Research

New connections in the prokaryotic toxin-antitoxin network: relationship with the eukaryotic nonsense-mediated RNA decay system

Vivek Anantharaman and L Aravind

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

Correspondence: L Aravind. E-mail: aravind@ncbi.nlm.nih.gov

Published: 26 November 2003

Genome ***Biology*** 2003, **4:**R81

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/12/R81

Received: 21 August 2003
Revised: 13 October 2003
Accepted: 10 October 2003

Abstract

Background: Several prokaryotic plasmids maintain themselves in their hosts by means of diverse post-segregational cell killing systems. Recent findings suggest that chromosomally encoded copies of toxins and antitoxins of post-segregational cell killing systems - such as the RelE system - might function as regulatory switches under stress conditions. The RelE toxin cleaves ribosome-associated transcripts, whereas another post-segregational cell killing toxin, ParE, functions as a gyrase inhibitor.

Results: Using sequence profile analysis we were able unify the RelE- and ParE-type toxins with several families of small, uncharacterized proteins from diverse bacteria and archaea into a single superfamily. Gene neighborhood analysis showed that the majority of these proteins were encoded by genes in characteristic neighborhoods, in which genes encoding toxins always co-occurred with genes encoding transcription factors that are also antitoxins. The transcription factors accompanying the RelE/ParE superfamily may belong to unrelated or distantly related superfamilies, however. We used this conserved neighborhood template to transitively search genomes and identify novel post-segregational cell killing-related systems. One of these novel systems, observed in several prokaryotes, contained a predicted toxin with a PilT-N terminal (PIN) domain, which is also found in proteins of the eukaryotic nonsense-mediated RNA decay system. These searches also identified novel transcription factors (antitoxins) in post-segregational cell killing systems. Furthermore, the toxin Doc defines a potential metalloenzyme superfamily, with novel representatives in bacteria, archaea and eukaryotes, that probably acts on nucleic acids.

Conclusions: The tightly maintained gene neighborhoods of post-segregational cell killing-related systems appear to have evolved by *in situ* displacement of genes for toxins or antitoxins by functionally equivalent but evolutionarily unrelated genes. We predict that the novel post-segregational cell killing-related systems containing a PilT-N terminal domain toxin and the eukaryotic nonsense-mediated RNA decay system are likely to function via a common mechanism, in which the PilT-N terminal domain cleaves ribosome-associated transcripts. The core of the eukaryotic nonsense-mediated RNA decay system has probably evolved from a post-segregational cell killing-related system.

Background

Post-segregational cell killing (PSK) is a widespread mechanism that aids several plasmids to maintain themselves in their bacterial hosts [1-4]. Operons containing genes for interacting toxin-antitoxin (T-A) pairs that are borne on these plasmids, are the basis for PSK. Typically, the first gene in these operons encodes a labile antitoxin, which also acts as a transcriptional regulator of the operon, while the second gene encodes a stable toxin. Usually, the antitoxin forms a physical complex with the toxin and neutralizes its action. A variation on this theme is seen in the form of the unstable anti-sense RNAs, which act as inhibitors of translation of the toxin mRNAs. If the plasmid is lost, the antitoxin is rapidly degraded while the stable toxin lingers on, killing cells that lack the plasmid. Thus, plasmids with systems for PSK cause their host cells to become addicted to them [1-4]. Additionally, several of these T-A systems are also found on prokaryotic chromosomes, where they may have alternative regulatory functions [5].

A systematic survey of such T-A operons and their mechanisms was presented in the seminal work of Gerdes in 2000 [6]. Subsequently, there have also been some important studies that have elucidated the biochemical details regarding the action of several toxins. One of these toxins, ParE, was shown to act as an inhibitor of the DNA gyrase, and it induced formation of DNA-gyrase covalent complexes, which could inhibit replication and damage the integrity of the chromosome [7]. In contrast, the RelE and Doc toxins were shown to be inhibitors of translation [5,8]. More recently, it was demonstrated that the RelE protein cleaved transcripts associated with the ribosome, by specifically targeting codons associated with the ribosomal A-site [9]. RelE displays codon-specificity by showing highest preference for UAG among the stop codons and UCG and CAG among the sense codons [9]. Interestingly, this inhibition of translation by RelE is reversed by the transfer-messenger RNA (tmRNA), which acts as a regulator of protein stability in bacteria [10]. These studies have also suggested that the chromosomal versions of these antitoxin-toxin pairs could function as regulatory switches that control gene expression under poor growth conditions.

Although Gerdes proposed that all T-A operons could have a common origin [6], an objective evaluation of the evolutionary relationships of these proteins and the origin of these systems has not been conducted. The availability of a large number of prokaryotic genome sequences allows us to use a variety of computational approaches to address the problem of the origin and evolution of these systems. One approach, involving sensitive sequence searches using profile methods, allows the detection of distant relationships, which were hitherto not detected [11-13]. Additionally, it also enables objective evaluation of relationships, based on statistical significance of the detected similarities and multiple alignment-derived secondary structure predictions. A second approach involves the use of comparative genomics to detect conserved gene neighborhoods, gene or domain fusions, and to extract functional and evolutionary information from these contextual connections [14-18]. This approach is particular useful in the case of the prokaryotic PSK systems because of the strong coupling of the toxin and antitoxin genes in a single operon. Our objective in applying these analyses was to discover new functional connections that may not have been previously uncovered in experimental studies on these systems. Given the recent experimental results suggesting a specific role for these systems in the regulation of cellular responses to stress [9,10,19], we were also interested in identifying novel genomic versions of PSK-related systems with a wide phyletic distribution.

As a result of our analyses we were able to uncover several new T-A systems and establish an evolutionary relationship between them and the eukaryotic nonsense-mediated RNA degradation system. We also present evidence that the RelE and ParE families of toxins, despite their very distinct modes of action, have been ultimately derived from a common ancestor. Furthermore, we show that the Doc toxin defines a large family of enzymes that could potentially act on RNA and function as regulators of translation in both prokaryotes and eukaryotes.

Results and discussion

Unification of the RelE and ParE families and identification of new related families of proteins

As *Escherichia coli* RelE and its close relatives are amongst the functionally best-characterized toxins of the PSK systems, with a wide phyletic pattern in bacteria and archaea [6], we chose them as the starting point of our investigation of the general cellular functions and natural history of these systems. In order to determine the deep evolutionary affinities of the RelE proteins, we initiated a sequence profile search of the non-redundant (NR) protein database (National Center for Biotechnology Information, Bethesda, USA) using the PSI-BLAST program (threshold for inclusion in profile = 0.01, iterated till convergence) [11]. At convergence, this search recovered a large number of homologs of RelE-including all the previously described versions - from a variety of bacteria and archaea. We selected distinct representatives from the newly-detected members and transitively searched the NR database with these proteins as queries. As these proteins are typically small (85-110 residues in length) and divergent, several searches initiated with different seed sequences were required to exhaustively identify distant homologs of RelE. For example, RelE (gi: 16129522, *E. coli*) recovers a *Staphylococcus aureus* protein (gi: 15925446, ortholog of *E. coli* YoeB) in the third iteration (e= 6e-04), a *Campylobacter fetus* protein (gi: 28974229, ortholog of *E. coli* YafQ) in the fourth iteration (e = 2e-04), a *Microbulbifer degradans* protein (gi: 23028223, ParE family) in the fourth iteration (e = 0.004) and a *Magnetococcus* protein (gi: 23001539, with the RelE-related segment fused to a SF-I helicase module) in the

fifth iteration (e = 0.001). To further ensure the detection of highly divergent members, all unique members detected in these searches were included in a single PSI-BLAST PSSM that was used to iteratively search the NR database till convergence. As result of this procedure, we were able to recover over 150 distinct homologs (less than 92% identical) of RelE. Reciprocal searches started with diverse proteins detected in the above procedure recovered a common set of obvious RelE-related 'intermediate' sequences supporting these relationships. For example, a reciprocal search with a protein from *Bacteroides thetaiotaomicron* (gi: 29350140), which is consistently recovered from various starting sequences that were detected in the above searches, recovers other divergent RelE-related proteins (for example, *Nostoc punctiforme* protein gi: 23129164) in the third iteration (e= 0.001) and the *E. coli* RelE itself in the fifth iteration (e = 3e-06). These sequences were then clustered using the BLASTCLUST program and individual clusters were aligned using the T_coffee program [20]. These alignments were used to predict individually the secondary structure for each of these clusters with the PHD program [21]. A very similar arrangement of the predicted secondary structure elements between diverse groups of these proteins further reinforced their relationships.

A striking aspect of these searches was the establishment of the relationship between the ParE (typified by the plasmid RK2-encoded toxin, ParE) [22] and RelE families of toxins that were previously believed to be unrelated. These toxins have very different targets of action: ParE acts at the level of DNA replication and recombination by interfering with the action of gyrase [7], whereas RelE acts on RNA at the level of translation [5]. This observation suggested that despite a common origin and significant sequence similarity, these PSK toxins could have diverged into different functional roles. Hereinafter, we refer to this superfamily of proteins, which includes the toxin families defined by RelE, ParE and other evolutionarily-related proteins that were detected in the above searches, as the RelE/ParE superfamily. The majority of proteins in this superfamily are of similar length and appear to fold into a single globular domain.

A multiple sequence alignment of the entire RelE/ParE superfamily (Figure 1) was constructed by combining the alignments of the individual clusters using the Profile Consistency Multiple Sequence Alignment (PCMA) program and refining it based on PSI-BLAST pair-wise alignments and secondary structure predictions. The predicted secondary structure which is conserved throughout this superfamily defines an $\alpha + \beta$ fold with a single amino-terminal strand, followed by a bi-helical hairpin and at least three strong strands at the carboxyl terminus. This secondary structure pattern does not appear to be consistent with that of the MazF/Kid/CcdB superfamily of toxins [23-26], which adopts a SH3 barrel fold. Furthermore, no statistically significant relationship can be established between the profiles of the MazF/Kid/CcdB superfamily toxins and the RelE/ParE superfamily. Hence, even though both CcdB and ParE function as gyrase inhibitors, they are likely to fold into very distinct three-dimensional structures.

The multiple alignment of the RelE/ParE family shows that much of the conservation is associated with the residues forming the core of the conserved, predicted secondary structure elements (Figure 1). Two charged or polar residues, one associated with the first conserved helix and the second associated with the end of the carboxy-terminal-most strand, are also strongly conserved throughout the superfamily. A third, slightly less conserved polar residue is also seen to be associated with the second universally predicted strand of these proteins. This conservation of a charged residue is consistent with the nucleic acid-associated role of the functionally characterized proteins of this family, and could mediate interactions with RNA or DNA. However, beyond this general similarity, the ParE and RelE proteins have very different modes of action. Experimental studies have suggested that ParE inhibits the gyrase by trapping it with DNA in a stable complex, but so far there has been no report of any catalytic activity in ParE. In contrast, RelE and its homologs have been shown to cleave mRNA only when it is associated with the ribosome, but not free mRNAs [5,9]. This suggests that certain members of this superfamily may possess catalytic activity under certain circumstances, and the conserved polar residues could contribute to this activity. In particular, the charged residue, which occurs at the carboxyl terminus of the last strand in these proteins, is an attractive candidate for a potential catalytic residue in the RelE proteins. In light of the relationship between the ParE and RelE families of proteins it would be of some interest to investigate the possibility of an unexplored DNA-cleaving activity in members of the ParE family, analogous to the ribosome-associated RNAse activity of RelE.

We then investigated the evolutionary history of the RelE/ParE superfamily by exploring its phyletic and phylogenetic diversity. The superfamily is widely distributed in the currently sequenced prokaryotes: at least a single member is encoded by the chromosome or one of the large genomic partitions in several bacterial and most archaeal lineages (Figure 2). Additionally, plasmids, particularly those from proteobacteria, encode their own RelE/ParE-related proteins. However, no members of this superfamily could be detected in eukaryotes. This phyletic pattern could mean that the superfamily had its origin early the evolution of one of the prokaryotic lineages, followed by dissemination via plasmids. However, it is also possible that at least one representative of this superfamily was present in the last universal common ancestor and secondarily lost in the eukaryotes.

We determined the major lineages of the RelE/ParE superfamily through single-linkage clustering of the proteins with the BLASTCLUST program and construction of neighbor-joining phylogenetic trees with a multiple alignment of all complete members of the superfamily. Several distinct

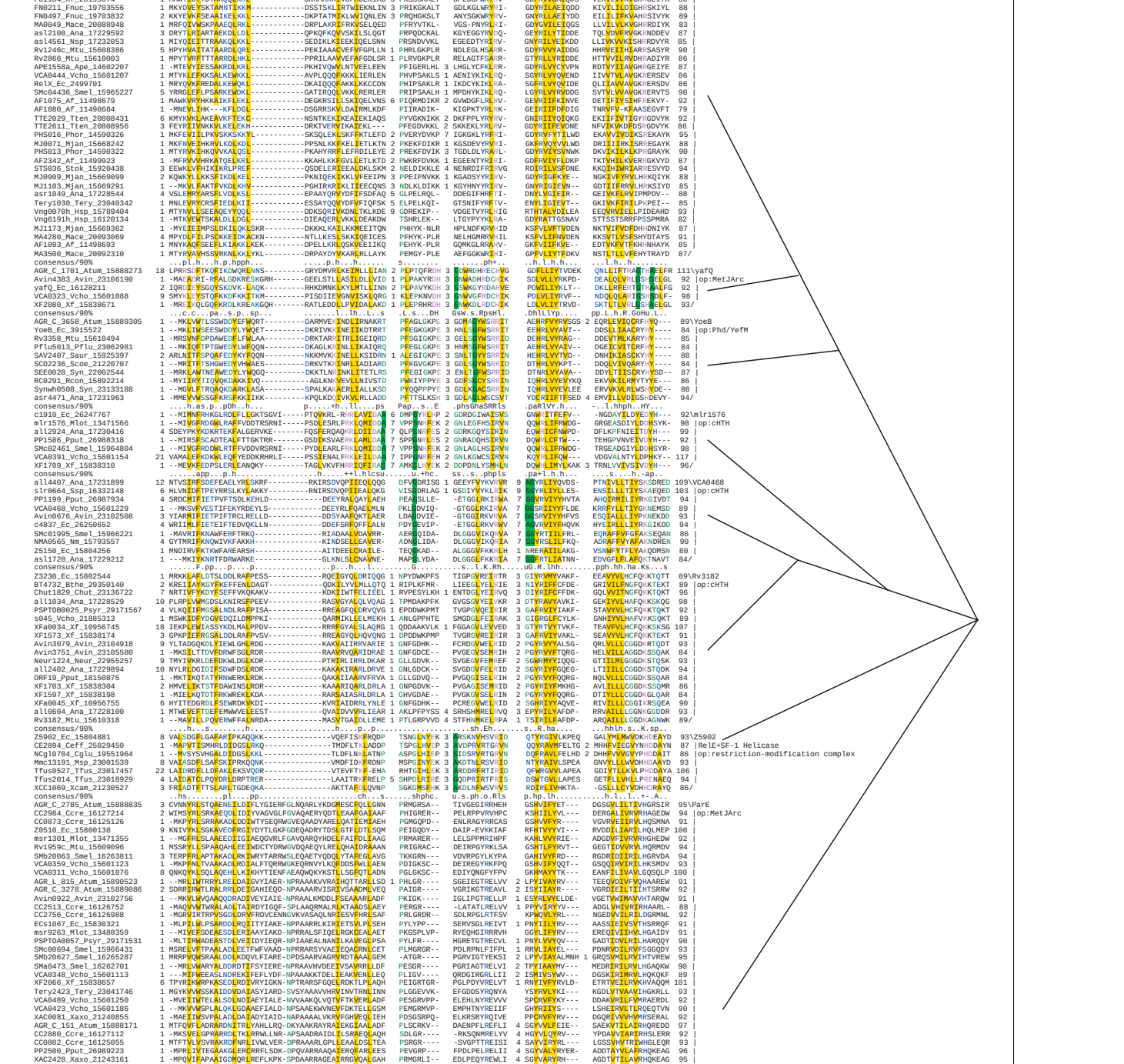

Figure 1 *(see legend on next page)*

Figure 1 *(see previous page)*
Multiple alignment of the RelE/ParE superfamily. Multiple sequence alignments of the different families of RelE/ParE were constructed using T-Coffee [20] and PCMA [50] after parsing high-scoring pairs from PSI-BLAST search results. The PHD-secondary structure [21] is shown above the alignment with E representing a β strand, and H an α-helix. The consensus of the individual families and the entire superfamily is shown, and the names of each family are shown on the right. The 90% (or 80%) consensus shown below the alignment was derived using the following amino acid classes: hydrophobic (h: ALICVMYFW, yellow shading); the aliphatic subset of the hydrophobic class (l: ALIVMC, yellow shading); aromatic (a: FHWY, yellow shading); small (s: ACDGNPSTV, green); the tiny subset of the small class (u: GAS, green shading); polar (p: CDEHKNQRST, blue); alcohol subset of polar (o: ST, blue); charged subset of polar (c: DEHKR, pink); positive subset of polar (+: HKR, pink); and negative subset of polar (-: DE, pink). An amino acid in capitals like 'G', or 'E' shows the completely conserved amino acid in that group. The operon information (op) and/or the domain architecture information are shown on the right for each family. The limits of the domains are indicated by the residue positions, in bold, on each side. The numbers within the alignment are non-conserved inserts that have not been shown. The sequences are denoted by their gene name followed by the species abbreviation and GenBank Identifier. The phylogenetic relationship between the families is shown as a tree to the right. The species abbreviations are: Af, *Archaeoglobus fulgidus*; Ape, *Aeropyrum pernix*; Hsp, *Halobacterium* sp.; Mace, *Methanosarcina acetivorans*; Mjan, *Methanocaldococcus jannaschii*; Phor, *Pyrococcus horikoshii*; Stok, *Sulfolobus tokodaii*; Ana, *Anabaena* sp.; Atum, *Agrobacterium tumefaciens*; Avin, *Azotobacter vinelandii*; Bthe, *Bacteroides thetaiotaomicron*; Ccre, *Caulobacter crescentus*; Ceff, *Corynebacterium efficiens*; Cglu, *Corynebacterium glutamicum*; Chut, *Cytophaga hutchinsonii*; Ec, *Escherichia coli*; Fnuc, *Fusobacterium nucleatum*; Mlot, *Mesorhizobium loti*; Mmag, *Magnetospirillum magnetotacticum*; Msp, *Magnetococcus* sp.; Mtu, *Mycobacterium tuberculosis*; Neur, *Nitrosomonas europaea*; Nm, *Neisseria meningitidis*; Paer, *Pseudomonas aeruginosa*; Pflu, *Pseudomonas fluorescens*; Pput, *Pseudomonas putida*; Psyr, *Pseudomonas syringae*; Rcon, *Rickettsia conorii*; Saur, *Staphylococcus aureus*; Scoe, *Streptomyces coelicolor*; Smel, *Sinorhizobium meliloti*; Ssp, *Synechocystis* sp.; Syn, *Synechococcus* sp.; Tery, *Trichodesmium erythraeum*; Tfus, *Thermobifida fusca*; Tten, *Thermoanaerobacter tengcongensis*; Vcho, *Vibrio cholerae*; Xaxo, *Xanthomonas axonopodis*; Xcam, *Xanthomonas campestris*; Xf, *Xylella fastidiosa*.

families could be delineated within the superfamily, of which two of the largest families were the RelE and ParE families. Most of the families could be distinguished by means of certain lineage-specific conserved residues (Figure 1). As previously noted [6], the RelE family had the widest phyletic spread with members in several bacterial and archaeal lineages. A small, proteobacteria-specific family, typified by the YafQ protein of *E. coli* (YafQ family), is the one that is most closely related to the RelE family. These two families are unified by the presence of shared polar residue at the beginning of strand 1 (Figure 1). The ParE family is restricted to bacteria, but is widely distributed with representatives in proteobacteria, cyanobacteria, actinomycetes and cytophagales. The ParE family is distinguished by the presence of a single polar residue in the second conserved strand, whereas all other members of the RelE/ParE superfamily possess two conserved polar residues in this strand (Figure 1). Two of the remaining families, one typified by the protein Rv3182 from *Mycobacterium tuberculosis* (Rv3182 family) and one defined by the YoeB protein of *E. coli* (YoeB family), are fairly widespread across a range of bacterial lineages (Figure 1) and are primarily encoded by the main chromosome. The remaining smaller families are far more sporadic in their distribution, and chiefly occur only in proteobacteria and cyanobacteria (Figure 1). One of the most divergent families of the ParE-RelE superfamily is typified by the Z5902 protein (Z5902 family) from the enterohemorrhagic strain of *E. coli* (O157:H7), and has sporadic representatives from a number of unrelated bacteria such as *Magnetobacterium*, *Corynebacterium* and *Thermobifida*. All members of this family occur fused to a carboxy-terminal superfamily I (SF-I) helicase module, and represent one of the rare instances when the ParE-RelE domain occurs in a multidomain protein (Figure 3).

Wider phyletic spread of the RelE family and its relatives, as compared to the ParE family, may suggest that the former group represents the more ancient member of the superfamily, with the ParE lineage being secondarily derived in bacteria. This would imply that the RNA-cleaving activity is likely to be the primitive function of this superfamily, with a secondary innovation of gyrase inhibitor activity in the ParE family. The sporadic, but widespread phyletic patterns of several families, and differences in representation between strains of the same species (for example, *E. coli*), suggest a potential role for lateral transfer in the spread of these genes. At the same time, the extensive occurrence of genes for this superfamily in the chromosomal partitions of the genomes, and not merely on plasmids, supports the proposal that they may be widely used as cellular regulators. Thus, the acquisition of members of the RelE/ParE superfamily through lateral transfer could be a means by which certain strains could rapidly evolve a new regulatory pathway that helps in adapting their gene expression to unique environmental stresses.

Gene-neighborhood analysis of the RelE/ParE superfamily and identification of PSK-like systems encoding PilT-N terminal (PIN) domain proteins

Given the tight coupling of the toxin-antitoxin gene pairs, we investigated contextual information derived from their gene neighborhoods [14-18]. We concentrated on the newly identified members of the RelE/ParE superfamily to glean previously unknown contextual connections to other genes. Upstream genes encoding transcription factors of the MetJ/Arc superfamily accompany both RelE and ParE families [6,27,28]. This transcription factor serves as the antitoxin, which not only regulates the transcription of genes in the T-A operon, but also physically binds to the toxins and counters their actions [6]. A systematic survey of all the newly identified members of the RelE, YafQ and ParE families showed that the majority of the genes encoding these proteins were associated with upstream genes for MetJ/Arc transcription factors (Figures 1,3). In contrast, a range of novel gene neighborhood associations was observed in several of the newly identified families of the RelE/ParE superfamily.

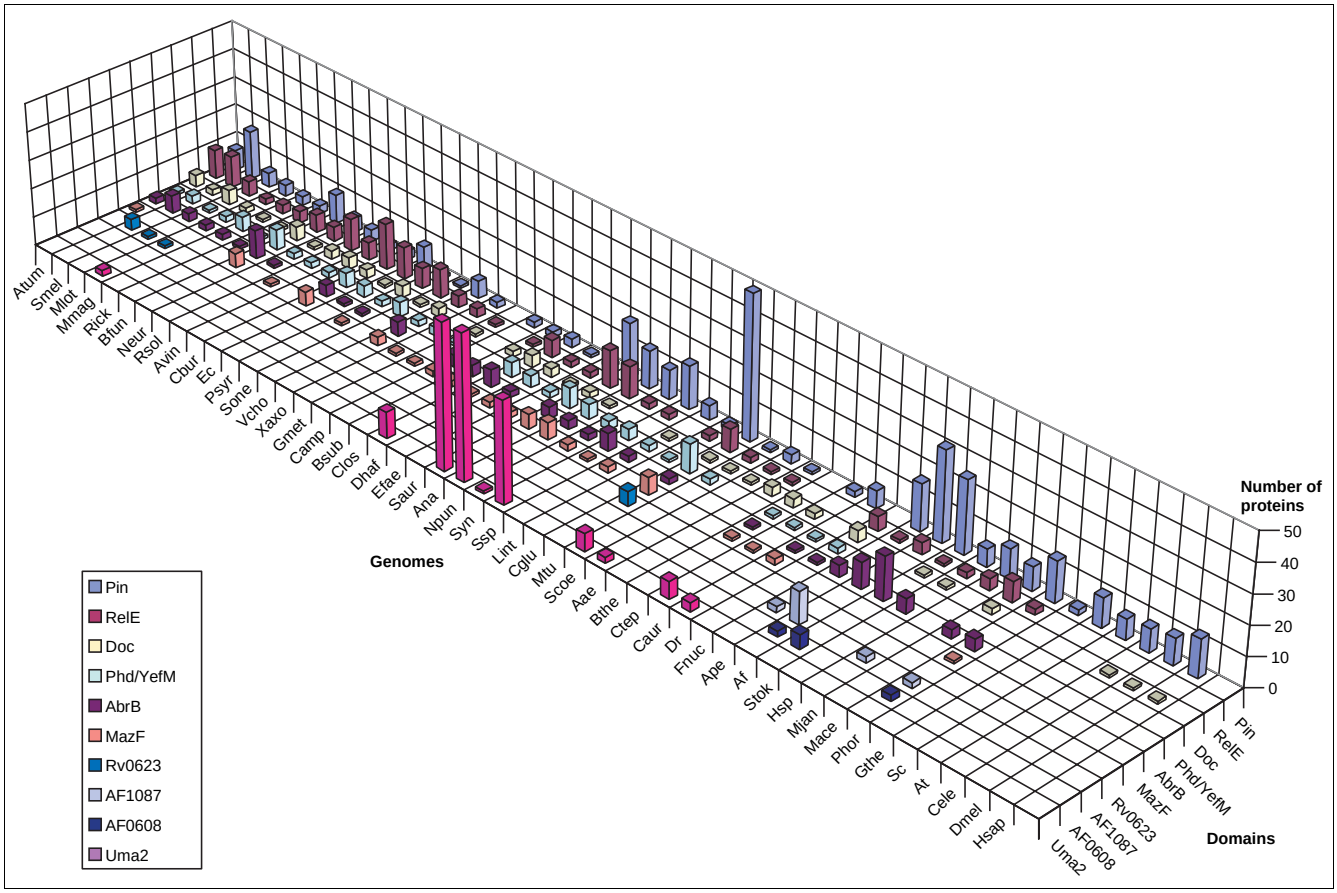

Figure 2

Relative abundance of some major families of toxins, associated transcription factors (antitoxins) and the UMA2 superfamily in various genomes. The number of proteins containing PIN, RelE/ParE, Doc, Phd/YefM, AbrB, MazF/CcdB/KiD, Rv0623, AF0319 and AF0608 domains in different genomes is indicated for each genome. The species abbreviations are as shown in Figure 1 and additionally: Aae, *Aquifex aeolicus*; Bfun, *Burkholderia fungorum*; Bsub, *Bacillus subtilis*; Camp, *Campylobacter*; Caur, *Chloroflexus aurantiacus*; Cbur, *Coxiella burnetii*; Clos, *Clostridium*; Ctep, *Chlorobium tepidum*; Dhaf, *Desulfitobacterium hafniense*; Dr, *Deinococcus radiodurans*; Efae, *Enterococcus faecalis*; Gmet, *Geobacter metallireducens*; Lint, *Leptospira interrogans*; Npun, *Nostoc punctiforme*; Rick, *Rickettsia*; Rsol, *Ralstonia solanacearum*; Sone, *Shewanella oneidensis*; Gthe, *Guillardia theta*; Sc, *Saccharomyces cerevisiae*; At, *Arabidopsis thaliana*; Cele, *Caenorhabditis elegans*; Dmel, *Drosophila melanogaster*; Hsap, *Homo sapiens*.

Genes for proteins belonging to the YoeB family of the RelE/ParE superfamily were consistently associated with upstream genes that coded small proteins (~75-90 residues) that were unrelated to the MetJ/Arc superfamily. We investigated this family of small proteins further by initiating iterative PSI-BLAST searches seeded with the *E. coli* YefM protein, which is their archetypal representative. These searches showed that they formed a group of bacterial and phage proteins that included the previous characterized DNA-binding proteins, like Phd from phage P1 and DnaT [29,30]. Reciprocal searches initiated with the Phd protein recovered YefM and those of its relatives that are encoded by genes co-occurring with genes for the YoeB family of RelE/ParE related toxin homologs. Hereinafter, we refer to these proteins as the Phd/YefM superfamily. The Phd/YefM superfamily is characterized by a conserved domain that is approximately 70 to 75 residues in length. This domain is predicted to bind DNA based on the experimental studies on the phage P1 Phd protein and the *E. coli* DnaT protein, which functions in DNA replication [29-32]. Secondary structure prediction based on the multiple sequence alignment (Figure 4) revealed that the DNA-binding domain of the Phd/YefM superfamily is likely to adopt an $\alpha + \beta$ fold with amino- and carboxy-terminal helices flanking a central β-hairpin. This secondary structure pattern does not suggest any direct relationship to the MetJ/Arc or HTH folds, suggesting that the Phd/YefM domain may define a unique DNA-binding fold. The Phd protein is a transcription regulator of the toxin Doc, and functions as the antitoxin of the phage P1 plasmid PSK system. Though the Phd-Doc PSK system is functionally analogous to the RelE/ParE systems, the toxin Doc is unrelated to the RelE/ParE superfamily (see below). However, based on the organization of the gene neighborhoods in YoeB family (Figure 3), the Phd/YefM proteins encoded by the upstream genes are predicted to function as transcriptional regulators and antitoxins of the YoeB proteins. Interestingly, the Phd/YefM domain is also fused to

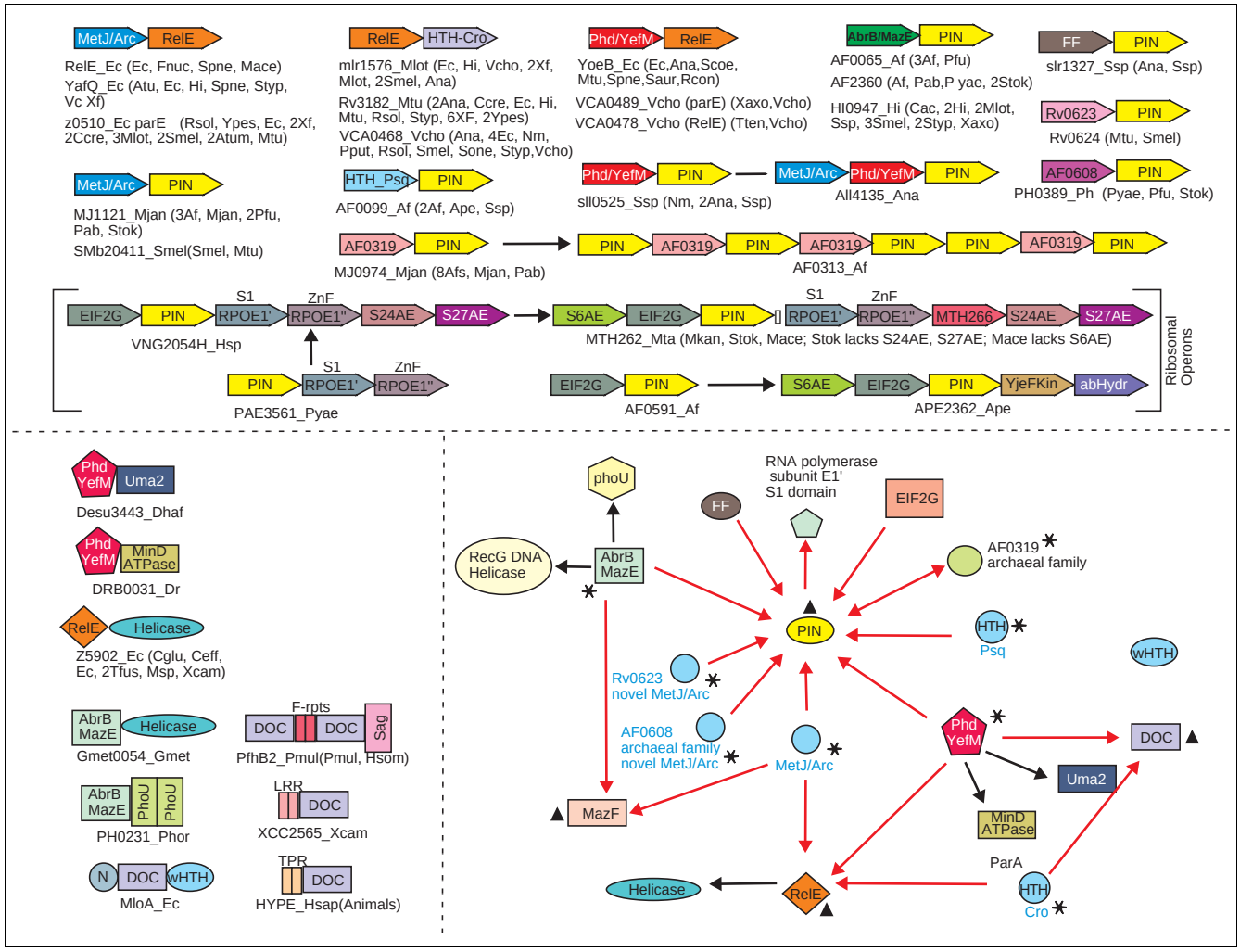

Figure 3

Contextual information and an ordered graph of gene neighborhood and domain architectures of the PSK network. The top panel shows the gene neighborhoods (predicted operons) for some of the PSK systems and other relevant gene clusters. The arrows indicate the direction of transcription. For each gene neighborhood, representative gene names are given below the depicted operon and the phyletic distribution of the operons is provided in brackets. The organisms are abbreviated as Figures 1 and 2. If an organism has more than one representative of a given PSK system, that number is appended before the organism's abbreviation. If one of the organisms has additional functionally relevant genes in the neighborhood, then these neighborhoods are shown separately, and linked to the core conserved gene neighborhood with an arrow. The lower right panel shows the ordered graph for the contextual information contained in conserved gene neighborhoods and domain fusions. The red edge in the graph denotes a neighboring gene, while the black edge denotes domain fusion. The direction of the edge denotes the order of the genes or the order of the fusion of domain in the polypeptide. Members of the vast assemblage of DNA-binding domains that share common structural features, namely the HTH (helix-turn-helix) and the RHH (ribbon-helix-helix) folds, have been colored blue. The triangles indicate toxins and the stars indicate anti-toxins/transcription factors. Domain architectures of a select set of proteins discussed in the text are shown in the lower left panel. The domain abbreviations are: abhydr, alpha/beta hydrolase; eif2G, translation initiation factor eIF-2, gamma subunit; FF, protein-protein interaction domain from human hypa/fbp11; Frpts, tetratricopeptide repeats; HTH Psq, HTH of the pipsqueak variety; LRR, leucine rich repeats; N, amino-terminal alpha-helical domain found in MloA-like proteins; RpoE1, DNA-directed RNA polymerase subunit E9; RpoE2, DNA-directed RNA polymerase subunit E99; S6E, ribosomal protein S6E; S24E, 30S ribosomal protein S24E; S27AE, 30S ribosomal protein S27AE; Sag, *Yersinia/Haemophilus* virulence surface antigen; TPR, tetratricopeptide repeats; YjeFKin, YjeF-like ribokinase. The species abbreviations are as shown in Figures 1, 2 and additionally: Pab, *Pyrococcus abyssi*; Pfu, *Pyrococcus furiosus*; Pyae, *Pyrobaculum aerophilum*; Hsom, *Haemophilus somnus*; Pmul, *Pasteurella multocida*; Spne, *Streptococcus pneumoniae*; Styp, *Salmonella typhimurium*; Ypes, *Yersinia pestis*.

the MinD ATPase in *Deinococcus* that is involved in chromosomal partitioning in bacteria, and domains of the Uma2 superfamily in *Desulfitobacterium* (Figure 3). Given that the Phd/YefM domain is a DNA-binding domain, it is possible that it is has been secondarily recruited in certain bacteria to tether other catalytic activities, such as the MinD to DNA. The Uma2 domain family shows a lineage specific expansion in cyanobacteria, *Streptomyces* and *Desulfitobacterium* (Figure

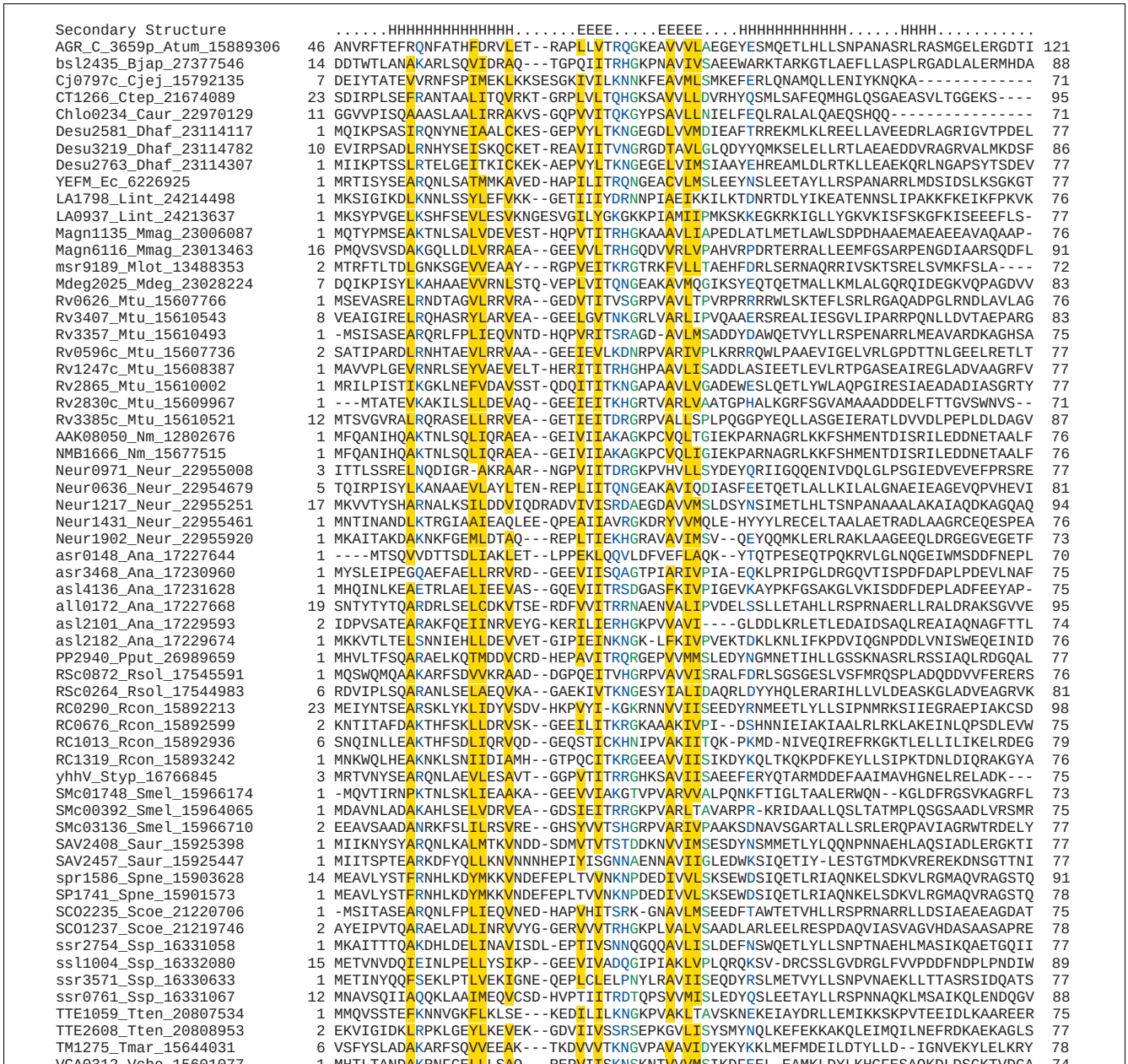

Figure 4 *(see legend on next page)*

Figure 4 *(see previous page)*
Multiple alignment of Phd/YefM. The labeling and coloring conventions are as followed in Figure 1. The species abbreviations are as shown in Figure 1, 2 and additionally: Bjap, *Bradyrhizobium japonicum*; Cjej, *Campylobacter jejuni*; Mdeg, *Microbulbifer degradans*; Spne, *Streptococcus pneumoniae*; Styp, *Salmonella typhimurium*; Tmar, *Thermotoga maritima*; Ypes, *Yersinia pestis*.

2). The proteins of this superfamily contained conserved acidic residues (data not shown), suggesting that it might also function as an uncharacterized enzyme that acts on DNA.

Genes encoding members of the Rv3182, mlr1576, VCA0468 families of the RelE/ParE superfamily were consistently associated with conserved downstream genes that encoded small proteins (90-110 residues) unrelated to either the Phd/YefM or MetJ/Arc superfamilies (Figure 3). PSI-BLAST searches initiated with these proteins showed that they all contained a conserved helix-turn-helix domain related to the lambda cro protein (cHTH domain). This suggested that they are likely to be DNA-binding proteins that act as transcription regulators of the upstream genes, which encoded members of the RelE/ParE superfamily. By analogy to the other PSK systems, these cHTH proteins are also expected to function as antitoxins countering the action of the products of their upstream genes. However, given the 'reverse' organization with respect to the classical PKS systems, it is conceivable that the functional interaction between the cHTH transcriptional regulator and the toxin component is different in these systems.

One possibility, which is supported by the specific relationship between these cHTH proteins and cro/cI repressors, is that these proteins act as repressors of the toxin gene. The degradation of the repressor under certain conditions could then allow the expression of the toxin component. The Z5902 family of the RelE/ParE superfamily, where the RelE/ParE domain is fused to a carboxy-terminal SF-I helicase module, differs from all other families in its predicted operon organization. These proteins typically co-occur with genes for another large helicase of superfamily II (SF-II), a restriction endonuclease and a DNA methylase. This implies that these proteins could constitute a novel restriction-modification complex, in which the RelE/ParE domain could function as a DNA-binding domain.

The above observations suggested that there is considerable unity in the organization of these toxin-antitoxin gene systems: typically these comprise of two small genes, in which one member of the pair encodes a toxin and the other encodes a DNA-binding protein that functions as an antitoxin and a transcription factor. However, the transcription factor and toxin in a functional comparable pair might belong to entirely unrelated superfamilies of proteins. Thus, genes of the RelE/ParE superfamily may be associated with genes for transcription factors belonging to either the MetJ/Arc or Phd/YefM or cHTH superfamilies. Likewise, a survey of the operonic associations for transcription factors showed that the Phd/YefM might be associated with at least two unrelated toxin superfamilies, namely RelE/ParE and Doc (see below). Nevertheless, this strongly coupled operon architecture in the form of a gene-dyad encoding a transcription factor and a toxin, appears to be a unique signature of PSK and related regulatory systems. Hence, to detect other potentially novel transcription factors and toxins, we systematically surveyed the gene neighborhoods of transcription factors which were close homologs of those associated with the RelE/ParE-superfamily toxins in order to find organizations similar to the PSK systems. We then transitively extended this scanning of gene neighborhoods on the homologs of any potential toxin candidates that were detected in the first screen and sought to detect any other transcription factors they may be associated with these newly predicted toxin-like genes. In particular, we concentrated on only those potential toxin or transcription factors that are conserved across a wide range of cellular genomes. Figure 3 illustrates the network of contextual connections that were recovered in these screens in the form of a directed graph. Previously observed associations such as that of MetJ/Arc transcription factors with toxins of the MazF superfamily [25], and Phd/YefM transcription factors with toxins of the Doc family were recovered in these screens supporting the effectiveness of this procedure.

Importantly, the screening procedure recovered a novel widespread family of small proteins (~100 residues, typified by MJ1121) that was consistently found downstream of genes for MetJ/Arc transcription factors (Figure 3). In this respect they closely resembled the operons of the RelE/ParE and MazF superfamily PSK systems. Sequence profile searches initiated with MJ1121 and its relatives showed that these small proteins comprised entirely of a RNA-binding domain, which we had previously described as the PilT-N terminal (PIN) domain [33-36]. Transitive analysis of the gene neighborhoods, using this class of solo PIN domain proteins as the pivot, showed that those versions which were not encoded by genes downstream of MetJ/Arc transcription factors were associated with other sets of conserved upstream or downstream genes (Figure 3). Analysis of these genes showed that two groups of solo PIN-protein-encoding genes were flanked by genes for transcription factors of the Phd/YefM and AbrB superfamilies [37,38], which are also found in other PSK operons as antitoxins and transcriptional regulators with other unrelated toxin genes (Figure 3). For example, MazF, the archetypal member of the MazF/CcdB/KiD superfamily of toxins, is encoded by a gene that is operonic with the MazE gene, which encodes an antitoxin of the AbrB superfamily of transcription factors. Two other groups of solo PIN-encoding genes were associated with upstream genes encoding

conserved proteins, typified by AF0608 from *Archaeoglobus* and RV0623 from *Mycobacterium tuberculosis*, respectively (Figure 3). Secondary structure prediction based on multiple alignments for these gene products showed that they comprised of small globular domains (Figure 5a,5b) with a conserved extended region followed by two helices. This secondary structure, together with the conservation pattern of the residues in these families, strongly suggested that they might define novel transcription factor families possessing a 'ribbon-helix-helix' fold as seen in the MetJ/Arc superfamily [28]. Yet another group of solo PIN domain proteins, typified by AF0099, were encoded by genes associated with upstream genes which encoded predicted DNA-binding proteins containing HTH domains belonging to the Pipsqueak family [28]. Finally, one group of solo PIN-protein-encoding genes was consistently associated with upstream genes encoding a family of small proteins that did not show detectable similarity to any known family of transcription factors. A multiple alignment of this family, with AF0319 as an archetypal member, reveals a simple $\alpha + \beta$ fold with a highly conserved amino-terminal region enriched in positively charged residues (Figure 5c). Based on the contextual precedence offered by the other T-A operons, we predict that AF0319 defines a novel class of transcription factors that regulate the expression of the PIN protein-encoding genes.

Based on this web of contextual connections offered by gene neighborhoods (Figure 3) we predict that the above-detected group of solo PIN domain proteins defines a toxin-like component of novel PSK-related regulatory systems. These predicted PSK-related systems with the PIN domain are as widespread as the systems with proteins of the RelE/ParE superfamily in both archaea and bacteria.

Functional and evolutionary connections of the PIN and Doc domains and eukaryotic nonsense-mediated mRNA decay

In contrast to the RelE proteins that are restricted to prokaryotes, the PIN domain is found in all three superkingdoms of life. This suggested that the PSK-related regulatory systems with PIN domain proteins might throw light on the more general roles of such systems. Given the RNA-binding role for the PIN domain [34-36], it is likely that these systems elicit their action by acting upon some RNA substrate. Importantly, a highly-conserved solo PIN domain protein is encoded by the archaeal super-operons that contain genes for ribosomal proteins and translation GTPases, like eIF3γ (Figure 3). This contextual connection implies that this version of the solo PIN domain is likely to function in the translation process in association with the ribosome and eIF3γ. This observation, along with the analogy to the Doc, RelE and possibly the MazF systems, implies that the PSK-related systems with PIN domains might function as translation inhibitors. The PIN domain proteins from eukaryotes suggest a deeper functional analogy between the PIN and RelE domains. These eukaryotic PIN domain proteins, such as SMG-7 from *Caenorhabditis elegans* and Nmd4p from yeast, are known to participate in the process of nonsense codon mediated decay (NMD) of mRNA [36,39-41]. In eukaryotes, this system specifically targets mRNAs with stop codons for degradation [42,43]. This suggests that the prokaryotic PSK-related systems with PIN domain proteins are likely to target transcripts in a process analogous to NMD of mRNA. There has been an earlier proposal that the PIN domain may be related to 39R59 exonucleases [36]. However, even though these two domains may have a common fold, they show differences in the conserved residues that constitute their active sites (additional data file 1) [34]. Hence, it possible that certain PIN domains, analogous to the RelE domains, cleave RNA only when it is associated with the ribosome. Thus, we predict that a ribosome-associated RNAse activity is likely to be the common mechanism of action for the solo PIN proteins in NMD as well as in prokaryotic PSK-related systems.

The above observations suggest that the crucial PIN domain protein of the NMD system is perhaps a remnant of an ancient PSK-type regulatory system. The emergence of the nucleus in eukaryotes, and the uncoupling of translation and transcription could have caused the PIN domain protein to be released from the tight regulatory circuit involving a coupled antitoxin transcription factor. Our earlier studies have suggested that other key components of the NMD system and the eukaryotic translation initiation systems have evolved from a common group of ancestral proteins [44]. The evolution of interactions with this eukaryote-specific complex might have contributed to the decoupling of the solo PIN domain proteins from the ancestral PSK-related system, and led to their incorporation into the NMD system.

We examined other superfamilies of toxins to determine if they included widely distributed members with a general functional significance similar to the solo PIN domain proteins. Several PSK-systems have a very limited phyletic distribution [6] and are not further detailed here because they are unlikely to throw light on broadly deployed regulatory mechanisms. The well-known MazF/CcdB/Kid superfamily is widely represented in the bacterial superkingdom [25] and a single archaeal genus, *Pyrococcus*, but not in eukaryotes (Figure 2). As the structures of several proteins from this superfamily are currently available, we searched the PDB database [45] with them to detect other related structures. These searches indicated that although the MazF/CcdB/Kid domain possessed a SH3-barrel fold, they were not closely related to any other members of this fold. Hence, it is likely that these domains represent a specialized version of the SH3-barrel fold that was derived in the bacteria.

The Doc toxin of the Phd-Doc PSK system has been hitherto detected only in P1-like phages and related mobile DNA elements from γ-proteobacteria [6]. Our sequence profile searches with the PSI-BLAST program recovered several homologs of Doc from several proteobacterial lineages, low

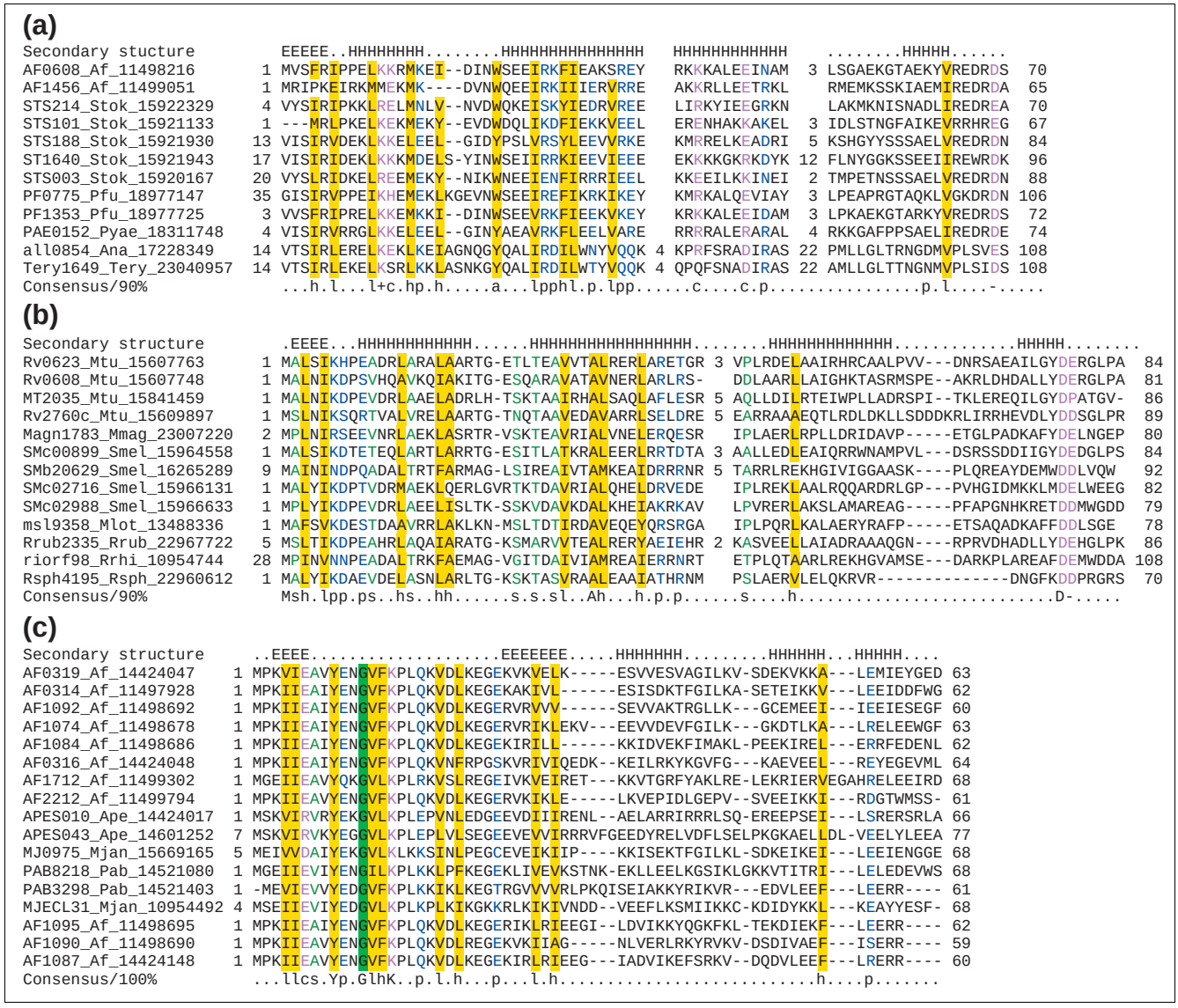

Figure 5
Multiple alignment of novel transcription factors associated with the PSK operons. **(a)** AF0608 family, **(b)** Rv0623 family and **(c)** AF0319 family. The labeling and coloring conventions are as followed in the legend to Figure 1. The species abbreviations are as shown in Figure 1 and additionally: Pab, *Pyrococcus abyssi*; Pfu, *Pyrococcus furiosus*; Pyae, *Pyrobaculum aerophilum*; Rrhi, *Rhizobium rhizogenes*; Rrub, *Rhodospirillum rubrum*; Rsph, *Rhodobacter sphaeroides*.

GC Gram positive bacteria, actinobacteria, cyanobacteria, spirochetes, *Aquifex*, *Fusobacterium*, some archaeal lineages and animals, with statistically significant expect values ($e < 0.001$). Amongst these newly-detected homologs of Doc were proteins such as the Fic protein from *E. coli* [46,47], and the huntingtin associated protein E (HYPE) [48]. The conserved region shared by all these proteins was approximately 125 to 150 residues long, and appeared to define a novel globular domain that we refer to, hereinafter, as the Doc domain.

A multiple alignment of the Doc domain superfamily (Figure 6) shows that these proteins share several nearly absolutely-conserved charged or polar residues, and the proteins are predicted to assume an α-helical fold. The amino-terminal half contains a highly-conserved histidine and a basic residue (almost always arginine), while the carboxy-terminal half contains a characteristic motif with a HX3 [DE]XNXR (where X is any amino acid) signature (Figure 6). This conservation pattern suggests that the Doc domain is a catalytic domain, with the charged or polar residues constituting the catalytic residues. While this pattern of residues does not match those seen in the active sites of any known class of α-helical enzymes, the conserved histidines and asparagine could form a metal chelating site. A mutant version of the Doc protein, in which the amino-terminal-conserved histidine is disrupted, loses its toxin activity [30]. This suggests that the catalytic

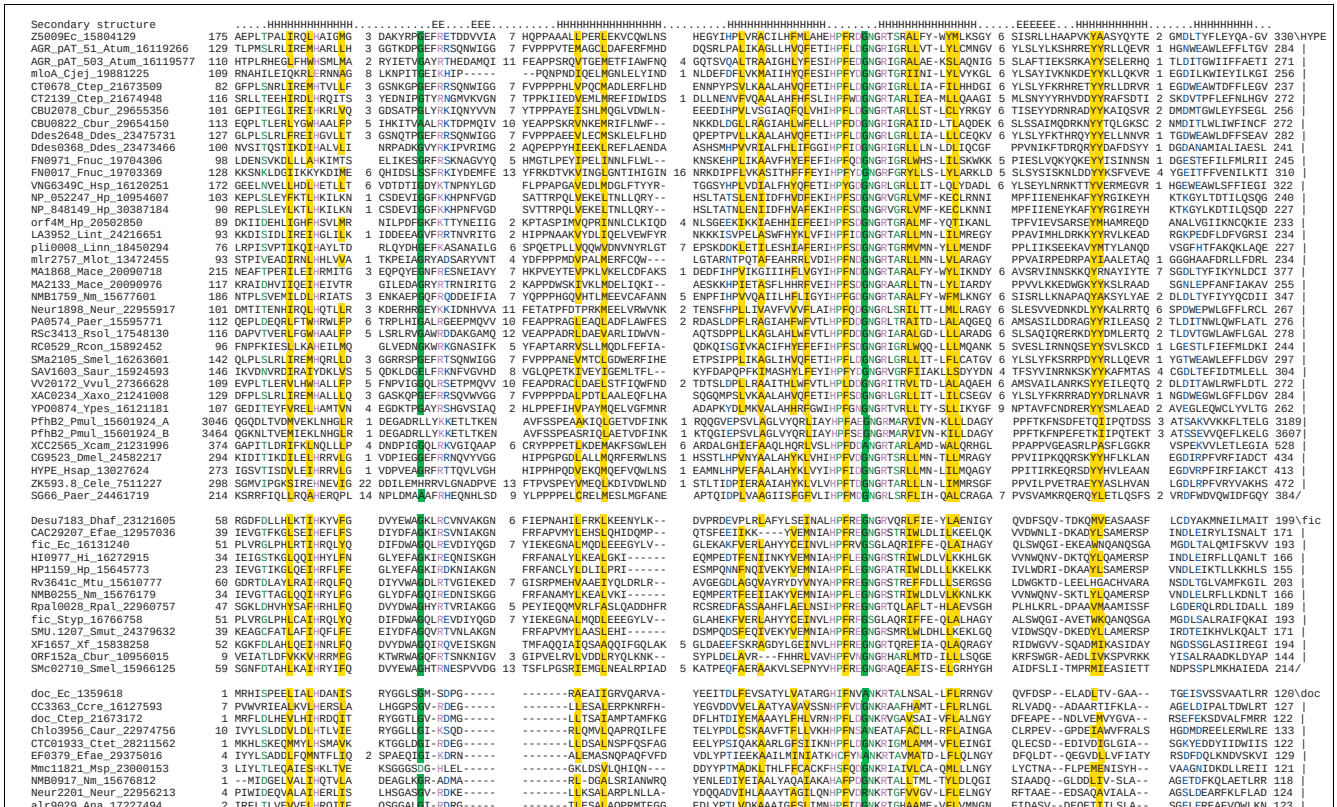

Figure 6

Multiple alignment of the Doc domain. The three major families of the Doc domain superfamily have been delineated by small blank spacers. The labeling and coloring conventions are as followed in the legend to Figure 1. The species abbreviations are as shown in Figure 1, Figure 2 and additionally: Cjej, *Campylobacter jejuni*; Ctet, *Clostridium tetani*; Ddes, *Desulfovibrio desulfuricans*; Hi, *Haemophilus influenzae*; Hp, *Helicobacter pylori*; Linn, *Listeria innocua*; Rpal, *Rhodopseudomonas palustris*; Smut, *Streptococcus mutans*; Spne, *Streptococcus pneumoniae*; Styp, *Salmonella typhimurium*; Vvul, *Vibrio vulnificus*; Ypes, *Yersinia pestis*.

activity of the Doc protein is required for its toxicity. Experimental evidence has suggested that Doc blocks a step in translation [2,8]. This observation, along with the predicted enzymatic nature for the Doc domain, suggests that it might possibly act as a nuclease that blocks translation by cleaving transcripts. Alternatively, it is possible that it acts as an uncharacterized RNA-processing enzyme that modifies transcripts and makes them unusable for translation.

A phylogenetic analysis of the Doc superfamily reveals that it contains three distinct families (Figure 6). The first family contains the Doc protein from phage P1 and its homologs from several bacterial genomes. Typically, upstream genes of this family (Figure 3). All these proteins contain a minimal stand-alone version of the Doc domain. The second family, typified by the animal HYPE protein is also found in several bacteria and some archaea. These proteins contain a longer insert after the conserved amino-terminal motifs (Figure 6) and are typically multidomain proteins. The animal HYPE contains a amino-terminal tetratricopeptide repeat (TPR) module, whereas most prokaryotic versions are fused to a carboxy-terminal DNA-binding winged HTH (wHTH) domain [28]. Interestingly, a single bacterial protein, XCC2565 from *Xanthomonas*, has leucine-rich repeats (LRR, Figure 3) amino-terminal to the Doc domain. The presence of TPR repeats is reminiscent of similar TPR modules that are present amino-terminal to the PIN domain in NMD proteins such as Smg-7 [36]. The human HYPE protein interacts with the huntingtin protein, which also contains similar α-helical ARM repeats that adopt a superstructure similar to the TPR repeats [48]. While the physiological relevance of these interactions is unclear, it is plausible that the HYPE is part of an uncharacterized multiprotein complex in the animal cells that may have a regulatory role similar to the chromosomally encoded versions of the bacterial Doc systems. Although no transcription factor genes are seen accompanying the genes for the prokaryotic HYPE orthologs, the carboxy-terminal wHTH could possibly function as an inbuilt transcriptional regulator for these proteins. A single bacterial member of the

HYPE family, namely PfhB2 from *Pasteurella*, contains two Doc domains fused to several fibrinogen-type repeats and a conserved domain found in several bacterial agglutinins (Figure 3). This protein is likely to be an extracellular protein, and may represent an unusual case of recruitment of the Doc domain for a novel function, perhaps as a secreted nuclease or an enzyme for the processing of extracellular polysaccharides. The third family of Doc-related proteins is comprised of the *E. coli* Fic protein and its orthologs from diverse bacteria (Figure 6). Like the HYPE family, they also contain a longer insert in the Doc domain after the amino-terminal conserved motif (Figure 6). These clearly do not appear to be parts of a PSK-related system for they do not show any conserved operon architectures. Mutations in the Fic protein result in filamentous growth, indicating a role in cell division [46,47]. Based on the predicted catalytic activity for the Doc superfamily, it is possible that the Fic proteins may target specific transcripts when induced under certain growth conditions.

The above analysis suggests that there is considerable diversity amongst the T-A systems. Most widespread prokaryotic PSK or related systems appear to have been derived by mixing and matching a few major classes of toxins and antitoxins (Figure 2) that appear to have independent evolutionary origins. The major classes of toxins are the RelE/ParE superfamily, the MazF/CcdB superfamily, the Doc superfamily and the solo PIN domain superfamily (Figure 2). The major classes of antitoxin transcription factors are the MetJ/Arc superfamily and related ribbon-helix-helix fold proteins, the HTH superfamily, the AbrB superfamily and the Phd/YefM superfamily. This suggests that all PSK-related systems have not descended from a common ancestor, but have been assembled on different occasions from a relatively small pool of proteins. One simple hypothesis that could account for the observed pattern of gene neighborhoods is the *in situ* displacement of genes for functionally related proteins in a tightly maintained operon. In this process, the operon architecture is maintained due to the strong functional interactions of the encoded polypeptides, but the actual origin of the polypeptides encoded by it is not constrained. This is likely to happen if unrelated polypetides can perform the same function equally effectively. This is consistent with the functional identity of different superfamilies of antitoxins that act as transcription factors. The potential functional equivalence of several unrelated toxins, such as RelE, the PIN domain and Doc domain toxins, or ParE and CcdB suggests that even the toxin genes are viable candidates for *in situ* displacement by analogs. Thus toxin or antitoxin genes could be displaced *in situ* by functionally equivalent, but unrelated genes, while the operon architecture itself is preserved. This process is highly reminiscent of the displacement of functionally equivalent, but evolutionarily unrelated genes in certain DNA recombination related operons in bacteria and phages [49]. However, the case of the RelE/ParE superfamily suggests that toxin-antitoxin gene pairs could undergo vertical evolutionary divergence to acquire very distinct functions.

Finally, the abundant presence of PSK-related systems in prokaryotic chromosomes supports the original proposal of Gerdes and recent experimental studies that these systems could function as more generic regulatory systems [5,6,8,19]. In particular, they appear to have proliferated on the chromosomes of some prokaryotes, such as the RelE system in several proteobacteria and the PIN system in archaea, *Nostoc* and *Mycobacterium tuberculosis* (Figure 2).

Furthermore, in some cases, domains such as Doc, PIN, RelE/ParE and YefM proteins appear to have been incorporated in systems that function outside the context of classic PSK-related systems.

Conclusions

Using sequence profile analysis and contextual data derived from comparative genomics, we investigated the evolutionary relationships of prokaryotic T-A systems. As a result we were able to unify the functionally unrelated toxin families defined by the ParE and RelE proteins and detect several new families of this protein superfamily. The contextual information obtained from comparative genomics allowed us to identify several new operons of PSK-related systems. One of these encodes a protein with a solo RNA-binding PIN domain as the toxin component. We suggest that these PIN domain proteins function similarly to the RelE proteins in cleaving ribosome-associated transcripts. We predict that this is likely to be a common mode of action of the PIN domain containing PSK-related systems of prokaryotes and the NMD system that cleaves transcripts with stop codons in eukaryotes. We also show that the Doc toxin defines a large family of proteins that include the animal huntingtin-interacting HYPE proteins and the bacterial Fic proteins. These proteins are predicted to function as metalloenzymes that could potentially cleave RNA. Finally, we also describe several new families of associated transcription factors that are predicted to function as antitoxins in the newly identified PSK systems. These predictions are likely to aid in experimental investigation of poorly understood aspects of both eukaryotic and prokaryotic regulatory systems, including the process of nonsense mediated decay in eukaryotes.

Materials and methods

The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda) was searched using the BLASTP program [11]. Profile searches were conducted using the PSI-BLAST program with either a single sequence or an alignment used as the query, with a default profile inclusion expectation (E) value threshold of 0.01 (unless specified otherwise), and was iterated until convergence [11,13]. For all searches with compositionally biased proteins we used a statistical correction for this bias to reduce false positives in these searches. Multiple alignments were constructed using the T_Coffee [20] or

PCMA [50] programs, followed by manual correction based on the PSI-BLAST results. All large-scale sequence analysis procedures were carried out using the SEALS package [51].

Structural manipulations were carried out using the Swiss-PDB viewer program [52] and the ribbon diagrams were constructed with MOLSCRIPT [53]. Searches of the PDB database with query structures was conducted using the DALI program [54]. Protein secondary structure was predicted using a multiple alignment as the input for the PHD program [21]. Similarity-based clustering of proteins was carried out using the BLASTCLUST program [55]. Phylogenetic analysis was carried out using the maximum-likelihood, neighbor-joining and least squares methods [56,57]. Briefly, this process involved the construction of a least squares tree using the FITCH program [58] or a neighbor joining tree using the NEIGHBOR [57] or the MEGA program [59], followed by local rearrangement using the ProtML program of the Molphy package [57] to arrive at the maximum likelihood (ML) tree. The statistical significance of various nodes of this ML tree was assessed using the relative estimate of logarithmic likelihood bootstrap (ProtML RELL-BP), with 10,000 replicates. Gene neighborhoods were determined by searching the NCBI PTT tables with a script that was custom-written by the authors. Briefly the procedure involved collecting fixed neighborhoods centered on a set of query genes, followed by the clustering of their products using the BLASTCLUST program to determine related products. The presence of clusters of related genes amongst the neighbors of the query set implied the presence of conserved gene neighborhoods. This was used in combination with a previously reported screen for conserved gene neighborhoods [15,35]. These tables can be accessed from the genomes division of the Genbank database [60].

Additional data files

A complete list of all the novel proteins belonging to the various superfamilies discussed in this paper will be made available for download via [61]. A multiple alignment of selected PIN domains (Additional data file 1), including the predicted toxins of PSK-like systems is provided with the online version of this article.

References

1. Couturier M, Bahassi el M, Van Melderen L: **Bacterial death by DNA gyrase poisoning.** *Trends Microbiol* 1998, **6:**269-275.
2. Engelberg-Kulka H, Glaser G: **Addiction modules and programmed cell death and antideath in bacterial cultures.** *Annu Rev Microbiol* 1999, **53:**43-70.
3. Jensen RB, Gerdes K: **Programmed cell death in bacteria: proteic plasmid stabilization systems.** *Mol Microbiol* 1995, **17:**205-210.
4. Yarmolinsky MB: **Programmed cell death in bacterial populations.** *Science* 1995, **267:**836-837.
5. Christensen SK, Mikkelsen M, Pedersen K, Gerdes K: **RelE, a global inhibitor of translation, is activated during nutritional stress.** *Proc Natl Acad Sci USA* 2001, **98:**14328-14333.
6. Gerdes K: **Toxin-antitoxin modules may regulate synthesis of macromolecules during nutritional stress.** *J Bacteriol* 2000, **182:**561-572.
7. Jiang Y, Pogliano J, Helinski DR, Konieczny I: **ParE toxin encoded by the broad-host-range plasmid RK2 is an inhibitor of *Escherichia coli* gyrase.** *Mol Microbiol* 2002, **44:**971-979.
8. Hazan R, Sat B, Reches M, Engelberg-Kulka H: **Postsegregational killing mediated by the P1 phage 'addiction module' phd-doc requires the *Escherichia coli* programmed cell death system mazEF.** *J Bacteriol* 2001, **183:**2046-2050.
9. Pedersen K, Zavialov AV, Pavlov MY, Elf J, Gerdes K, Ehrenberg M: **The bacterial toxin RelE displays codon-specific cleavage of mRNAs in the ribosomal A site.** *Cell* 2003, **112:**131-140.
10. Christensen SK, Gerdes K: **RelE toxins from bacteria and Archaea cleave mRNAs on translating ribosomes, which are rescued by tmRNA.** *Mol Microbiol* 2003, **48:**1389-1400.
11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
12. Neuwald AF, Liu JS, Lipman DJ, Lawrence CE: **Extracting protein alignment models from the sequence database.** *Nucleic Acids Res* 1997, **25:**1665-1677.
13. Aravind L, Koonin EV: **Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches.** *J Mol Biol* 1999, **287:**1023-1040.
14. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402:**83-86.
15. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context.** *Genome Res* 2001, **11:**356-372.
16. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
17. Aravind L: **Guilt by association: contextual information in genome analysis.** *Genome Res* 2000, **10:**1074-1077.
18. Huynen M, Snel B, Lathe W 3rd, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res* 2000, **10:**1204-1210.
19. Hayes CS, Sauer RT: **Toxin-antitoxin pairs in bacteria: killers or stress regulators?** *Cell* 2003, **112:**2-4.
20. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302:**205-217.
21. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232:**584-599.
22. Roberts RC, Helinski DR: **Definition of a minimal plasmid stabilization system from the broad-host-range plasmid RK2.** *J Bacteriol* 1992, **174:**8119-8132.
23. Kamada K, Hanaoka F, Burley SK: **Crystal structure of the MazE/MazF complex: molecular bases of antidote-toxin recognition.** *Mol Cell* 2003, **11:**875-884.
24. de la Cueva-Mendez G: **Distressing bacteria: structure of a prokaryotic detox program.** *Mol Cell* 2003, **11:**848-850.
25. Mittenhuber G: **Occurrence of mazEF-like antitoxin/toxin systems in bacteria.** *J Mol Microbiol Biotechnol* 1999, **1:**295-302.
26. Hargreaves D, Santos-Sierra S, Giraldo R, Sabariegos-Jareno R, de la Cueva-Mendez G, Boelens R, Diaz-Orejas R, Rafferty JB: **Structural and functional analysis of the kid toxin protein from *E. coli* plasmid R1.** *Structure (Camb)* 2002, **10:**1425-1433.
27. Oberer M, Zangger K, Prytulla S, Keller W: **The anti-toxin ParD of plasmid RK2 consists of two structurally distinct moieties and belongs to the ribbon-helix-helix family of DNA-binding proteins.** *Biochem J* 2002, **361:**41-47.
28. Aravind L, Koonin EV: **DNA-binding proteins and evolution of transcription regulation in the archaea.** *Nucleic Acids Res* 1999, **27:**4658-4670.
29. Gazit E, Sauer RT: **Stability and DNA binding of the phd protein of the phage P1 plasmid addiction system.** *J Biol Chem* 1999, **274:**2652-2657.
30. Magnuson R, Yarmolinsky MB: **Corepression of the P1 addiction operon by Phd and Doc.** *J Bacteriol* 1998, **180:**6342-6351.
31. Allen GC Jr, Kornberg A: **Assembly of the primosome of DNA replication in *Escherichia coli*.** *J Biol Chem* 1993, **268:**19204-19209.
32. Hayes F: **A family of stability determinants in pathogenic bacteria.** *J Bacteriol* 1998, **180:**6415-6418.

33. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV: **Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell.** *Genome Res* 1999, **9:**608-628.
34. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism.** *Nucleic Acids Res* 2002, **30:**1427-1464.
35. Koonin EV, Wolf YI, Aravind L: **Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach.** *Genome Res* 2001, **11:**240-252.
36. Clissold PM, Ponting CP: **PIN domains in nonsense-mediated mRNA decay and RNAi.** *Curr Biol* 2000, **10:**R888-R890.
37. Huffman JL, Brennan RG: **Prokaryotic transcription regulators: more than just the helix-turn-helix motif.** *Curr Opin Struct Biol* 2002, **12:**98-106.
38. Vaughn JL, Feher V, Naylor S, Strauch MA, Cavanagh J: **Novel DNA binding domain and genetic regulation model of *Bacillus subtilis* transition state regulator abrB.** *Nat Struct Biol* 2000, **7:**1139-1146.
39. Cali BM, Kuchma SL, Latham J, Anderson P: **smg-7 is required for mRNA surveillance in *Caenorhabditis elegans*.** *Genetics* 1999, **151:**605-616.
40. Anders KR, Grimson A, Anderson P: **SMG-5, required for *C. elegans* nonsense-mediated mRNA decay, associates with SMG-2 and protein phosphatase 2A.** *EMBO J* 2003, **22:**641-650.
41. Domeier ME, Morse DP, Knight SW, Portereiko M, Bass BL, Mango SE: **A link between RNA interference and nonsense-mediated decay in *Caenorhabditis elegans*.** *Science* 2000, **289:**1928-1931.
42. Wagner E, Lykke-Andersen J: **mRNA surveillance: the perfect persist.** *J Cell Sci* 2002, **115:**3033-3038.
43. Schell T, Kulozik AE, Hentze MW: **Integration of splicing, transport and translation to achieve mRNA quality control by the nonsense-mediated decay pathway.** *Genome Biol* 2002, **3:**reviews1006.1-1006.6.
44. Aravind L, Koonin EV: **Eukaryote-specific domains in translation initiation factors: implications for translation regulation and evolution of the translation system.** *Genome Res* 2000, **10:**1172-1184.
45. **PDB - Protein Data Bank** [http://www.rcsb.org/pdb/]
46. Komano T, Utsumi R, Kawamukai M: **Functional analysis of the fic gene involved in regulation of cell division.** *Res Microbiol* 1991, **142:**269-277.
47. Kawamukai M, Matsuda H, Fujii W, Utsumi R, Komano T: **Nucleotide sequences of fic and fic-1 genes involved in cell filamentation induced by cyclic AMP in *Escherichia coli*.** *J Bacteriol* 1989, **171:**4525-4529.
48. Faber PW, Barnes GT, Srinidhi J, Chen J, Gusella JF, MacDonald ME: **Huntingtin interacts with a family of WW domain proteins.** *Hum Mol Genet* 1998, **7:**1463-1474.
49. Iyer LM, Koonin EV, Aravind L: **Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52.** *BMC Genomics* 2002, **3:**8.
50. Pei J, Sadreyev R, Grishin NV: **PCMA: fast and accurate multiple sequence alignment based on profile consistency.** *Bioinformatics* 2003, **19:**427-428.
51. **SEALS Home Page** [http://www.ncbi.nlm.nih.gov/CBBresearch/Walker/SEALS/index.html]
52. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723.
53. Kraulis PJ: **Molscript.** *J Appl Cryst* 1991, **24:**946-950.
54. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, **233:**123-138.
55. **BLASTCLUST - BLAST score-based single-linkage clustering** [ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.txt]
56. Felsenstein J: **Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.** *Methods Enzymol* 1996, **266:**418-427.
57. Hasegawa M, Kishino H, Saitou N: **On the maximum likelihood method in molecular phylogenetics.** *J Mol Evol* 1991, **32:**443-445.
58. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
59. Kumar S, Tamura K, Jakobsen IB, Nei M: **MEGA2: molecular evolutionary genetics analysis software.** *Bioinformatics* 2001, **17:**1244-1245.
60. **NCBI Entrez Genome** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome]
61. **Supplementary information and additional files** [ftp://ftp.ncbi.nih.gov/pub/aravind/rele/]