupA-like ATP-grasp and paralogous ATP-grasps. 1–4 ATP-grasp genes in an operon
Cyanophycin synthetase	Storage polypeptide biosynthesis	Cyanophycinase (flavodoxin fold), TIM Barrel amidohydrolase. Several operons have two paralogous ATP-grasp genes. Most versions consist of an N-terminal ATP-grasp fused to a C-terminal Mur family ligase

Table 1 (*continued*)

Family/sub-family	Pathway	Functional partners, operonic associations and comments
ATP-grasp amidoligases		
New Cyanophycin synthetase-like fused to an N-terminal GNAT domain; *P.aeruginosa* PA3460 (gi: 15598656)	Predicted storage polypeptide biosynthesis[a]	GNAT, distinct M20-type peptidase ("M42"-like), asparagine synthetase; asparagine is predicted to be one of the amino acids in the polypeptide
DabC-like; *A. friuliensis* (gi: 126635114)	Friulimicin biosynthesis[a]	DabA-like PLPDE, DabB-like fumarase/argininosuccinate lyase, threonine aldolase, homoserine (GHMP) kinase,
Distinct Friulimicin-biosynthesis-like (circularly permuted ATP-grasp); *Frankia* FRAAL4660 (gi: 111224051)	Predicted peptide antibiotic biosynthesis[a]	PqqC-oxidoreductase, PLPDE aminotransferase, M24-like metallopeptidase, E1, 2-oxoglutarate dependent dioxygenase, GNAT. plu2191 combines the PqqC oxidoreductase, PLPDE aminotransferase and the ATP-grasp in a single protein.
BtrJ; *B. circulans* (gi: 70720850)	Butirosin/predicted peptide antibiotic biosynthesis[a]	gamma-Glutamyl cyclotransferase (BtrG), non-ribosomal peptide synthetases, ACP, BtrH-like peptidase, PLPDE fold decarboxylase
Distinct Butirosin-biosynthesis-like with circularly permuted ATP-grasp; *Burkholderia* Bcep18194 (gi: 78066459)	Predicted secondary metabolite biosynthesis[a]	Acyl ACP synthase, glycine C-acetyltransferase-like PLPDE, two circularly permuted ATP-grasp genes
Bacilysin synthetase	Peptide antibiotic biosynthesis	Prephenate dehydratase (BacA), cupin superfamily dioxygenase (BacB), amino acid dehydrogenase (BacC), PLPDE aminotransferase (ywfG), peptide transporter
PvsA-like ATP-grasp, *Vibrio* (gi: 194541429)	Siderophore biosynthesis	PvsB/D type ligases, siderophore transporters and receptors
Marinostatin/Microviridin cyclizing enzymes	Ribosomally synthesized peptide modification	Two paralogous ATP-grasps, ABC transporter with peptidase domain, GNAT, marinostatin-like peptides
Other peptide cyclizing/modifying systems	Ribosomally synthesized[a] Peptide modification	Depending on the operons a single ATP-grasp ligase is associated with O-methyltransferases, SAM radical enzymes, multiTM proteins, McbC-like oxidoreductases and diverse peptides
$COOH–NH_2$ ligase domain associated systems		
PuuA	Putrescine utilization	PuuD (GAT-I family amidohydrolase)
FluG	Predicted condensation of a glutamate and an amino group-bearing molecule[a]	TIM Barrel fold amidohydrolase, APG superfamily permease, GAT-I family amidohydrolase; in eukaryotes often fused to an N-terminal TIM-Barrel fold amidohydrolase
GCS1 (glutamate-cysteine ligase)	Glutathione biosynthesis	Glutathione synthetase, insulinase (LuxS-like)-like peptidase, gamma-glutamyltranspeptidase (NTN), M20-like peptidase, zincin. Versions of this domain in firmicutes are fused to a cyanophycin synthetase-like ATP-grasp (see above)
Classical GCS2	Predicted peptide biosynthesis[a]	RNA methyltransferase, occasionally glutamine synthetase-like ATP-grasp
Roseiflexus RoseRS_2615-like GCS2 (gi: 148656736)	Predicted peptide biosynthesis[a]	*Roseiflexus* RoseRS_2616-like ATP-grasp (see above)
RLAN domain associated GCS2	Predicted peptide biosynthesis[a]	*P. aeruginosa* PA1944 type ATP-grasp (see above)
Pup-ligase (PafA); GCS2 family	Pupylation (peptide tagging)	PUP, proteasomal peptidase (NTN-hydrolase fold), proteasomal AAA + ATPase, proteasomal chaperone PAC2
GCS2 associated with inactive formyl-glycine synthesizing enzyme; *Mycobacterium* Rv3704c (gi: 15610840)	Predicted amine utilization pathway[a]	Inactive formyl-glycine synthesizing enzyme with an N-terminal DinB protease and C-terminal C-type lectin domain, Rossmann fold methylase, GAT-II (NTN-hydrolase fold)
Pseudomonas Patl_3664-like	Predicted amine utilization pathway[a]	Inactive formyl-glycine synthesizing enzyme with an N-terminal DinB protease and C-terminal C-type lectin domain that is further fused to a Rossmann methylase domain
Desulfitobacterium DSY4546-like (gi: 89897292)	Predicted amine utilization pathway[a]	gamma-Glutamyl cyclotransferase (Aig-2/BtrG), GAT-I peptidase (flavodoxin fold). Versions in *Clostridium* are fused to an N-terminal SWIM domain
Circularly permuted $COOH–NH_2$ ligase	Peptide tag biosynthesis[a]	Inactive Rv2411c-like ATP-grasp, alpha-E (see above). Usually fused to an N-terminal transglutaminase domain
Phage phiEco32-$COOH–NH_2$ ligase-1 phi32_84 (gi: 167583639)	Predicted cell wall modification[a]	See phage phiEco32-encoded ATP-grasp for more details; laterally transferred to eukaryotes, fused to YEATS domain in stramenopiles
Phage phiEco32-$COOH–NH_2$ ligase-2	Predicted cell wall modification and spore coat biosynthesis[a]	YheC/D-like ATP-grasp, CotE in Gram-positive bacteria.
Vibrio MARTX toxin	Cross-linking of host actin	Found in multi domain proteins fused to downstream papain-like peptidase and alpha/beta hydrolase domains
Gcn5-like amino acetyltransferase (GNAT) associated systems		
Phenylalanyl/leucyl transferase	Peptide-tag biosynthesis N-end rule pathway	Leucyl/phenylalanyl tRNA
Argininyl transferases	Peptide-tag biosynthesis N-end rule pathway	Argininyl tRNA. Sporadically it might be fused to the above enzyme.
Fem/MurM ligase	Peptidoglycan biosynthesis	tRNA, FemA, FemB, Fem/MurM ligase

Table 1 (*continued*)

Family/sub-family ATP-grasp amidoligases	Pathway	Functional partners, operonic associations and comments
Mycobacterial lysyl tRNA synthetase associated	Predicted cell surface peptide biosynthesis[a]	Fused or operonically associated with a multi-TM membrane protein and Lysyl tRNA synthetase.
Fem/Mur M ligase associated with ATP-grasp; *Synechococcus* CYA_1909	Predicted peptide biosynthesis[a]	Found only cyanobacteria in place of GCS2-type ligase that is typically seen in these operons (see above, *Roseiflexus* RoseRS_2615-like)
GNAT fused to a papain-superfamily cysteine protease	Predicted peptide biosynthesis[a]	See *P. aeruginosa* PA1944-like ATP-grasp for functional partners

[a] New information presented in this study regarding this amidoligase system.

Fig. 3 Domain architectures of Amidoligases. Proteins are denoted by their gis (GenBank ID) and species names. Domain architectures are grouped according to various contextual themes discussed in the text. Domains are shown as cartoon representations and are not drawn to scale. Peptidases are shown as colored hexagons. Standard domain abbreviations are used wherever possible. For other non-standard abbreviations refer to ESI. X represents a poorly characterized globular domain.

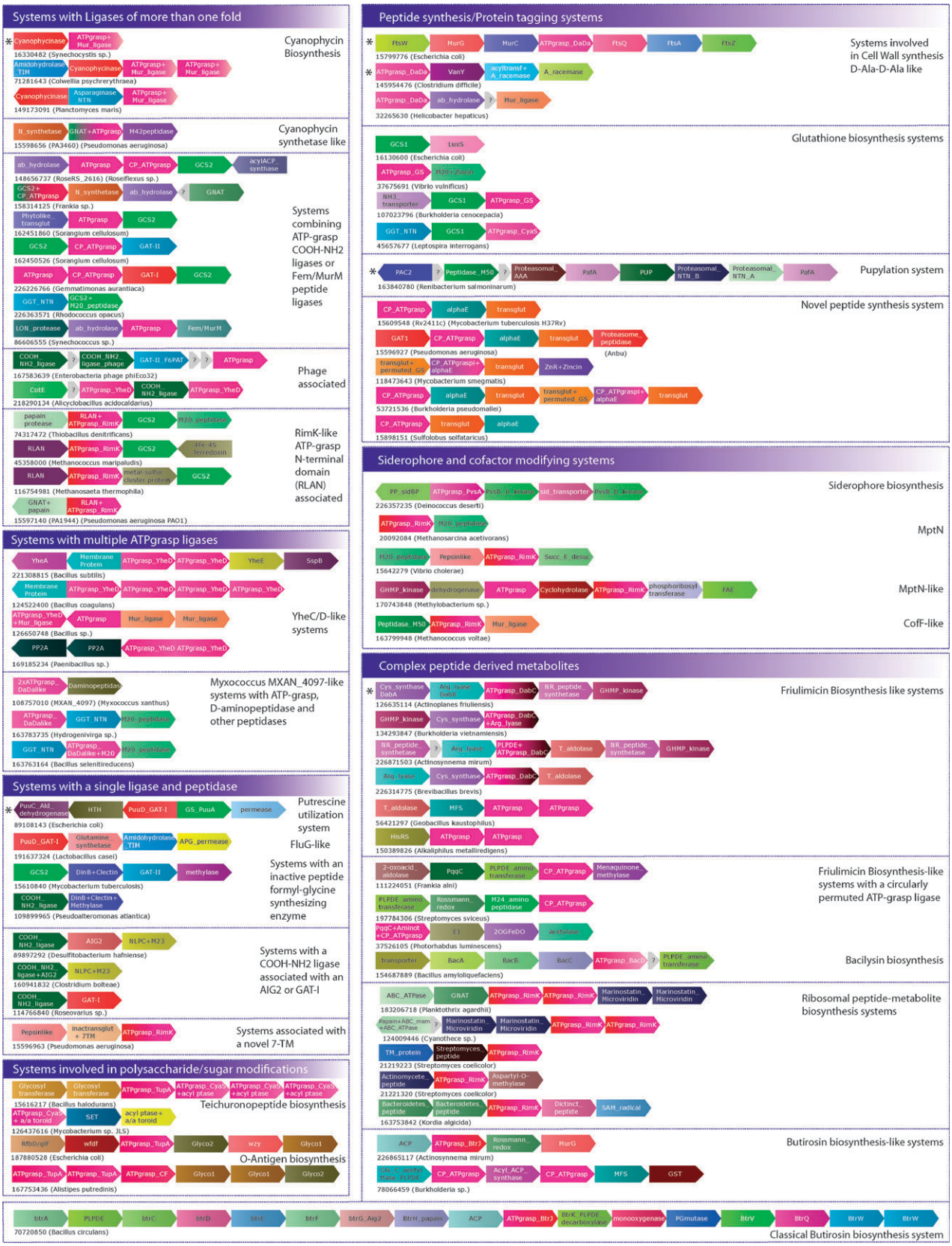


Fig. 4 Gene neighborhoods of various amidoligase systems described in this study. Genes are represented as arrows that point from the 5′ to the 3′ end of the coding sequence. Operons are labeled with the gi and species name of a representative gene that typically belongs to one of the major amidoligase folds described in the study. Gene names of prototypic examples are further depicted in brackets. Genes belonging to the same protein fold share the same color. Gene neighborhoods are arranged according to the functional themes described in the text. Well known neighborhoods are marked with an asterisk next to them.

to co-occur in the same operon (Fig. 4). For example, cyanophycin synthetases are frequently found in a conserved gene neighborhood combining two distinct paralogous versions of this enzyme. In the case of the D-Ala–D-Ala ligase operons we often find them co-occurring with genes for Mur family amino-acid ligases (Fig. 4).

Thus, the common denominator of the combined evidence from domain architectures and predicted operons was the association between genes encoding distinct peptide/amide-bond forming enzymes and/or associations with genes encoding a diverse array of structurally unrelated peptidases/amidases (represented as a network in Fig. 5). These associations could have emerged due to several distinct selective forces: (1) Coupled enzymes catalyzing opposite reactions potentially form a switch that maintains a certain concentration of the peptide/amide product depending on concentrations of precursors and cellular requirements. This appears to be case in the cyanophycin synthetase-cyanophycinase and glutathionyl spermidine synthetase-amidase pairs.[36,57] (2) The amidase could release ammonia for formation of a new amide linkage (*e.g.* in CPS and GatA).[54,58] (3) The Pup-proteasome system differs from all these contexts in being the only one having not just a peptidase but also an ATP-dependent protein unfolding apparatus (the proteasomal ATPase). This is consistent with its distinct proteolytic function that not just hydrolyzes a peptide linkage but also unfolds and degrades the Pup-tagged proteins.[12,13] The combination of two distinct peptide/amide ligases is indicative of the formation of two or more successive, distinct versions of such bonds involving more than one carboxylate or amino group-bearing moiety (*e.g.* the successive linkages in glutathione).[33] This is specifically supported by the observation that the linked peptide/amide ligases often belong to unrelated folds or are usually distantly related if of the same fold (Fig. 4 and 5).

The contextual linkages of enzymes involved in synthesis of stand-alone peptides or peptide tags on proteins are typically distinct from those involved in basic metabolism. The latter often show extensive linkages to other enzymes involved in these basic metabolic pathways and usually lack linkages to peptidases or second peptide ligases. For example, those catalyzing amide-bond formation in purine metabolism (*e.g.* PurD, PurK and PurT) are linked to other purine metabolism genes and those involved in amino acid biosynthesis are linked to other genes in this pathway (*e.g.* LysX subfamily is linked to lysine biosynthesis genes) (ESI). In other cases the extended gene-neighborhood context might be reflective of the other functional links of peptide ligases and associated peptidases. For instance, the peptide ligases involved in bacterial cell wall biogenesis (*e.g.*D-Ala–D-Ala ligase) are linked to several other cell-wall related genes such as FtsQ or amino acid racemases (Fig. 4) that generate D-amino acid substrates. Thus, contextual associations provide a means of distinguishing ligases that form distinct peptides or related amides, and also anchoring them to the larger pathway in which they might participate (Fig. 4 and 5).

These contextual associations are comparable to those of the adenylating enzymes of E1-like fold, which provided a powerful means to decipher their functional contexts with considerable precision.[10] The E1s involved in molybdopterin and thiamin biosynthesis are never linked to peptidases. Those involved in cysteine and siderophore biosynthesis or peptide modifications are linked to peptidases, while those involved potential Ub transfer systems are linked to both E2 homologs and peptidases.[10]

Contextual associations predict novel peptide synthesis, modification and tagging systems

We then explored the domain architectures and operonic links of the poorly characterized ATP-grasp, COOH–NH_2 ligase and Fem/MurM domains to identify matches to the functionally informative contextual templates described above (Fig. 5). As a consequence we were able to identify about 23 potential biosynthetic systems for distinct peptides and related amide containing metabolites, which showed a great variety of phyletic patterns in bacteria and certain bacteriophages (Fig. 4, Table 1). We describe below some of the major examples of these systems along with the predicted activities or functions.

A novel peptide synthesis and potential tagging system widely distributed in prokaryotes

One of the most widespread systems recovered in this study was defined by a conserved gene neighborhood distributed across most lineages of proteobacteria, actinomycetes, cyanobacteria, chlamydiae/verrucomicrobia and chloroflexi (Fig. 4). It is also found more sporadically in crenarchaea, spirochetes, firmicutes and deinococci. The core of this system is characterized by a predicted operon encoding: (1) An ATP-grasp protein (prototyped by the *M. tuberculosis* Rv2411c) belonging to the large clade of circularly permuted versions (Fig. 2 and 5; ESI). (2) A unique alpha-helical protein with two copies of an absolutely conserved motif with a glutamate–arginine (ER) dipeptide signature (Fig. 6). (3) A transglutaminase protein (papain fold) and a NTN-hydrolase peptidase closely related to the classical proteasomal peptidase. This proteasomal peptidase homolog had earlier been described as a novel bacterial proteasome termed 'Anbu'.[59] Studies on nitrogen starvation in *Pseudomonas putida* show that this peptidase and the linked transglutaminase are highly expressed upon nitrogen starvation.[60] (4) Several representatives of this operon also encode a distinctive zincin-like metallopeptidase (Fig. 4, ESI). (5) Even more sporadically the system might also contain a linked gene for an amidotransferase of the GAT-I family (flavodoxin fold amidase; see SCOP database; http://scop.mrc-lmb.cam.ac.uk/scop/)). Additionally, in several bacteria, the above core operon is linked to a second predicted operon which encodes: (1) a version of the unique alpha-helical protein found in the above operons, which in this case is fused at the N-terminus to an inactive circularly permuted ATP-grasp domain (Fig. 2–5). (2) A distinctive circularly permuted COOH–NH_2 ligase which is fused to an N-terminal transglutaminase domain.

The original interpretation of the Anbu peptidase as the constituent of a single subunit bacterial proteasome is unsupported.[59] Firstly, all know proteasomes and related systems (*e.g.* ClpAB, HslUV, FtsH, and different Lon-like systems) contain AAA+ ATPase subunits that are necessary

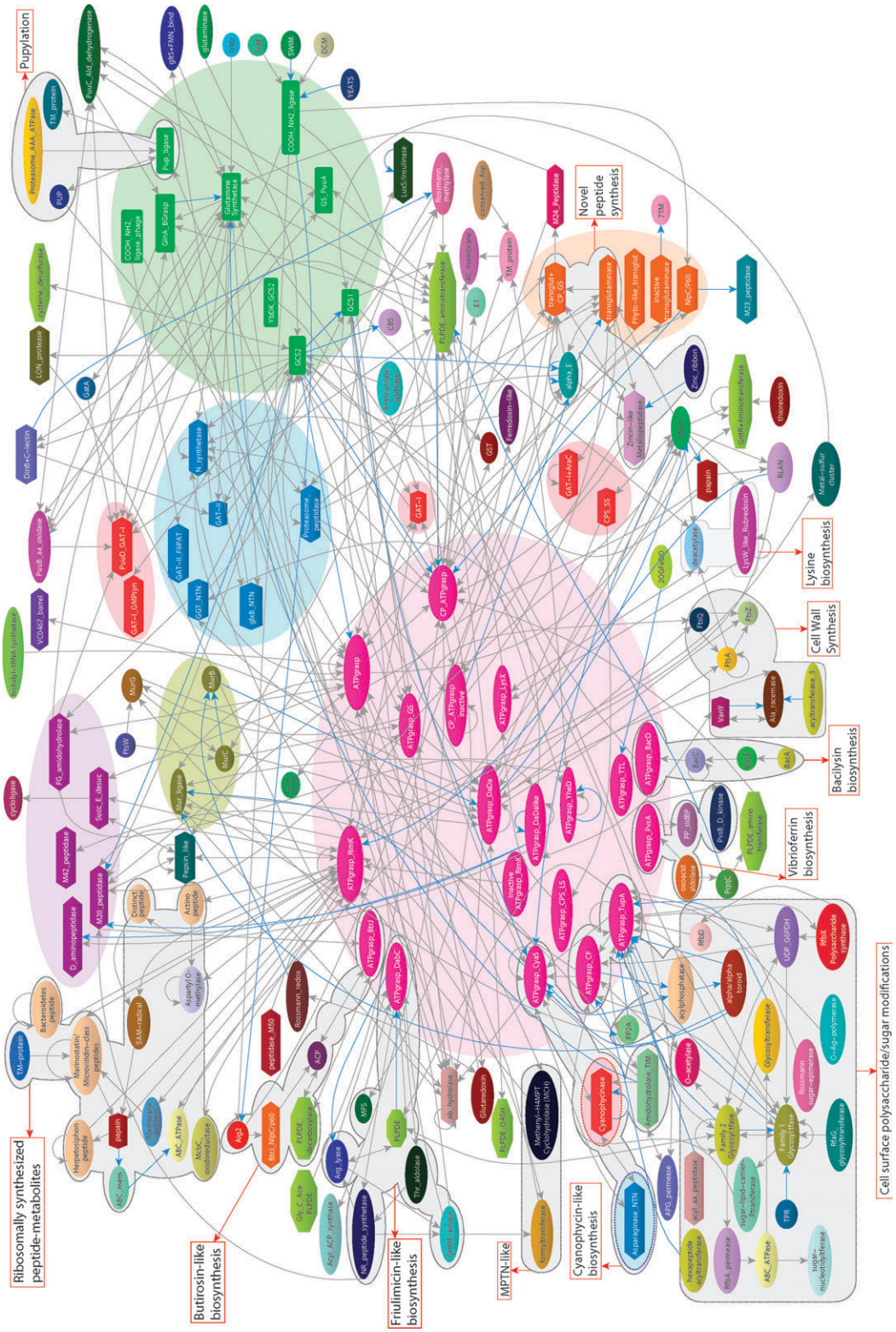


Fig. 5 Network of contextual linkages derived from gene neighborhoods and domain architectures. Grey arrows indicate information from gene neighborhoods and blue arrows indicate information derived from domain architectures. The direction of the arrows shows the order of genes in operons from 5′ to 3′, or order of domains from N-terminal to C-terminal. Domains that share the same fold are shown in the same shape and color. Thus, ATP-grasps are colored magenta, glutamine synthetases green, tranglutaminases orange, M20 peptidases purple, NTN-hydrolase fold peptidases blue, Flavodoxin fold GAT-I family red, and Mur ligases olive green. The shape of the domains corresponds to their functional role. For example, peptidases are shown as colored hexagons and the PLPDE-dependent enzymes in colored octagons. Domains are further grouped in two distinct themes. The principal domains that share a common fold are enclosed in colored oval shapes with light background and dotted lines and the corresponding nodes in darker shades of the same color. Further domains involved in a common pathway are enclosed within a grey background and dotted lines. Standard abbreviations have been used for gene names. The supplementary material (ESI) gives a full list of the abbreviations.

Fig. 6 Multiple sequence alignment of the alpha-E domain. Sequences are denoted by their gene names, species abbreviations and gi. Since the alpha-E domain is typically found as a tandem duplication, the N- and C-terminal domains of this duplication are grouped separately. Solo alpha-E domains are shown at the bottom. The depiction of secondary structure elements, species and consensus abbreviations, and coloring scheme are as in Fig. 2.

to unfold the protein substrates prior to proteolysis.[14,15] In prokaryotes genes of the AAA+ ATPases of these systems show a strong contextual linkage to those of the proteolytic subunits either in the same operon (*e.g.* the classical proteasome and HslUV) or *via* fusion in the same protein (*e.g.* Lon proteases).[15,61] However, not even in a single instance over 450 detected occurrences of this conserved operon did we observe a genuine linkage to an AAA+ ATPase. In contrast, this set of conserved gene neighborhoods, combining one or more peptide-ligase homologs (*i.e.* of the ATP-grasp and COOH–NH_2 ligase folds) with peptidases unconnected to ATPases showed a clear resemblance to the peptide synthesis systems, such as those implicated in peptidoglycan, cyanophycin or glutathione biosynthesis (Fig. 4, ESI). Hence, rather than being a proteasome-like protein degrading system, the Anbu peptidase is likely to merely be a peptide-cleaving system similar to peptidases in the above-stated systems. Absence of other basic metabolism-related genes in these gene neighborhoods, as well as the presence of intact glutathione and glutamine biosynthesis pathways in many of the organisms containing the above gene-neighborhoods further argues against a primary role for it in amino acid or glutathione biosynthesis. This system also does not show any linkage to genes encoding distinctive secondary metabolite-related enzymes (*e.g.* as seen in the bacilysin biosynthesis operon[37,62] or the acyl-carrier and amino-glycoside linking enzymes seen in the case of the butirosin biosynthesis operons (Fig. 4)[38]). Unlike the pupylation or E1–E2 systems, there is no evidence for a small protein in this system that might be conjugated to a target.

Taken together these observations indicate that the ATP-grasp ligase, the Anbu peptidase and other frequently linked genes such as the COOH–NH_2 ligase and transglutaminases constitute a novel system, distinct from all the previously characterized peptide synthesis systems and probably does not use any modified amino acids or specialized metabolites typical of antibiotics. We postulate that the ATP-grasp and COOH–NH_2 ligase (if present) in this system catalyze two distinct peptide bond formations. A key difference from all other peptide synthesis systems is the presence, without exception, of the unique alpha helical protein, which is only encoded in the predicted operons of this novel system and nowhere else. It is strongly linked to the ATP-grasp protein, either as its immediate neighbor or is fused to an inactive version of it (when a second peptide-ligase, the circularly permuted COOH–NH_2 ligase is present) suggesting that it physically interacts with the ATP-grasp and the COOH–NH_2 ligase. Analysis of this protein showed that it contains two divergent copies of a unique alpha-helical domain (termed hereinafter the alpha-E domain; Fig. 2 and 6); the archaeal versions contain a single stand-alone copy of the same domain. Each alpha-E domain contains 6 core helices of which the first helix contains the absolutely conserved ER signature (Fig. 6). In light of this absolutely conserved glutamate and the predicted physical interaction with the two distinct peptide ligases it is tempting to speculate that it serves as a substrate for elongation of a peptide *via* the gamma-carboxylate of its side chain. This proposal is consistent with the use of glutamate side chains as substrates in eukaryotic proteins such as tubulin by peptide tagging ATP-grasp enzymes.[4,5] The presence of two peptidase genes in practically all versions of these operons (Fig. 4) suggests that two successive peptidase reactions are necessary for removal of the peptide product. Alternatively, the transglutaminase superfamily protein might indeed function in cross-linking the peptide to lysine side chains or other amino groups.

Thus, the weight of the contextual evidence supports a role for this widespread conserved gene-neighborhood in peptide synthesis; the resulting peptide could be added as a tag to the unique alpha-E protein in this system. Such a tag could either regulate the assembly or interactions of the alpha-E domain protein (*e.g.* as in tubulin) or serve as an amino acid storage mechanism.

Other novel predicted peptide synthesis in bacteria

We additionally recovered several conserved gene neighborhoods defining systems that appeared to be *bona fide* peptide synthesis systems distinct from previously characterized ones (Table 1). The larger contextual connections showed that these systems participate in specialized processes that range from formation of small stand-alone peptides to possible cross-linking of proteins. Most of these showed restricted phyletic distributions and some of them appeared to be evolutionarily mobile as they were found sporadically in phylogenetically distant organisms. We discuss below some major examples of these systems along with the evidence in support of their role. Some of these systems contain predicted peptide ligases of a single kind *i.e.* just ATP-grasp or COOH–NH_2 ligases, whereas other systems are of a mixed type, including fusions of, or neighborhoods combining, one or more copies of different types of ligases.

Systems with multiple ATP-grasp ligases. A system comprised entirely of multiple ATP-grasp type ligases is defined by a conserved gene neighborhood found in sporulating firmicutes of the *Bacillus* lineage. This neighborhood contains a core comprised of 2–4 successive genes encoding distinct ATP-grasp proteins belonging to a distinct family within the peptide ligase clade (prototyped by *B. subtilis* YheC (gi: 16078043) and YheD (gi: 16078042); Fig. 2 and 4, Table 1 and ESI). The YheC/D family shows further gene neighborhood linkages to spore-specific proteins such as SspB and YheA. However, no genes encoding peptidases were found in these neighborhoods (Fig. 4). Developmental studies in *B. subtilis* suggest that YheD is distributed in two rings around the forespore and eventually forms an envelope around the entire forespore.[63] The analogy with previously characterized operons suggests that the YheC/D system catalyzes the formation of a peptide, with each ATP-grasp protein probably adding a distinct residue. The absence of peptidases in this suggests that these peptides are formed as part of a terminal maturation process with no immediate remodeling or degradation of these peptides. In light of the specific association of the YheC/D system with the forespore coat it appears plausible that the peptide chains synthesized by these enzymes either link spore-wall proteins by cross-links or else link spore-wall proteins to the underlying cortical peptidoglycan by specialized cross-links. Consistent with this latter

suggestion, in some members of the *Bacillus* clade the YheC/D is closely linked to Mur family ligases involved in peptidoglycan biosynthesis (Fig. 4). *Paenibacillus* has several paralogous versions of the YheC/D operon containing from a single to four copies of the ATP-grasp gene, one of which is linked to a gene for the PP2A phosphatase domain protein present in the capsular polyglutamate synthesis system[64] (Fig. 4). Hence, the peptide ligase activity of this YheC/D ATP-grasp might be linked to capsular biogenesis in this organism.

Another class of conserved gene neighborhoods with multiple ATP-grasp ligases (prototyped by MXAN_4097 (gi: 108757010) from *Myxococcus xanthus*) has a more sporadic distribution in certain proteobacteria, aquificae, crenarchaea and firmicutes (Table 1, ESI). These operons typically contain two ATP-grasp encoding genes linked to an aminopeptidase that is specific to peptide-bonds between D-amino acids (Fig. 3 and 4). This suggests that the peptide synthesized by this system is likely to contain a D-amino acid. Although the ATP-grasp protein is related to the D-Ala–D-Ala ligase, the majority of organisms with this system have a conventional cell-wall specific D-Ala–D-Ala ligase suggesting that it synthesizes a D-amino acid-containing metabolite distinct from the regular peptidoglycan peptides. Certain versions of this operon (*e.g.* in Rhizobia and aquificae) show a replacement of the above aminopeptidase by other peptidases such as those of the gamma-glutamyltranspeptidase family (NTN-hydrolase fold) or multiple versions of peptidases of the phosphorylase fold (''M20''-like).

Systems which combine ligases of more than one fold. We found several distinct operon types that combined different peptide ligases of different folds. The first of these is typified by a conserved gene-neighborhood found primarily in actinobacteria, chlorobi and proteobacteria (prototyped by PA3460, gi: 15598656, from *Pseudomonas aeruginosa*) encoding enzymes for the synthesis of a predicted storage polypeptide similar to cyanophycin. The core of this system is a cyanophycin synthetase-like protein with an ATP-grasp domain fused to a Gcn5-like acetyltransferase domain (Fig. 3 and 4). This GNAT domain is reminiscent of those found in the amino acyl tRNA-dependent peptide ligases and might catalyze a second peptide condensation reaction distinct from that catalyzed by the ATP-grasp domain. Additionally, these operons encode an asparagine synthetase and a metallopeptidase of the phosphorylase fold (''M20''/''M42''-like), which is unrelated to the flavodoxin fold cyanophycinase found in classical cyanophycin operons (Fig. 4). Typically, this operon also encodes an asparagine synthetase implying that the putative storage polypeptide produced by this system is distinct from cyanophycin and includes asparagine as one of the amino acids.

Several novel operon types combine genes encoding ATP-grasp and COOH–NH_2 ligases with those for different peptidases and generally resemble the prototype provided by the glutathione biosynthesis systems (Table 1, Fig. 4). Among components of the glutathione biosynthetic system, GCS1 is present in an operon with an ATP-grasp protein of the glutathione synthetase family (GshB) in genomes of certain proteobacteria and spirochaetes. In firmicutes and certain other proteobacteria, GCS1 is fused in a single polypeptide to a distinct family of ATP-grasp domains that are related to the cognate domains of the cyanophycin synthetase family[51] (Fig. 2 and 3, ESI). The presence of GCS1 with an additional ATP-grasp correlates well with the demonstrated production of glutathione in particular bacterial lineages, whereas the stand-alone GCS1 found in certain lineages is likely to catalyze the synthesis of the related metabolite gamma-glutamylcysteine.[35] A functionally equivalent thiol, mycothiol, is produced *via* the combined action of a cysteinyl tRNA synthetase paralog and a glycosyltransferase in several bacteria lacking glutathione. The biosynthetic pathway for mycothiol is present in actinomycetes[35,65] and chloroflexi (genes detected in this study). Most of the novel operons combining ATP-grasp and COOH–NH_2 ligases detected by us occur in organisms with intact glutathione or mycothiol biosynthesis pathways or those known to lack these metabolites.[35,65] Hence, they might have a role distinct from glutathione biosynthesis.

One group of these predicted operons (prototyped by the ATP-grasp gene RoseRS_2616 from *Roseiflexus*, gi: 148656737) is found in several phylogenetically distant bacteria such as the chloroflexi, *Gemmata*, *Gemmatimonas*, *Solibacter*, *Sorangium*, *Bacteroides*, firmicutes, actinobacteria and cyanobacteria (Table 1, Fig. 4). The typical neighborhood of this group encodes: (1) 1–2 distinct ATP-grasp proteins, one of which is circularly permuted and related to the version associated with the alpha-E proteins (see above). (2) One COOH–NH_2 ligase related to GCS2. (3) A peptidase of the alpha/beta hydrolase fold (Table 1, Fig. 4). Variant versions contain other peptidases in place of the alpha/beta hydrolase. For example, the two distinct versions of this operon from *Sorangium*, respectively contain a *Phytopthora*-type transglutaminase or a GAT-II like amidohydrolase (NTN-hydrolase fold) in place of the alpha/beta hydrolase fold peptidase (Fig. 4). In contrast, in *Gemmatimonas*, the peptidase is a GAT-I-like amidohydrolase (flavodoxin fold) and in *Rhodococcus opacus* it is a metallopetidase of the phosphorylase fold. In cyanobacteria we observe a displacement of the GCS2 family peptide ligase by one of the Fem/MurM family of the GNAT fold. Based on the glutathione template the presence of up to three distinct predicted ligases in the longest of these operons suggests they might synthesize peptides with a maximum of four residues.

A second distinct predicted operon-type combining ATP-grasp and COOH–NH_2 ligase genes is prototyped by that in enterobacteriophages such as phiEco32 (phi32_84; gi:167583639; Fig. 4). These phage gene-neighborhoods encode two highly divergent versions of the COOH–NH_2 ligase domain that are not closely related and an ATP-grasp ligase, implying that it could potentially catalyze three distinct peptide condensation reactions (Fig. 4, Table 1). The first of the COOH–NH_2 ligases of this neighborhood also has eukaryotic representatives in certain stramenopile algae and fungi. The algal versions are fused to an N-terminal YEATS domain, which is specifically found in eukaryotic chromatin proteins (Fig. 3),[66] suggesting that it might synthesize a peptide tag in chromatin proteins. The second COOH–NH_2 ligase encoded by the phage gene-neighborhood is also found in certain

endospore-forming firmicutes, where it is fused to or co-occurs in a predicted operon with one or more YheC/D-like ATP-grasp domains and spore-coat proteins such as CotE (Fig. 4, Table 1). The phage versions are linked to an NTN-hydrolase fold peptidase specifically related to the glutamine: D-hexose-6-phosphate amidotransferase. Thus, this system could potentially modify the cell-wall by linking a novel peptide and thereby prevent other phages from accessing the host. The comparable operon in sporulating firmicutes appears to be an analog of the classical YheC/D system that probably operates on the spore-coat.

A third operon-type which combines ATP-grasp and COOH–NH_2 ligases is found sporadically in both archaea and bacteria such as bacteroidetes, planctomycetes, verrucomicrobia and proteobacteria (prototyped by *Thiobacillus denitrificans* Tbd_1454, gi: 74317472 neighborhood, Fig. 4, Table 1). Its core encodes an ATP-grasp of the RimK family and a COOH–NH_2 ligase of the GCS2 family. The ATP-grasp protein of this system is distinguished by a fusion to a novel conserved N-terminal domain with an alpha + beta fold, which in the euryarchaea is encoded by a stand-alone gene that is an immediate neighbor of the cognate ATP-grasp gene (see ESI). The bacterial versions of these operons additionally encode two peptidases respectively of the phosphorylase fold (M20-family) and the papain-fold cysteine protease (Fig. 4). The archaeal versions belong to a more extended gene cluster, usually embedded in the highly conserved ribosomal operon, and encodes 4Fe-4S ferredoxin or another conserved metal-sulfur cluster protein (*e.g.* AF2306 from *Archaeoglobus fulgidus*, Fig. 4). In several bacteria we found a variant of this operon type that lacked the GCS2 family ligase, instead containing a gene for a GNAT superfamily protein in its place (Fig. 4). Interestingly, this GNAT protein is nearly always fused to a papain-fold cysteine protease similar to those in the GCS2-encoding version of the operon. Based on this we propose that this GNAT domain might act as a peptide ligase similar to those of the Fem/MurM family or acetylate the peptide produced by this system. The presence of at least two potential peptide ligases in most versions of this system suggests that it synthesizes a peptide with at least three amino acids. The contextual features of the archaeal versions of the operon suggest that the ribosomal protein, the 4Fe–4S ferredoxin or the second metal cluster protein (Fig. 3 and 4) are potential substrates for peptide-tagging by this system. The unique N-terminal domain of the RimK-like ATP-grasps might have a key role in recognizing a common class of substrates modified by this system.

Systems with a single ligase and peptidase. This type of system is prototyped by the PuuA–PuuD pair of a proteobacterial putrescine utilization system.[32,67] Here, the COOH–NH_2 ligase PuuA condenses alpha-L-glutamate and putrescine by forming an amide linkage, which is followed by oxidation of the linked putrescine moiety to gamma-amino butyrate. Then PuuD, which is an amidohydrolase of the GAT-I superfamily (flavodoxin fold), hydrolyzes the isopeptide bond in gamma-glutamyl-gamma-aminobutyrate. In this study we detected several distinct dyads of potential amide/peptide forming enzymes with a peptidase/amidase that resemble the PuuA–PuuD pair. One of these, typified by the *Aspergillus* FluG, is present mainly in actinobacteria, sporadically in cyanobacteria, firmicutes, deinococci, chloroflexi and proteobacteria, and in crown group eukaryotes such as fungi, plants and *Dictyostelium*.[68] This system combines a COOH–NH_2 ligase, which like PuuA is closely related to glutamine synthetase, with a TIM-barrel fold amidohydrolase[55] either in a single polypeptide or in an operon (Fig. 2 and 3). FluG diverged from glutamine synthetases in bacteria (ESI) and appears to have been laterally transferred from actinobacteria to the ancestor of crown group eukaryotes. Experimental characterization of it in both actinobacteria and fungi suggests that it is not involved in glutamine synthesis.[68,69] The predicted bacterial operons encoding FluG also typically encode an amino acid/polyamine transporter[70] and less frequently PuuD. Hence, like the PuuA-based system, FluG and its linked amidohydrolase might also constitute a polyamine or amino acid utilization system that acts *via* condensation of glutamate to an amino-group-bearing moiety. FluG is believed to be required for production of a soluble developmental signal in *Aspergillus* conidiation.[69] We propose that this signal is likely to be a compound such as aminobutyrate produced by the action of FluG and the amidohydrolase.

Two other comparable systems combine, in a conserved gene-neighborhood, a gene for a COOH–NH_2 ligase with a gene encoding a protein containing an N-terminal DinB and a C-terminal domain with the C-type-lectin-fold (Fig. 3 and 4). While the latter proteins are closely related to each other across these two systems, the COOH–NH_2 ligases of the respective systems are only distantly related. The first of these, found mainly in actinobacteria, contains a COOH–NH_2 ligase of the GCS2 family (*e.g.* Rv3704c; wrongly annotated as GshA in the nr database). The second is found only in gamma-proteobacteria and has a distinctive COOH–NH_2 ligase that forms a small family of its own (*e.g.* Patl_3664 from *Pseudoalteromonas atlantica*). In the latter system, the protein with DinB and C-type lectin fold domains is further fused to a C-terminal AdoMet-dependent methyltransferase of the Rossmann fold (Fig. 4). The C-type-lectin-fold domain in both the systems is specifically related to a version of the domain found in the peptide formyl-glycine synthesizing enzyme, but lacks the key catalytic residues of that enzyme.[71] This suggests that these C-type-lectin-fold domains are likely to be non-enzymatic and only bind a peptide. We had previously predicted the DinB domains to be a novel alpha-helical hydrolase domain with a catalytic site formed by 3 conserved histidines.[72] By analogy to other comparable systems, such as PuuA–PuuD and FluG, we predict that the DinB domain in these systems functions as a novel peptidase/amidase. A further widespread single-ligase system from firmicutes, certain proteobacteria and bacteroidetes (*e.g.* DSY4546 gene neighborhood from *Desulfitobacterium hafniense*, Fig. 4 and 5) groups in an operon a novel peptide/amide ligase of the COOH–NH_2 ligase fold (ESI) with a gamma-glutamyl cyclotransferase. This family of COOH–NH_2 ligases are also found combined in a predicted operon with a GAT-I- like peptidase (flavodoxin fold) in a mutually exclusively set of proteobacteria (Fig. 4, ESI). This mutual exclusivity of the gamma-glutamyl cyclotransferase with the GAT-I peptidases

suggests that they perform an equivalent role. In syntactical terms all these three gene neighborhoods are reminiscent of the PuuA-PuuD system. Hence, we predict that these systems are also likely to be involved in amine utilization with the COOH–NH_2 ligating a glutamate to the amine followed by its eventual removal either *via* a peptidase reaction or a cyclotransferase reaction catalyzed by the cyclotransferase to release oxoproline. Alternatively, a subset of these systems might have a role in producing novel small molecule metabolites from amino group-containing compounds.

We also identified a single-ligase system centered on a RimK family ATP-grasp ligase (*e.g.* PA1766 from *Pseudomonas aeruginosa*, Fig. 4), which is combined with genes encoding a pepsin superfamily peptidase and a novel membrane protein with seven membrane-spanning segments (7-TM) and an N-terminal extracellular inactive transglutaminase domain. This predicted operon is found in proteobacteria, cyanobacteria and planctomycetes (Fig. 4, ESI) and its ATP-grasp ligase is distinguished from the paralogous classical RimK proteins which are often found in a distinct context (see below). The architecture of the 7-TM protein encoded by this operon, with a large extracellular domain potentially involved in ligand-binding, is suggestive of a membrane-associated receptor. However, the predicted intracellular loops of the 7-TM region contain several absolutely conserved glutamates and arginines and a glycine-rich loop suggesting that it might potentially function as a membrane-associated enzyme which responds to an extracellular stimulus. The tight linkage of the three genes indicates that this RimK-like ATP-grasp might carry out a peptide ligase activity strictly in connection with the 7-TM protein—it could either modify the intracellular regions of the 7-TM protein or alternatively modify a small molecule, such a peptide or F420/pterin-like molecule which interacts with the 7-TM protein. The peptidase in this system is likely to reverse the modification as in several other such systems discussed earlier.

ATP-grasp and COOH–NH_2 ligase domains in synthesis of complex peptide-derived metabolites

We identified several predicted amide/peptide bond-forming enzymes in this study, which are embedded in gene-neighborhoods indicative of a function in the synthesis of complex secondary metabolites (Fig. 1 and 5) including diverse antibiotics. Some exemplars of such pathways have been characterized to differing degrees but their diversity and reaction mechanisms remain incompletely understood. The previously studied pathways include those involved in biosynthesis of friulimicin produced by *Actinoplanes friuliensis*, bacilysin by *Bacillus pumilus*, butirosin by *B. circulans*, teichuronopeptide-type metabolites and the siderophore vibrioferrin by *Vibrio*.[37–39]

Systems prototyped by the Friulimicin and bacilysin biosynthesis gene clusters. In the friulimicin system the ATP-grasp protein DabC is encoded in the Dab gene cluster involved in the biosynthesis of a key amino acid, 2,3-Diaminobutyric acid, which is found twice in the sequence of this 11-residue antibiotic. DabC was found to be required along with DabA, a pyridoxal phosphate dependent enzyme (PLPDE) related to cystathionine-beta lyase, and DabB, a fumarase related to argininosuccinate lyase.[73] Such a system was also noted in rhizobia, which are known to contain 2,3-diaminobutyrate,[73] but the mechanism for the synthesis of 2,3-diaminobutyrate remains unknown. We detected comparable conserved gene neighborhoods in several other actinomycetes, sporulating firmicutes and certain proteobacteria (*e.g.* *Burkholderia*) (ESI). We observed that these gene neighborhoods additionally contained one or both of two key threonine biosynthesis enzymes, namely threonine aldolase that synthesizes threonine from glycine and acetaldehyde, and the homoserine kinase, indicating that a key substrate for this pathway was threonine (Fig. 4). Based on this, we could reconstruct the synthesis of 2,3-diaminobutyrate *via* an aminotransferase reaction involving threonine and aspartate catalyzed by DabA and the subsequent release of fumarate through the action of the fumarase DabB (Fig. 1). The ATP-grasp protein, in addition to occurring in an operon with the above enzymes, might also be found fused to the fumarase or the aminotransferase. Hence, it is likely that it catalyzes a reaction closely linked to the synthesis of 2,3-diaminobutyrate. Accordingly we predict that it catalyzes the protection of the 2-amino group of 2,3-diaminobutyrate by ligating an amino acid to it (Fig. 1). In the case of friulimicin, this amino acid could be the first asparagine found in its sequence.[73] In most actinobacteria these gene neighborhoods additionally contain one or more multidomain non-ribosomal peptide synthetases, suggesting that in each of these cases the 2,3-diaminobutyrate is incorporated into a larger peptide antibiotic (Fig. 4).[73] However, in some organisms, *e.g.* *Brevibacillus brevis*, there is no such association with the multidomain peptide-ligases, suggesting that 2,3-diaminobutyrate might just be part of a dipeptide synthesized by the action of the ATP-grasp enzyme. A variant of these operons, *e.g.* those found in *Mesorhizobium loti*, *Alkaliphilus metalliredigens*, *Serratia proteamaculans* and *Geobacillus* are characterized by the presence of two adjacent genes encoding paralogous ATP-grasp proteins (Fig. 4, Table 1). It is likely that the second ATP-grasp in these operons catalyzes a further peptide ligation reaction, potentially resulting in a tripeptide. The version of this gene neighborhood in *A.metalliredigens* is further tightly linked to a paralog of the histidinyl tRNA synthetase suggesting that this enzyme might also catalyze the incorporation of a histidine residue in the peptide (Fig. 4). The majority of the gene neighborhoods also encode a linked transporter, which might play a role in the efflux of the peptides synthesized by them.

We also uncovered another entirely uncharacterized set of gene neighborhoods which appear to define a biosynthetic system comparable to the prototype provided by the friulimicin 2,3-diaminobutyrate biosynthesis system described above (Fig. 5). This gene neighborhood is found in different actinobacteria and the myxobacterium *Haliangium* (*e.g.* the FRAAL4660, gi: 111224051, neighborhood of *Frankia alni*, Fig. 4). This neighborhood encodes a circularly permuted ATP-grasp, a PLPDE related to 2-oxo acid aminotransferases, a 2-oxo acid-forming aldolase (related to 4-hydroxy-2-oxovalerate aldolase[74]) and a PqqC family oxidoreductase[75] (Fig. 4). By analogy to the friulimicin 2,3-diaminobutyrate

biosynthesis pathway, we propose that the aldolase synthesizes a 2-oxo acid, followed by the action of the aminotransferase to generate an amino acid. The presence of the PqqC enzyme suggests that the side chain of this amino acid might be further desaturated by the action of this enzyme. A subset of these neighborhoods also possesses a methyltransferase that might also modify the amino acid. It is likely that the ATP-grasp protein in this system, as suggested in the classical friulimicin-like pathways (see above), ligates a second amino acid to the modified amino acid synthesized by the former enzymes. Gene deletion studies in *Streptomyces fradiae* suggest that this gene cluster is required for the anti-microbial capability of this organism supporting a role in synthesis of an antibiotic.[76]

Other sporadic, tightly linked gene clusters also combine a comparable group of enzymes. One such from *Streptomyces sviceus* encodes a circularly permuted ATP-grasp, an amino acid dehydrogenase, a metallopeptidase (M24) and a PLPDE aminotransferase, which acts on 2-oxo acids (Fig. 4, ESI). Another gene cluster from *Photorhabdus luminescens* encodes a giant protein (plu2191) which combines three domains, namely a PqqC oxidoreductase, a PLPDE aminotransferase and a circularly permuted ATP-grasp (Fig. 3 and 4). The aminotransferase is specifically related to those involved in the synthesis of 2,4-diaminobutyrate. This gene cluster further includes an E1-like adenylating enzyme, a 2-oxoglutarate-dependent dioxygenase and a GNAT superfamily protein. Here again it is likely that the synthesis of an amino acid formed by the transfer of an amino group to a 2-oxo acid by the PLPDE is followed by ligation of an additional amino acid by the ATP-grasp to form a dipeptide. Other enzymes in these systems (Fig. 3 and 4) suggest that there are likely to be further modifications of the amino acids catalyzed by them. This proposal for the synthesis of dipeptide metabolites with modified amino acids is supported by a comparable system found in several *Bacillus* species, which synthesizes the dipeptide antibiotic bacilysin.[62] Here the ATP-grasp is linked to genes encoding a prephenate dehydratase (BacA), a cupin superfamily dioxygenase (BacB), an amino acid dehydrogenase (BacC), a PLPDE aminotransferase (ywfG) and a peptide transporter (Fig. 4). Based on the reactions predicted to be catalyzed by the latter set of enzymes we could completely explain the synthesis of the highly modified amino acid anticapsin by successive dehydration, reduction and oxygenation and amino transfer steps (Fig. 1).

Systems prototyped by the butirosin biosynthesis gene clusters. The antibiotic butirosin is a complex metabolite that is produced by combining an aminoglycoside and a highly modified peptide derivative.[38,77] Several components of this pathway have been studied in the context of the synthesis of other aminoglycoside metabolites[78] but not all components related to synthesis of the peptide portion are fully understood. The core of the peptide synthesis part of the butirosin system has an ATP-grasp enzyme (BtrJ), which catalyzes two steps, namely the gamma-glutamylation of an acyl carrier protein (ACP), BtrI, and the subsequent ligation of a second glutamate to the amino group of first ACP-linked glutamate (Fig. 1). Additionally, the system also has a second ACP, BtrV, which might receive the peptide just prior to its transpeptidation to the aminoglycoside. The second glutamate in this system appears to have a role in protecting the NH_2 group of the ACP-linked glutamate because it is removed by a gamma-glutamyl cyclotransferase, BtrG (Aig2 superfamily), in the final stage of butirosin biosynthesis.[38] The transpeptidase reaction, which transfers the peptide from the ACP to the aminoglycoside, is catalyzed by BtrH that does not contain any previously characterized domain.[38] Using sequence profile searches with the PSI-BLAST program we detected a large number of homologs of BtrH in various bacteria and unified it with the NlpC/p60 fold peptidases of the papain-like fold (ESI). All members of the BtrH family of proteins contain a conserved cysteine and histidine characteristic of catalytically active versions of the papain-like fold. Members of the BtrH family are often found linked in operons or fused in the same polypeptide to non-ribosomal peptide synthetases, which along with polyketide synthetases, also utilize an ACP base in their synthetic reactions.[38,78] It is likely that they function in the context of ACPs as trans-peptidases involved in transfer of peptide linkages during peptide metabolite biosynthesis (*e.g.* transglutaminase or lecithin:retinol acyl transferase[50]). The transpeptidase activity proposed for the BtrH family is thus comparable to that of the transglutaminase family protein admF in andrimid biosynthesis.[79]

We uncovered several gene clusters that appear to resemble the core of the butirosin peptide ligation system in different endospore-forming firmicutes, actinobacteria and a *Burkholderia* species (*e.g.* AmirDRAFT_03860; gi: 226865117 from *Actinosynnema mirum* and Bcep18194_A4990 from *Burkholderia* sp.). While most of these gene clusters encode a gene for a BtrJ-like ATP-grasp protein linked to a gene for an ACP, the version from *Burkholderia* encodes two circularly permuted ATP-grasp proteins. These gene-neighborhoods differ from the conventional butirosin system in lacking the BtrH-like transpeptidase and the gamma-glutamyl cyclotransferase. However, a subset of these gene clusters encodes a MurG family glycosyltransferase which might combine the peptide formed by the remainder of the system to a carbohydrate moiety. Furthermore, the presence of a BtrK-like decarboxylase in a subset of these systems suggests that the first glutamate might be decarboxylated to form an amine as in the case of the butirosin pathway. In the *Burkholderia* version of this neighborhood there is a tightly linked gene encoding a glycine C-acetyltransferase, which is known to synthesize L-2-amino-3-oxobutanoate (Fig. 4). It is possible that this modified amino acid is used as part of the peptide formed by this system.

Systems modifying ribosomally synthesized peptides. The related peptides marinostatin and microviridin respectively produced by *Alteromonas* and *Microcystis* are defensive protease inhibitors that appear to be deployed against predators such as crustaceans.[41] Their precursors are synthesized ribosomally and subsequently modified by a pair of paralogous ATP-grasp enzymes that catalyze three cyclization reactions (Table 1). One of these enzymes forms a conventional amide linkage between a conserved lysine and glutamate

side chain in these peptides, whereas the other enzyme synthesizes two lactone linkages between serine/threonine and other conserved acidic residues.[42] Marinostatin/microviridin homologs are also widely encoded in several other cyanobacteria, bacteroidetes and myxobacteria[41] (and this study), and these peptides preserve the same pattern of a conserved lysine, two alcoholic and three acidic residues suggesting they are similarly cyclized. In the majority of these organisms 1–7 marinostatin/microviridin peptides and the two ATP-grasp enzymes are encoded in a gene neighborhood. Occasionally these neighborhoods may also encode a GNAT protein that acetylates the NH_2 end of the peptide and an ABC transporter with a fused papain-like peptidase domain that is required both to cleave the pre-peptide and transport it out of the cell. We observed that the two paralogous ATP grasp proteins of this system belong to a distinct family of ATP-grasp enzymes related to the more widespread RimK family (Fig. 2, Table 1 and ESI). Analysis of other members of this peptide-modifying ATP-grasp family that do not co-occur with marinostatin/microviridin revealed at least 10 distinct sub-groups, most of which are encoded by genes linked to those for other distinctive peptides. Most of these subgroups are also distinguished from the marinostatin/microviridin gene clusters in specifying only a single ATP-grasp enzyme.

Peptides encoded by four of these subgroups are characterized by the presence of one to several repetitive modules with conserved alcoholic (usually threonine) and acidic residues with spacing similar to that seen in the marinostatin/microviridin peptides (ESI). This strongly suggests that they are cyclized by the accompanying ATP-grasp *via* 1–2 lactone linkages. Further, at least one of these groups also shows a conserved lysine in addition to multiple conserved acidic residues suggesting cyclization *via* amide linkages (*e.g.* *Pseudomonas syringae* Psyr_2650; gi: 66045886). Several of these gene neighborhoods also encode a multi-TM protein (Fig. 4 and ESI) which could be required for the efflux of these peptides. These systems are widely distributed across several phylogenetically distant bacterial groups suggesting that they have been dispersed by lateral transfer due to the selective advantages they provide. Of particular interest is our identification of such potentially cyclized peptides in human, insect and plant pathogens such as enteropathogenic *E. coli* O127:H6 (gi: 215487193) *Bacillus thuringiensis* (gi: 228924946), and *Pseudomonas syringae* (see above), suggesting that such modified peptides might have a notable role in pathogenesis of these organisms. The remaining subgroups that encode associated peptides show no detectable similarity to marinostatin/microviridin or the peptides of the above describe four subgroups. However, they possess their own unique sets of acidic, lysine and alcoholic residues (ESI) suggesting comparable cyclizations. Most of these subgroups are restricted in their distribution. For example, one of the predicted peptides is widely conserved across actinobacteria, whereas another is restricted to bacteroidetes, and yet others are found only in the genus *Streptomyces* or encoded in multiple tandem copies in *Herpetosiphon*. In the case of the actinobacterial peptide the conserved gene-neighborhood always encodes a member of the aspartyl O-methyltransferase which methylates the beta-COOH group of aspartate. This suggests that the side chain of one of the conserved acidic residues in this peptide is modified *via* methylation. Such a secondary modification is also likely in the case of the bacteroidetes peptides. These operons often additionally encode radical SAM-dependent enzyme. This might catalyze a methylthiolation or desaturation or heterocyclic ring formation to modify a side chain like other members of this family such as MiaB, RimO, coproporphyrinogen III oxidase and biotin synthase (ESI). Other modifications are indicated in certain actinomycete operons which encode McbC-like flavin-dependent oxidoreductases, which are predicted to catalyze formation of oxazole or thiazole rings in other ribosomally encoded peptide metabolites.[10] A single subgroup shows apparently no peptide-encoding gene linked to the ATP grasp, but instead encodes a second degenerate paralog which conserves only the pre-ATP-grasp domain.

These newly detected ribosomally synthesized peptides and their modifying systems are found in diverse bacteria, which include lineages such as actinobacteria which are known to produce numerous antibiotics. Indeed some of these peptides might function as antibiotics; alternatively they could also be diffusible signals used by these developmentally complex bacteria. In the case of bacteroidetes such as *Kordia algicida* they could potentially be involved in their algicidal properties or provide an anti-predatory mechanism. Presence of multiple tandem copies of these peptides in organisms such as *Microscilla*, *Herpetosiphon* and *Kordia*, or multiple repeats in the same polypeptide is consistent with such a defensive role which could select for diversity in the peptides.

Systems prototyped by the teichuronopeptide biosynthesis pathway. Several bacteria contain distinctive polysaccharide-peptide conjugates in addition to peptidoglycan on their cell surfaces.[80] One such is teichuronopeptide, a highly acidic copolymer of glucuronic acid and amino acids such as glutamate that contributes to alkaliphily[81] of organisms such as *Bacillus halodurans* (*Bacillus lentus*). Experimental studies have implicated the TupA gene in the biosynthesis of this product[82] but the mode of action of this protein has not been understood. We detected an ATP-grasp related to the D-Ala–D-Ala ligase version in TupA (Fig. 2, ESI) and accordingly suggest that it is the ligase required for synthesis of the polyglutamate portion of the teichuronopeptide (Fig. 1). The *B.halodurans* TupA is present in an operon that additionally encodes three paralogous proteins with an ATP-grasp related to the cyanophycin synthetase family, fused to a C-terminal acylphosphatase domain (Fig. 4). These genes are further embedded within a larger gene cluster involved in teichoic acid biosynthesis and transport. A comparable combination of genes is also seen in alkali resistant bacteria such as *Dethiobacter alkaliphilus* and *Oceanobacillus*, and the polycyclic aromatic hydrocarbon degrading *Mycobacterium* sp. JLS. The gene neighborhoods from *D.alkaliphilus* and *Mycobacterium* sp. JLS are characterized by the presence of a third type of ATP-grasp that is related to the RimK family (Fig. 3 and 4). It is conceivable that the additional ATP-grasp proteins in the gene neighborhood catalyze further linkages between amino acids or between amino acids and sugars in

teichuronopeptides. The lateral transfer of this neighborhood might have been important in the emergence of alkali resistance in various distantly-related bacteria. The version in *Mycobacterium* sp. JLS is different in that the ATP-grasp proteins contain fusions to alpha–alpha toroid domains (Fig. 3 and 4). Such alpha–alpha toroids are also fused to a poly-gamma glutamate synthetase-type Mur-ligase in *Mycobacterium sp. KMS*. It possible that these proteins are involved in the synthesis of the highly modified cell surface mycolic acid derivatives in these mycobacteria with the alpha–alpha toroids acting as scaffolds for the synthesis of these molecules.

Members of the TupA family are also detected in a wide range of bacteria such as firmicutes, actinobacteria, proteobacteria, spirochaetes, bacteroidetes, fusobacteria and cyanobacteria (Table 1). These include wfdG and wfdR involved in *E.coli*/*Shigella* O-antigen biosynthesis operons, and the *Streptococcus pneumoniae* wcyV involved in capsular biosynthesis.[83,84] These TupA family genes are combined with genes that encode proteins involved in biosynthesis of cell-surface polysaccharides such as the O-Antigen in proteobacteria and the capsule in firmicutes (Fig. 4 and 5, ESI,Table 1). The TupA family ATP-grasp is also fused in some of these organisms to family 1 and family 2 glycosyltransferases, and capsular biosynthesis-type PP2A-fold phosphatases (Fig. 3). These operons might also encode multiple paralogous copies of TupA or the RimK-related ATP-grasp protein found in the above-described neighborhoods from *D.alkaliphilus* and *Mycobacterium* sp. JLS. Comparable operons with just this family of RimK-related ATP-grasp genes are also found sporadically in bacteria and one euryarchaeon in combination with other capsular biosynthesis genes. Studies in *Proteus* and *Providencia* have shown that sugars of the cell surface O-antigen are further aminoacylated by D- and L-aspartic acid residues.[85,86] We predict that ATP-grasp genes in these operons catalyze this ligation of amino acids to sugar moieties in these polymers. The wide phyletic distribution of TupA-family centered and related operons suggests that sugar/sugar acid and amino acid conjugates are a common feature of the capsules and other distinctive cell surface polymers of a large number of bacteria. The presence of up to four ATP-grasp genes in some of these operons suggests peptide chains with complexity comparable to the peptide linkages in peptidoglycan might be present in some of these polymers.

Siderophore and cofactor modifying systems. The siderophore vibrioferrin[39] belongs to a large class of siderophores that are condensation products of amino acids and 2-oxo acids.[87] The primary condensation reactions of these siderophores are catalyzed by the PvsB/D type ligases which belong to the serine/threonine/tyrosine (STY) kinase fold.[88] Previously, only the vibrioferrin biosynthesis system from *Vibrio* species was characterized as containing an additional ligase of the ATP-grasp fold.[87] We found comparable operons in various other proteobacteria, actinobacteria and *Deinococcus* (Fig. 4). Most of these operons are rather stereotypic and combine the ATP-grasp gene with genes encoding STY kinase fold ligases with siderophore transporters and receptors. This suggests that the system for synthesis of vibrioferrin-like siderophores has been widely disseminated across phylogenetically distant bacteria by lateral transfer.

We also obtained evidence for new functional links for the MptN and CofF families of ATP-grasp peptide ligases and for the presence of parallel systems in bacteria. Both the MptN and CofF families of peptide ligases appear to be derived from the more universal LysX family in the archaea (ESI). The LysX-like lineage appears to be the archaeal equivalent of the RimK lineage of the bacteria and the two were probably represented by a common ancestral version in LUCA. We observed that the MptN family, which catalyzes the ligation of a glutamate residue to tetrahydromethanopterin to form tetrahydrosarcinapterin,[33] has been laterally transferred to the bacterial lineages (*e.g.* planctomycetes and proteobacteria). These bacterial versions show a strongly conserved neighborhood with enzymes such as methylene tetrahydromethanopterin cyclohydrolase, the formaldehyde activating enzyme and tetrahydromethanopterin formyl-transferase (Fig. 4, ESI) which are involved in the biosynthesis of tetrahydromethanopterin. Thus, it is likely that these bacteria produce a similarly modified pterin co-factor. Further, the bacterial operons contain a second ATP-grasp gene, suggesting that there might be an additional glutamylation of the co-factor in these organisms. The CofF family ATP-grasp catalyzes the addition of the third glutamate residue to the coenzyme F420,[33] whereas the first two are added by an unrelated ligase termed CofE.[34] We detected a Mur family peptide ligase in the conserved gene-neighborhoods containing the CofF gene (Fig. 4), which might catalyze the ligation of the fourth glutamate which is known to occur in F420.[34] This is reminiscent of the Mur ligase which ligates glutamate to pterin-derived folates of bacteria.[33] Additionally, the MptN gene neighborhoods might encode a M20 family metallopeptidase (phosphorylase fold), whereas the CofF gene neighborhoods encode a M50-like metallopeptidase (Fig. 4). Thus, these peptide-modified cofactors could be regulated by removal of the peptide moieties by these linked peptidases.

We found that the RimK lineage, the bacterial counterpart of the LysX-MptN-CofF clade of the archaea, are also tightly linked in predicted operons to two distinct peptidases, one of the pepsin superfamily and the second related to succinyl-glutamate desuccinylase of the phosphorylase fold (see SCOP database; http://scop.mrc-lmb.cam.ac.uk/scop/). Interestingly, we found in several bacteria this gene-neighborhood contains genes encoding 5-formyltetrahydrofolate cyclo-ligase which is a key enzyme in folate metabolism.[89] At least in these bacteria, it is likely that the glutamylation of folate or a related pterin-derivative is catalyzed by the RimK-like enzymes. While in most bacteria the Mur ligase, folylpolyglutamate synthetase, is believed to catalyze the ligation of glutamates to folates, it is conceivable that as in the MptN and CofF systems there might be more than one glutamate ligase catalyzing successive glutamylations. It is also possible that RimK might catalyze glutamylation of other co-factors in these organisms. Whatever the case, these observations do suggest the possibility that the common ancestor of the LysX-like and RimK-like families might have already catalyzed glutamylation of cofactors in LUCA.

Evolutionary considerations

Origin of non-ribosomal peptide/amide formation activity and relationship to other reactions. The ribosomal synthesis of peptides is remarkable in being catalyzed by one of the two universal ribozymes (the other being RNAse P) that might have been inherited directly from the earliest protein-synthesizing systems of the so called "RNA-world".[90] While several non-ribosomal peptide ligases emerged subsequently, this ribozyme was never displaced by a protein catalyst. A survey of non-ribosomal peptide synthesis systems shows that this activity has emerged independently in several distinct folds: (1) ATP-grasp, (2) STY kinase, (3) COOH–NH2 ligase (glutamine synthetase-like), (4) GNAT (acetyltransferase fold), (5) Mur ligases of the P-loop kinase superfamily, (6) Condensation domain of peptide synthetases, (7) the E1–E2–(HECT E3) ligase system and (8) CofE-like proteins. Of these the ATP-grasp and the STY kinases are related folds sharing a common module[44,46] while the rest are unrelated to each other. Thus, the multiple origins of the peptide ligase activity are clearly convergent.

Phyletic patterns of the ATP-grasp, the STY kinase-like, the COOH–NH_2 ligase, the GNAT and P-loop NTPase domains indicate that they were already present in the last universal common ancestor (LUCA) (see ESI).[52] In case of the ATP-grasp fold, three of the families traceable to LUCA are involved in purine metabolism (PurD, PurK and PurT, ESI). Further, the only representative of the STY kinase-like fold which can be confidently traced to LUCA is the SAICAR synthetase (PurC) (ESI) that retains the primitive features of the fold shared with the ATP-grasp.[91] Of these PurD, PurT and PurC catalyze amide bond formation reactions similar to peptide ligation. This suggests that the common ancestor of the ATP-grasp and STY kinase-like folds was probably a generic ligase that functioned in the context of purine metabolism. Thus, an amide-bond forming activity related to peptide ligation was probably an ancestral feature of the STY-kinase and ATP-grasp folds and emerged well-before LUCA itself. The inference of the common ancestor of the RimK and LysX-MptN-CofF in LUCA suggests that *bona fide* peptide ligase activity had emerged in the ATP-grasp fold by this time. This probably marks the first emergence of a genuine non-ribosomal peptide synthesis mechanism.

The amide-forming reactions in glutamine synthesis appear to be an ancestral feature of the COOH–NH_2 ligase fold and emerged prior to LUCA (ESI). However, *bona fide* peptide ligation reactions such as those in glutathione synthesis and pupylation appear to have emerged only in the bacterial lineage. While the activity of the SAICAR synthetase suggests an ancestral amide-forming reaction in the STY-kinase like fold, the peptide-ligases involved in siderophore biosynthesis (see above) are closer to the protein kinases than to SAICAR synthetase.[88] Hence, these peptide ligases appear to represent a secondary re-acquisition of this activity from a kinase ancestor in the bacterial lineages. A similar kinase to peptide-ligase transition is postulated for the Mur ligases in the bacterial superkingdom, albeit from the entirely unrelated P-loop fold.[52] Interestingly, these peptide ligases also emerged in the context of bacterial peptidoglycan biosynthesis and cofactor glutamylation just as seen in some of the above peptide ligases. While present in LUCA, GNATs acquired their role as peptide ligases only in the bacterial lineage in the context of peptidoglycan metabolism (Fem ligases) and the N-end rule. Among the components related to ubiquitin-ligation, E1 enzymes were present in LUCA but functioned as enzymes which adenylated and thiocarboxylated Ubls.[10] A rudimentary form of peptide ligation emerged as a part of a distinctive bacterial cysteine biosynthesis pathway catalyzed by E1 enzymes. But only upon partnering with the E2 enzymes, which first emerged in the bacterial superkingdom, did they become part of the Ub-peptide ligation system.[8,10]

Other peptide-ligase folds appear to be purely lineage-specific innovations. The condensation domain of the giant non-ribosomal peptide synthetases appears to have emerged from the CoA-dependent acyltransferases probably in the actinobacteria as a part of the biosynthetic pathway for antibiotics and other secondary metabolites. Subsequently they appear to have been widely disseminated across bacteria. On at least one occasion a class-I amino acyl tRNA synthetase appears to have been recruited to form an amide linkage like a peptide bond in synthesis of mycothiol in certain bacterial lineages (see above).[35,65] The presence of a paralog of the class-II histidinyl tRNA synthetase in a potential peptide synthesis operon (see above) suggests that there might have been independent recruitments of tRNA synthetases for such reactions. While most innovations of peptide-ligases appear to have occurred in bacteria, thus far only one such innovation is seen in the archaea—the distinctive CofE fold.[34] However, it appears to have been laterally transferred to several bacteria, where it is often fused to an asparagine synthetase domain (ESI) indicating that it might participate in peptide ligation reactions other than the modifications of cofactors which are observed in archaea.

Interestingly, many of the folds in which peptide ligases emerged also share commonalities in the other reactions they catalyze. On one hand, the ATP-grasp, STY kinase, the P-loop kinase superfamily and the COOH–NH2 ligases include both kinases and ligases. On the other hand the GNAT fold, condensation domain of peptide synthetases and the E1–E2 system generally utilize diverse thioester intermediates; a similar reaction is also observed in certain members of the ATP-grasp fold as in BtrJ in butirosin biosynthesis or the succinyl CoA synthetase.[20,45,92,93] As described above in the STY kinase there appears to have been transitions in both direction from ligases to kinases and back. Whereas Mur ligases have emerged from a kinase ancestor of the P-loop fold,[52] the kinases of the COOH–NH2 ligase fold (*e.g.* arginine and creatinine kinases) have a more restricted distribution and are likely to have been derived relatively late in bacteria from an ancestral GatB-type enzyme (ESI). In the ATP-grasp fold, other than in one of the domains of the carbamoyl phosphate synthetase, the decoupling of ligase and kinase activities rarely took place. Thus, these folds appear to have exploited their basic ability to utilize the free energy of phosphate linkages or carbon-sulfur linkages in either stand-alone form (kinase reaction) or as part of a more complex reaction (peptide/amide bond formation).

Diversification of non-ribosomal peptide synthesis systems in bacteria. In an earlier work on evolution of E1 enzymes we observed that their diversification was closely linked to the pathways to which they were recruited.[10] Versions involved in ancient metabolic pathways tended to be conservative, whereas those involved in synthesis or modification of secondary metabolites such as peptide antibiotics and signaling molecules tended to show an enormous diversity both in terms of the E1 enzyme itself as well as its predicted operonic associations. However, these diverse operonic associations tended to draw from a relatively small common group of enzymes such as acetylases, methylases and functionally related but structurally distinct peptidases.[10] We observed a remarkably parallel diversification of the peptide ligases considered in this study. Most ancient representatives of these folds involved in amino acid and purine metabolism are rather conservative. The versions involved in glutathione and peptidoglycan biosynthesis show greater lineage-specific diversification, but those involved in the various, more sporadically distributed peptide and secondary metabolite synthesis/modification systems appear to be far more diversified in their operonic associations and architectures. Nevertheless, there are some stereotypic contextual linkages that appear to be preserved throughout the diversification of these peptide ligases (Fig. 5, Table 1).

The widespread nature of these connections suggests two possible explanations that are not mutually exclusive: (1) multiple convergent assemblies of operons or domain architectures with similar syntax involving peptide ligases and peptidases due to the selective pressure for tight functional cooperation. (2) Duplications of an operon or architecture prototype followed by *in situ* displacement of particular components by evolutionarily distinct but functionally equivalent counterparts (usually the peptidases and in some cases peptide ligases). While it is difficult to differentiate between the two, certain examples support one or the other scenarios. In the case of the cyanophycin synthetase and the related cyanophycin synthetase-like system it is clear that the core ATP-grasp domains are closely related but the second ligase domain is of a different fold. Given that they are both likely to catalyze two ligation steps it is plausible that there has been an *in situ* displacement of the second ligase by a functional equivalent. Similarly in the case of the firmicute glutathione forming enzyme it appears that the classical glutathione synthetase ATP-grasp ligase was displaced by a ligase related to the cyanophycin synthetase ATP-grasp domain. However, it is quite possible that the more general connections between different peptidases and ligases emerged convergently due to selective pressures of functional cooperation (*e.g.* Pup ligase and proteasomal peptidases or the cell-wall ligases and VanY peptidases).

Further, as in the case of the E1 system, we noted that the diversification of these peptide ligases in the context of secondary metabolite and peptidoglycan metabolism was accompanied by emergence of operonic associations with a relatively small pool of enzymes, such as PLPDE aminotransferases, methylases and acetylases (Fig. 5). This phenomenon is mainly observed in phylogenetically diverse, non-autotrophic bacteria with large genomes and a complex metabolism (ESI). The increased number of duplications and lateral transfers that accompany an increase in genome size seem to provide multiple paralogous copies of genes that serve as the evolutionary raw material for generation of secondary metabolism pathways. Organisms with large genomes also tend to divide more slowly. Hence, production of antibiotics and other secondary metabolites, which might inhibit other microbes to prevent resource competition, repel predators or kill other cells to release their nutrients, provides these organisms with a major selective advantage. Most of these pathways appear to form around a core set of genes encoding one or more ATP-grasp ligases (less frequently a ligase of some other fold) and a peptidase. This core is combined with either genes encoding a ribosomally synthesized peptide or genes catalyzing the formation of a modified amino acid. This latter set appears to be subjected to the greatest lineage-specific diversity (Fig. 4 and 5), probably in response to the selective pressure of resistance against the secondary metabolites. There is also evidence for larger scale recombination of secondary metabolite biosynthesis pathways. For example, the butirosin, friulimicin and vibrioferrin systems appear to have emerged from the coalescence of a simple peptide synthesis system based on the ATP-grasp domains (comparable to the antibiotic bacilysin) with multiple components from other distinct systems. In the butirosin system we have the confluence of 3 distinct modules: (1) classical aminoglycoside biosynthesis, (2) the BtrH-like transpeptidase from non-ribosomally synthesized peptide antibiotic biosynthesis pathways and (3) the ATP-grasp-dependent peptide biosynthesis system (Fig. 4 and 5). In friulimicin there is a combination of multidomain non-ribosomal peptide synthetases with the ATP-grasp centered system, whereas in cell-surface polymer biosynthesis there is a combination of polysaccharide biosynthesis systems with the ATP-grasp-based peptide synthesis system.

Origin of peptide tagging of proteins from peptide synthesis systems. Multiple, functionally diversified peptide tags, which are added to proteins, are universally found in eukaryotes. While most major eukaryotic peptide tags are traceable to the last eukaryotic common ancestor (LECA), their provenance has until recently been largely unclear. Comparable systems do not appear to be prevalent in archaea. While little is known of peptide tags in bacteria beyond the few well-characterized examples, several recent studies are making it clear that the precursors of the eukaryotic peptide tagging systems lie in the bacterial world.[8,10,12] The current study, taken together with earlier studies, presents the following consistent picture across different groups of peptide tags ligated by structurally unrelated folds of enzymes: the earliest representatives of these folds catalyzed reactions mechanistically related to peptide ligation, but in entirely distinct contexts, such as cofactor, amino acid and nucleotide metabolism. Subsequently, in bacterial evolution these ancient metabolic enzymes appear to have spawned a diversity of enzymes producing peptides and related amide-bond metabolites ranging from the pan-bacterial peptidoglycan to lineage-specific antibiotics and siderophores. From within this diversity of peptide and amide synthesis systems actual peptide tagging systems appear

to have independently emerged in at least four structurally distinct scaffolds, namely the GNATs, COOH–NH_2-ligases, ATP-grasp and the E1–E2 system. Identification of the alpha-E domain-associated system and the stramenopile YEATS domain-linked COOH–NH2-ligases in the current study suggests that there are multiple examples of such modifications that remain unexplored.

In this study we detected the first bacterial homologs of the eukaryotic TTL in a number of free-living bacteria (*e.g.* TK90DRAFT_2815, gi: 224818354 from *Thioalkalivibrio*, ESI), which in certain cases are fused to a 2-oxoglutarate-dependent dioxygenase related to prolyl hydroxylases. It is conceivable that these enzymes were modifying target proteins in bacteria both by peptide tags and oxidative modification of side chains (*i.e.* if fused to the dioxygenase). Such a combination of modifying activities is also seen in eukaryotes where we have previously reported fusions of the TTL domain to the SET protein methyltransferase domain[94] (Fig. 3). Thus, eukaryotic TTLs which catalyze a range of peptide-tagging reactions on proteins might have emerged from an ancestral bacterial version. Peptide tags added by TTLs are traceable to the last eukaryotic common ancestor (LECA) and are required for the assembly of quintessential eukaryotic structures such as the tubulin cytoskeleton and possibly also chromatin. Taken together these observations suggest that, along with ubiquitination, ATP-grasp-dependent modifications were acquired prior to LECA from the bacterial component perhaps in course of the symbiogenic origin of the eukaryotes. This, along with other previously reported observations,[95] strongly suggests that bacterial genetic contributions were behind the emergence of key features in quintessentially eukaryotic structures such as cytoskeleton and chromatin.

We were unable to obtain evidence for a functional linkage in bacteria between peptide-ligase-dependent peptide-tagging and ATP-dependent protein unfolding/degradation in systems other than pupylation and the N-end rule. Thus, the coupling of these systems probably evolved only on a few occasions in bacteria. Limited phyletic distribution of pupylation suggests that it possibly arose relatively recently from a system similar to the marinostatin/microviridin peptide modification system. Like these metabolites Pup is a small, largely disordered protein; however, instead of cyclization its associated ligase conjugates it to proteins. But the universality of peptide-tagging in targeting proteins for ATP-dependent unfolding and degradation in bacteria and eukaryotes is indicated by the presence, respectively, of tmRNA-based and the E1–E2-based systems in these two superkingdoms.[6,8,10] While there are a few archaeal peptide ligases which might be candidates for tagging of proteins, we did not find convincing evidence that any of them might be linked to protein degradation in archaea. This is strange since archaea do have robust ATP-dependent protein degradation systems (*e.g.* cognates of the eukaryotic proteasome). Hence, it is possible that they possess their own unidentified tagging system which might be RNA-dependent like the bacterial tmRNA system. Circumstantial support for this proposal is seen in the form of the previously observed linkage between the genes encoding proteasomal components and RNA-processing enzymes in archaea.[96]

General conclusions

While the there has been an explosion of genomic sequences from prokaryotes, there has not been a commensurate effort to understand the regulatory and metabolic novelties of most bacterial lineages. In particular, the potential of genomics in discovering novel natural products, which span a bewildering diversity from non-ribosomally synthesized storage poly-peptides to interesting low molecular weight secondary metabolites, has been under-utilized. The discovery of processes such as pupylation[13] and novel cofactor modification pathways[33] also highlights the diversity of regulatory mechanisms in non-model bacterial systems. Together with earlier studies on the prokaryotic antecedents of the Ub-system, we note that peptide/amide-bond forming ligase domains and their functional partners are a rich source of catalysts of the biochemical diversity generated by bacteria. While the gigantic multidomain peptide ligases and aminoglycoside biosynthesis pathways have been studied extensively as a potential source for new catalysts of antibiotic synthesis,[78,97] other peptide ligase systems have been less studied. In the current study we show that they define several novel pathways for secondary metabolism biosynthesis as well as possible regulatory pathways that involve peptide-tagging of target proteins. There are manifold ramifications of the findings presented here. Firstly, the general evolutionary principles related to the invention of peptide/amide-bond forming enzymes as well as peptide-tagging systems have been considerably clarified. Further, we uncover or clarify the biochemical mechanisms of certain poorly understood steps in synthesis of multiple antibiotics and cell-surface polymers. The data assembled here also serves as a repository for the experimental discovery of novel secondary metabolites and the biochemical engineering of antibiotics and related metabolites. Finally, presence of some of the systems uncovered in this study, for example the alpha-E domain containing ligases, in pathogens such as mycobacteria might help in directing experimental studies to better understand their pathogenesis.

Materials and methods

Structure similarity searches were conducted using the FSSP program,[98] and structural alignments were made using the MUSTANG program.[99] Protein structures were visualized and manipulated using the Swiss-PDB[100] and PyMol (http://pymol.sourceforge.net/) programs. Sequence profile searches were performed against the NCBI non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda, MD), and a locally compiled database of proteins from eukaryotes with completely or near-completely sequenced genomes. PSI-BLAST searches were performed using an expectation value (E-value) of 0.01 as the threshold for inclusion in the position-specific scoring matrix generated by the program;[101] searches were iterated until convergence. Profile-based HMM searches were performed using the newly released HMMER3 package (version beta 2).[102] Multiple alignments were constructed using the MUSCLE[103] and Kalign[104] programs, followed by manual correction based on

PSI-BLAST high-scoring pairs, secondary structure predictions, and information derived from existing structures. Protein secondary structure was predicted using a multiple alignment as the input for the JPRED2 program, which uses information extracted from a PSSM, HMM, and the seed alignment itself.[105] Pairwise comparisons of HMMs, using a single sequence or multiple alignment as query, against profiles of proteins in the PDB database were performed with the HHPRED program.[106] Similarity-based clustering was performed using the BLASTCLUST program [ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html] with empirically determined length and score threshold parameters. Gene neighborhoods in prokaryotes were obtained by isolating conserved genes immediately upstream and downstream of the gene in question showing separation of less than 70 nucleotides between gene termini. Neighborhoods were determined by searching NCBI PTT tables (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = Genome) with a custom PERL script. Phylogenetic analysis was carried out using neighborhood-joining and minimum evolution-based methods with gamma distributed rates and a JTT substitution matrix as implemented in the MEGA4 program.[107] The shape parameter α was estimated empirically through a series of experimental trials. Additionally maximum likelihood trees were obtained by first using the least-square method implemented in the FITCH program of the PHYLIP package[108] with subsequent local rearrangement using the PROTML program of the MOLPHY package.[109] All large-scale procedures were carried out using the TASS software package.[110]

Note added at proof

When this paper was being prepared for publication an uncharacterized representative of the COOH–NH2 ligase superfamily was characterized as the actin cross-linking enzymatic domain of the Vibrio MARTX toxins.[111] This activity is consistent with the predicted peptide/protein cross-linking activities of several representatives of this superfamily presented in this work.

Acknowledgements

L. M. I., S. A. and L. A. are supported by the intramural funds of the National Library of Medicine at the National Institutes of Health, USA.

References

1 A. Ciechanover, A. Orian and A. L. Schwartz, *BioEssays*, 2000, **22**, 442–451.
2 O. Kerscher, R. Felberbaum and M. Hochstrasser, *Annu. Rev. Cell Dev. Biol.*, 2006, **22**, 159–180.
3 A. Mogk, R. Schmidt and B. Bukau, *Trends Cell Biol.*, 2007, **17**, 165–172.
4 C. Janke, K. Rogowski, D. Wloga, C. Regnard, A. V. Kajava, J. M. Strub, N. Temurak, J. van Dijk, D. Boucher, A. van Dorsselaer, S. Suryavanshi, J. Gaertig and B. Edde, *Science (New York, N. Y.)*, 2005, **308**, 1758–1762.
5 J. van Dijk, K. Rogowski, J. Miro, B. Lacroix, B. Edde and C. Janke, *Mol. Cell*, 2007, **26**, 437–448.
6 K. C. Keiler, *Annu. Rev. Microbiol.*, 2008, **62**, 133–151.
7 W. K. Kang, T. Icho, S. Isono, M. Kitakawa and K. Isono, *MGG, Mol. Gen. Genet.*, 1989, **217**, 281–288.
8 L. M. Iyer, A. M. Burroughs and L. Aravind, *GenomeBiology*, 2006, **7**, R60.
9 K. E. Burns, S. Baumgart, P. C. Dorrestein, H. Zhai, F. W. McLafferty and T. P. Begley, *J. Am. Chem. Soc.*, 2005, **127**, 11602–11603.
10 A. M. Burroughs, L. M. Iyer and L. Aravind, *Proteins: Struct., Funct., Bioinf.*, 2009, **75**, 895–910.
11 A. Sauerwald, W. Zhu, T. A. Major, H. Roy, S. Palioura, D. Jahn, W. B. Whitman, J. R. Yates, 3rd, M. Ibba and D. Soll, *Science (New York, N. Y.)*, 2005, **307**, 1969–1972.
12 L. M. Iyer, A. M. Burroughs and L. Aravind, *Biol. Direct*, 2008, **3**, 45.
13 M. J. Pearce, J. Mintseris, J. Ferreyra, S. P. Gygi and K. H. Darwin, *Science (New York, N. Y.)*, 2008, **322**, 1104–1107.
14 L. M. Iyer, D. D. Leipe, E. V. Koonin and L. Aravind, *J. Struct. Biol.*, 2004, **146**, 11–31.
15 R. T. Sauer, D. N. Bolon, B. M. Burton, R. E. Burton, J. M. Flynn, R. A. Grant, G. L. Hersch, S. A. Joshi, J. A. Kenniston, I. Levchenko, S. B. Neher, E. S. Oakes, S. M. Siddiqui, D. A. Wah and T. A. Baker, *Cell*, 2004, **119**, 9–18.
16 J.-M. Peters, J. R. Harris and D. Finley, *Ubiquitin and the Biology of the Cell*, Plenum Press, New York, 1998.
17 K. Cadwell and L. Coscoy, *Science (New York, N. Y.)*, 2005, **309**, 127–130.
18 X. Dong, M. Kato-Murayama, T. Muramatsu, H. Mori, M. Shirouzu, Y. Bessho and S. Yokoyama, *Protein Sci.*, 2007, **16**, 528–534.
19 K. Suto, Y. Shimizu, K. Watanabe, T. Ueda, S. Fukai, O. Nureki and K. Tomita, *EMBO J.*, 2006, **25**, 5942–5950.
20 B. T. Dye and B. A. Schulman, *Annu. Rev. Biophys. Biomol. Struct.*, 2007, **36**, 131–150.
21 F. Striebel, F. Imkamp, M. Sutter, M. Steiner, A. Mamedov and E. Weber-Ban, *Nat. Struct. Mol. Biol.*, 2009.
22 K. Furukawa, N. Mizushima, T. Noda and Y. Ohsumi, *J. Biol. Chem.*, 2000, **275**, 7462–7465.
23 C. Lehmann, T. P. Begley and S. E. Ealick, *Biochemistry*, 2006, **45**, 11–19.
24 J. Xi, Y. Ge, C. Kinsland, F. W. McLafferty and T. P. Begley, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 8513–8518.
25 M. W. Lake, M. M. Wuebbens, K. V. Rajagopalan and H. Schindelin, *Nature*, 2001, **414**, 325–329.
26 M. J. Rudolph, M. M. Wuebbens, K. V. Rajagopalan and H. Schindelin, *Nat. Struct. Biol.*, 2001, **8**, 42–46.
27 A. M. Godert, M. Jin, F. W. McLafferty and T. P. Begley, *J. Bacteriol.*, 2007, **189**, 2941–2944.
28 S. F. Brady, C. J. Chao and J. Clardy, *Appl. Environ. Microbiol.*, 2004, **70**, 6865–6870.
29 E. V. Koonin and L. Aravind, *Curr. Biol.*, 1998, **8**, R266–269.
30 J. J. Abbott, J. Pei, J. L. Ford, Y. Qi, V. N. Grishin, L. A. Pitcher, M. A. Phillips and N. V. Grishin, *J. Biol. Chem.*, 2001, **276**, 42099–42107.
31 H. Oshikane, K. Sheppard, S. Fukai, Y. Nakamura, R. Ishitani, T. Numata, R. L. Sherrer, L. Feng, E. Schmitt, M. Panvert, S. Blanquet, Y. Mechulam, D. Soll and O. Nureki, *Science (New York, N. Y..)*, 2006, **312**, 1950–1954.
32 S. Kurihara, S. Oda, Y. Tsuboi, H. G. Kim, M. Oshida, H. Kumagai and H. Suzuki, *J. Biol. Chem.*, 2008, **283**, 19981–19990.
33 H. Li, H. Xu, D. E. Graham and R. H. White, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 9785–9790.
34 B. Nocek, E. Evdokimova, M. Proudfoot, M. Kudritska, L. L. Grochowski, R. H. White, A. Savchenko, A. F. Yakunin, A. Edwards and A. Joachimiak, *J. Mol. Biol.*, 2007, **372**, 456–469.
35 R. C. Fahey, *Annu. Rev. Microbiol.*, 2001, **55**, 333–356.
36 E. Aboulmagd, F. B. Oppermann-Sanio and A. Steinbuchel, *Arch. Microbiol.*, 2000, **174**, 297–306.
37 K. Tabata, H. Ikeda and S. Hashimoto, *J. Bacteriol.*, 2005, **187**, 5195–5202.
38 N. M. Llewellyn, Y. Li and J. B. Spencer, *Chem. Biol.*, 2007, **14**, 379–386.
39 T. Tanabe, T. Funahashi, H. Nakao, S. Miyoshi, S. Shinoda and S. Yamamoto, *J. Bacteriol.*, 2003, **185**, 6938–6949.

40 S. Liu, J. S. Chang, J. T. Herberg, M. M. Horng, P. K. Tomich, A. H. Lin and K. R. Marotti, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 15178–15183.
41 N. Ziemert, K. Ishida, A. Liaimer, C. Hertweck and E. Dittmann, *Angew. Chem., Int. Ed.*, 2008, **47**, 7756–7759.
42 B. Philmus, G. Christiansen, W. Y. Yoshida and T. K. Hemscheidt, *ChemBioChem*, 2008, **9**, 3066–3073.
43 T. E. Benson, D. B. Prince, V. T. Mutchler, K. A. Curry, A. M. Ho, R. W. Sarver, J. C. Hagadorn, G. H. Choi and R. L. Garlick, *Structure*, 2002, **10**, 1107–1115.
44 S. Balaji and L. Aravind, *Nucleic Acids Res.*, 2007, **35**, 5658–5671.
45 M. Y. Galperin and E. V. Koonin, *Protein Sci.*, 1997, **6**, 2639–2643.
46 N. V. Grishin, *J. Mol. Biol.*, 1999, **291**, 239–247.
47 Y. I. Wolf, I. B. Rogozin, A. S. Kondrashov and E. V. Koonin, *Genome Res.*, 2001, **11**, 356–372.
48 M. Huynen, B. Snel, W. Lathe, 3rd and P. Bork, *Genome Res.*, 2000, **10**, 1204–1210.
49 R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch and N. Maltsev, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 2896–2901.
50 V. Anantharaman and L. Aravind, *GenomeBiology*, 2003, **4**, R11.
51 B. E. Janowiak and O. W. Griffith, *J. Biol. Chem.*, 2005, **280**, 11829–11839.
52 D. D. Leipe, E. V. Koonin and L. Aravind, *J. Mol. Biol.*, 2003, **333**, 781–815.
53 G. Füser and A. Steinbuchel, *Macromol. Biosci.*, 2007, **7**, 278–296.
54 J. B. Thoden, H. M. Holden, G. Wesenberg, F. M. Raushel and I. Rayment, *Biochemistry*, 1997, **36**, 6305–6316.
55 L. Holm and C. Sander, *Proteins: Struct., Funct., Genet.*, 1997, **28**, 72–82.
56 I. A. Lessard and C. T. Walsh, *Chem. Biol.*, 1999, **6**, 177–187.
57 D. S. Kwon, C. H. Lin, S. Chen, J. K. Coward, C. T. Walsh and J. M. Bollinger, Jr, *J. Biol. Chem.*, 1997, **272**, 2429–2436.
58 A. W. Curnow, K. Hong, R. Yuan, S. Kim, O. Martins, W. Winkler, T. M. Henkin and D. Soll, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 11819–11826.
59 R. E. Valas and P. E. Bourne, *J. Mol. Evol.*, 2008, **66**, 494–504.
60 A. B. Hervas, I. Canosa and E. Santero, *J. Bacteriol.*, 2008, **190**, 416–420.
61 S. E. Chuang, V. Burland, G. Plunkett, 3rd, D. L. Daniels and F. R. Blattner, *Gene*, 1993, **134**, 1–6.
62 G. Steinborn, M. R. Hajirezaei and J. Hofemeister, *Arch. Microbiol.*, 2005, **183**, 71–79.
63 C. van Ooij, P. Eichenberger and R. Losick, *J. Bacteriol.*, 2004, **186**, 4441–4448.
64 T. Candela and A. Fouet, *Mol. Microbiol.*, 2006, **60**, 1091–1098.
65 M. Rawat and Y. Av-Gay, *FEMS Microbiol. Rev.*, 2007, **31**, 278–292.
66 I. Le Masson, D. Y. Yu, K. Jensen, A. Chevalier, R. Courbeyrette, Y. Boulard, M. M. Smith and C. Mann, *Mol. Cell. Biol.*, 2003, **23**, 6086–6102.
67 S. Kurihara, S. Oda, K. Kato, H. G. Kim, T. Koyanagi, H. Kumagai and H. Suzuki, *J. Biol. Chem.*, 2004, **280**, 4602–4608.
68 H. U. Rexer, T. Schaberle, W. Wohlleben and A. Engels, *Arch. Microbiol.*, 2006, **186**, 447–458.
69 B. N. Lee and T. H. Adams, *EMBO J.*, 1996, **15**, 299–309.
70 M. H. Saier, Jr, *Microbiology*, 2000, **146**(Pt 8), 1775–1795.
71 B. L. Carlson, E. R. Ballister, E. Skordalakes, D. S. King, M. A. Breidenbach, S. A. Gilmore, J. M. Berger and C. R. Bertozzi, *J. Biol. Chem.*, 2008, **283**, 20117–20125.
72 K. S. Makarova, L. Aravind, Y. I. Wolf, R. L. Tatusov, K. W. Minton, E. V. Koonin and M. J. Daly, *Microbiol. Mol. Biol. Rev.*, 2001, **65**, 44–79.
73 C. Muller, S. Nolden, P. Gebhardt, E. Heinzelmann, C. Lange, O. Puk, K. Welzel, W. Wohlleben and D. Schwartz, *Antimicrob. Agents Chemother.*, 2007, **51**, 1028–1037.
74 B. A. Manjasetty, J. Powlowski and A. Vrielink, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 6992–6997.
75 O. T. Magnusson, H. Toyama, M. Saeki, A. Rojas, J. C. Reed, R. C. Liddington, J. P. Klinman and R. Schwarzenbacher, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 7913–7918.
76 R. D. Woodyer, Z. Shao, P. M. Thomas, N. L. Kelleher, J. A. Blodgett, W. W. Metcalf, W. A. van der Donk and H. Zhao, *Chem. Biol.*, 2006, **13**, 1171–1182.
77 F. Kudo and T. Eguchi, *Methods Enzymol.*, 2009, **459**, 493–519.
78 U. F. Wehmeier and W. Piepersberg, *Methods Enzymol.*, 2009, **459**, 459–491.
79 P. D. Fortin, C. T. Walsh and N. A. Magarvey, *Nature*, 2007, **448**, 824–827.
80 S. Dumitriu, *Polysaccharides: Structural Diversity and Functional Versatility*, Marcel Dekker, New York, 1998.
81 R. Aono, *Biochem. J.*, 1990, **270**, 363–367.
82 R. Aono, M. Ito and T. Machida, *J. Bacteriol.*, 1999, **181**, 6600–6606.
83 B. Liu, Y. A. Knirel, L. Feng, A. V. Perepelov, S. N. Senchenkova, Q. Wang, P. R. Reeves and L. Wang, *FEMS Microbiol. Rev.*, 2008, **32**, 627–653.
84 S. D. Bentley, D. M. Aanensen, A. Mavroidi, D. Saunders, E. Rabbinowitsch, M. Collins, K. Donohoe, D. Harris, L. Murphy, M. A. Quail, G. Samuel, I. C. Skovsted, M. S. Kaltoft, B. Barrell, P. R. Reeves, J. Parkhill and B. G. Spratt, *PLoS Genet.*, 2006, **2**, e31.
85 N. A. Kocharova, S. N. Senchenkova, A. N. Kondakova, A. I. Gremyakov, G. V. Zatonsky, A. S. Shashkov, Y. A. Knirel and N. K. Kochetkov, *Biochemistry (Moscow)*, 2004, **69**, 103–107.
86 A. N. Kondakova, F. V. Toukach, S. N. Senchenkova, N. P. Arbatsky, A. S. Shashkov, Y. A. Knirel, B. Bartodziejska, K. Zych, A. Rozalski and Z. Sidorczyk, *Biochemistry (Moscow)*, 2003, **68**, 446–457.
87 G. L. Challis, *ChemBioChem*, 2005, **6**, 601–611.
88 S. Schmelz, N. Kadi, S. A. McMahon, L. Song, D. Oves-Costales, M. Oke, H. Liu, K. A. Johnson, L. G. Carter, C. H. Botting, M. F. White, G. L. Challis and J. H. Naismith, *Nat. Chem. Biol.*, 2009, **5**, 174–182.
89 S. Chen, A. F. Yakunin, M. Proudfoot, R. Kim and S. H. Kim, *Proteins: Struct., Funct., Bioinf.*, 2005, **61**, 433–443.
90 J. A. Doudna and T. R. Cech, *Nature*, 2002, **418**, 222–228.
91 K. A. Denessiouk, J. V. Lehtonen, T. Korpela and M. S. Johnson, *Protein Sci.*, 1998, **7**, 1136–1146.
92 F. Dyda, D. C. Klein and A. B. Hickman, *Annu. Rev. Biophys. Biomol. Struct.*, 2000, **29**, 81–103.
93 T. A. Keating, C. G. Marshall, C. T. Walsh and A. E. Keating, *Nat. Struct. Biol.*, 2002, **9**, 522–526.
94 L. M. Iyer, V. Anantharaman, M. Y. Wolf and L. Aravind, *Int. J. Parasitol.*, 2008, **38**, 1–31.
95 L. Aravind, L. M. Iyer and E. V. Koonin, *Curr. Opin. Struct. Biol.*, 2006, **16**, 409–419.
96 E. V. Koonin, Y. I. Wolf and L. Aravind, *Genome Res.*, 2001, **11**, 240–252.
97 S. A. Samel, M. A. Marahiel and L. O. Essen, *Mol. BioSyst.*, 2008, **4**, 387–393.
98 L. Holm and C. Sander, *Nucleic Acids Res.*, 1998, **26**, 316–319.
99 A. S. Konagurthu, J. C. Whisstock, P. J. Stuckey and A. M. Lesk, *Proteins: Struct., Funct., Bioinf.*, 2006, **64**, 559–574.
100 N. Guex and M. C. Peitsch, *Electrophoresis*, 1997, **18**, 2714–2723.
101 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
102 S. R. Eddy, *Bioinformatics*, 1998, **14**, 755–763.
103 R. C. Edgar, *Nucleic Acids Res.*, 2004, **32**, 1792–1797.
104 T. Lassmann, O. Frings and E. L. Sonnhammer, *Nucleic Acids Res.*, 2009, **37**, 858–865.
105 J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay and G. J. Barton, *Bioinformatics*, 1998, **14**, 892–893.
106 J. Soding, A. Biegert and A. N. Lupas, *Nucleic Acids Res.*, 2005, **33**, W244–248.
107 K. Tamura, J. Dudley, M. Nei and S. Kumar, *Mol. Biol. Evol.*, 2007, **24**, 1596–1599.
108 J. Felsenstein, *Cladistics*, 1989, **5**, 164–166.
109 M. Hasegawa, H. Kishino and N. Saitou, *J. Mol. Evol.*, 1991, **32**, 443–445.
110 V. Anantharaman, S. Balaji and L. Aravind, unpublished results.
111 B. Gessler, A. Bonebrake, K. L. Sheahan, M. E. Walker and K. J. Satchell, *Mol. Microbiol.*, 2009, **73**, 858–868.

