Bacterial toxin-antitoxin systems
Translation inhibitors everywhere

Julien Guglielmini1,2 and Laurence Van Melderen3,*
1 Microbial Evolutionary Genomics Group; Institut Pasteur; Paris, France; 2 CNRS; URA2171; Paris, France; 3 Génétique et Physiologie Bactérienne; Faculté de Sciences; Université Libre de Bruxelles; Bruxelles, Belgium

Keywords: RelE/ParE, Gin, selfish genes, horizontal gene transfer, endoribonuclease, DNA-gyrase

Toxin-antitoxin (TA) systems are composed of two elements: a toxic protein and an antitoxin which is either an RNA (type I and III) or a protein (type II). Type II systems are abundant in bacterial genomes in which they move via horizontal gene transfer. They are generally composed of two genes organized in an operon, encoding a toxin and a labile antitoxin. When carried by mobile genetic elements, these small modules contribute to their stability by a phenomenon denoted as addiction. Recently, we developed a bioinformatics procedure that, along with experimental validation, allowed the identification of nine novel toxin superfamilies. Here, considering that some toxin super-families exhibit dramatic sequence diversity but similar structure, bioinformatics tools were used to predict tertiary structures of novel toxins. Seven of the nine novel super-families did not show any structural homology with known toxins, indicating that combination of sequence similarity and three-dimensional structure prediction allows a consistent classification. Interestingly, the novel super-families are translation inhibitors similar to the majority of known toxins indicating that this activity might have been selected rather than more detrimental traits such as DNA-gyrase inhibitors, which are very toxic for cells.

The Toxin-Antitoxin System Framework

Toxin-antitoxin (TA) systems are diverse and widespread in bacterial genomes.1-3 These small modules are part of the mobilome, which is constituted by the huge collection of phages, plasmids, transposons, ICEs (integrative and conjugative elements) and other selfish entities. This gigantic gene pool shared by bacteria, even phylogenetically distant ones, is under constant flux and can exchange with chromosomes. It therefore makes a large contribution to bacterial evolution.

TA systems are composed of closely linked genes, encoding a toxic protein that can harm the cell and a labile antitoxin that either inhibits toxin expression (type I) or sequesters it in a harmless complex (type II and III). While the toxin is always a protein, the antitoxin can be either an RNA (type I and III) or a protein (type II) (for recent review, see ref. 4). Type II TA systems are organized in an operon, the upstream gene encoding the antitoxin. The expression of the two genes is regulated at the level of transcription by the antitoxin-toxin complex. Recently, exceptions to this framework were reported in the literature. ‘Reverse organization’ systems were described in which the toxin gene precedes that of the antitoxin within the operon as well as three-component systems in which the transcriptional regulation activity is performed by a third gene in the operon, preceding that of the antitoxin and the toxin (for a recent review, see ref. 4). Thus, diversity in gene organization within the group of type II systems is observed.

The Selfish Nature of Toxin-Antitoxin Systems

Type II systems were originally discovered on plasmids in which they participate to their stable maintenance in growing bacterial populations. They act by a mechanism denoted as post-segregational...
killing or addiction \cite{5,6} relying on the differential stability of the antitoxin and toxin proteins. In daughter-bacteria devoid of plasmid copy, the labile antitoxin is degraded thereby freeing the toxin from the harmless antitoxin-toxin complex. The free toxin will then interact with its target and inhibit cell growth and/or survival. This ensures plasmid prevalence within bacterial populations. Plasmid-encoded TA systems are also involved in plasmid-plasmid competition by eliminating bacteria containing plasmid of the same incompatibility group devoid of TA systems.\cite{7}

For systems integrated in bacterial chromosomes, two general functions are prevailing in the literature: stress response and DNA stabilization. For the stress response hypothesis, it comes in different flavors. It has been reported that the \textit{E. coli} mazEF system is an altruistic programmed cell death system that sacrifices part of the population in adverse conditions \cite{9,10} (for review, see ref. 8). This hypothesis is highly controversial since it is not a reproducible phenomenon.\cite{9,10} Other hypotheses related to persistence or to stress response against amino acid starvation or antibiotic treatments have been proposed.\cite{4,3,12} Regarding the stabilization hypothesis, it seems now clear that the main function of integrated TA systems is tightly linked to their addictive properties. They indeed contribute to the stability of ICEs or super-integrons as observed for plasmid-encoded systems.\cite{5,14} Another possibility that has not encountered much attention so far is that these systems might be devoid of any biological roles and may simply be selfish elements.\cite{9,10,15} Their stabilization properties might just be a consequence of their addictive behavior. Related to the selfish hypothesis, TA systems might also be involved in competition between mobile genetic elements as described above.\cite{7}

Interestingly, specific TA systems from the three types have been involved in protection against phages.\cite{16,17,18} Finally, given that an antitoxin can antagonize a toxin from another system in \textit{trans}, TA systems might contribute to the fitness of the replicon that carries them by eliminating competitors and/or surviving to the loss of competitors as proposed in the anti-addiction hypothesis.\cite{19} This could explain the evolutionary success of TA systems in the bacterial world.

### En Route for a Novel Classification

Originally, type II TA systems were classified into 10 families.\cite{3,20} This classification is based on the amino acid sequence similarity of the toxins and it was assumed that each toxin family is specifically associated with an antitoxin family. However, Anantharaman and Aravind showed that the association of a toxin family with different antitoxin families was frequently detected in bacterial genomes.\cite{2} As an example, genes from the RelE/ParE superfamily are associated with genes from the RelB, Phd, HigA or PasA antitoxin superfamilies and interestingly, with genes that do not belong to any known antitoxin super-families.\cite{2} Two ‘hybrid’ systems mixing RelE toxins with Phd and VapB antitoxins respectively were validated experimentally.\cite{9,10,21,22} This definitively proved that the concept of TA system families is inadequate and we proposed to consider toxin and antitoxin super-families independently.\cite{9} This ‘mix and match’ phenomenon opens the possibility that known toxin genes might be associated with genes representing novel antitoxins and vice versa. Based on this idea, we developed a bioinformatics tool based on the ‘guilt by association’ principle to explore bacterial genomes and identify families of toxin and antitoxin unrelated to known ones.\cite{3} Using this procedure, we predicted more than 500,000 toxin and antitoxin sequences in 2,181 bacterial genomes (chromosomes, plasmids and phages). Clustering of these sequences by the Markov cluster algorithm (MCL)\cite{25} defined more than 62,000 potential antitoxin and toxin super-families, although a majority of them contained very few sequences. We experimentally tested a tiny subset of these candidates using a simple in vivo assay in \textit{E. coli}: overexpression of the candidate toxin has to cause cell growth inhibition and co-expression of the cognate antitoxin has to alleviate cell growth inhibition. Twenty-three toxin and 18 antitoxin sequences originating from different and sometimes quite distant species from \textit{E. coli} were successfully validated. Unexpectedly, all these toxins inhibit translation in \textit{E. coli}, as do a vast majority of the known toxins.

The experimentally validated toxin sequences defined four type II toxin super-families for which a classical type II antitoxin was experimentally validated as well as five ‘solitary’ toxin super-families. For these super-families, we were unable to detect type II antitoxin activities for the ORFs flanking the toxin candidate that we experimentally validated. We cannot exclude that although identified by ‘guilt by association’, the antitoxin activity is encoded by a small RNA (like in type I and III systems) or other yet undefined types of activity. This could further exemplify the ‘mix and match’ phenomenon.

These nine super-families were denoted as GinA to GinI (for growth inhibition) and defined super-families unrelated to known ones based on the MCL clustering. A priori, the clustering procedure was satisfying since it grouped the CcdB and MazF sequences into the CcdB/MazF super-family although they share very little sequence similarity and were originally grouped on the basis of their common three-dimensional structures.\cite{24,25} It also grouped the ParE and RelE sequences into the large ParE/RelE super-family.\cite{2}

Using Phyre2\cite{26} and DALI,\cite{27} we searched for structural homologs to the GinA to GinI super-families and included the VapD, YafO, HicA and RnlA sequences in this analysis as no three-dimensional structure is available for these toxins (Table 1). Interestingly, it turned out that the GinB super-family shows significant predicted structural homology with the RelE toxin from \textit{T. thermophilus} (z score: 16.1; it is generally considered that 2 folds are similar when the z score is greater than 3.5; rmsd: 0.5, the lower the better) although this was neither detected by MCL nor in the CDD database (as GinB sequences do not match with the typical RelE COG2026 or Pfam05016).

Based on this and on preliminary experimental data indicating that GinB toxins induce mRNA cleavage, as do the RelE-like toxins (Goeders, Drèze and Van Melderen, unpublished data), we propose to include the GinB sequences in the ParE/RelE super-family. Interestingly, the ParE/RelE-fold appears to be quite widespread within mobile genetic elements, such as the RegB protein of phage T4\cite{28} and the ColE5 E5 toxin encoded by the ColE5 plasmid.\cite{29} Both proteins are involved in RNA degradation with RegB...
Structural homologs and conserved domains of the Gin, VapD, HicA, YafO and RnlA toxin super-families

| Super-family | Structural homologs | Conserved domains | References |
|--------------|---------------------|-------------------|------------|
| Type II      |                     |                   |            |
| GinA         | None                | Siphovirus Gp157  | 1, 32      |
|              |                     | protein family    |            |
|              |                     | Related to SFI    |            |
|              |                     | phage from        |            |
|              |                     | *Streptococcus*    |            |
|              |                     | *therophilus*      |            |
|              |                     | Thought to        |            |
|              |                     | protect against   |            |
|              |                     | phages           |            |
| GinB         | ReIE toxin of       | DUF213, COG4680:  | 1, 43      |
|              | *Thermus*           | uncharacterized   |            |
|              |                     | protein conserved |            |
|              |                     | in bacteria       |            |
| GinC         | Protein of unknown  | UPF0223, PRK04387:| 1          |
|              | function of *Bacillus*| uncharacterized   |            |
|              |                     | protein family    |            |
| GinE         | Putative RNA binding| UPF0150: protein  | 1          |
|              | protein in          | family that may   |            |
|              |                     | be involved in    |            |
|              |                     | RNA metabolism,   |            |
|              |                     | including RNA     |            |
|              |                     | binding and       |            |
|              |                     | cleavage          |            |
| GinF         | Pleckstrin domain   | None              | 1, 33      |
|              | of *Shewanella*     |                   |            |
|              |                     | *loihica* (PDB: 3dck)|          |
| GinG         | None                | None              | 1          |
| GinH         | None                | None              | 1          |
| VapD         | Sequence specific   | Cas2 protein      | 31         |
|              | endoribonuclease   | associated with   |            |
|              |                     | CRISPR; RNase     |            |
|              |                     | specific to U-rich|            |
| HicA         | Protein of unknown  | YcfA super-family:| 1, 20      |
|              | function of *Thermus*| hypothetical       |            |
|              |                     | proteins of unknown|            |
|              |                     | function; COG1724:|            |
|              |                     | predicted RNA     |            |
|              |                     | binding proteins  |            |
|              |                     | (dsRBD-like fold),|            |
|              |                     | HicA family       |            |
| YafO         | None                | None              | 44, 45     |
| RnlA         | None                | None              | 46         |
| Solitary     |                     |                   |            |
| GinE         | Putative RNA binding|                   | 1          |
|              | protein in          |                   |            |
|              |                      | *Lactobacillus*    |            |
|              |                     | *plantarum* (PDB: 3kw)|        |
| GinF         | Pleckstrin domain   |                   |            |
|              | of *Shewanella*     |                   |            |
|              |                     | *loihica* (PDB: 3dck)|       |
| GinG         | None                |                   |            |
| GinH         | None                |                   |            |
| GinI         | Protein of unknown  |                   | 1, 20      |
|              | function of *Thermus*|                   |            |
|              |                     | *therophilus* (PDB: 1whz)|     |
|              |                     | Endonuclease of    |            |
|              |                     | *Pyrococcus*       |            |
|              |                     | *furiosus* (PDB: 1dq3)|        |
| YafO         | None                |                   |            |
| RnlA         | None                |                   |            |
|               |                     |                   |            |

Structural homologs were identified using Phyre2 and DALI. Conserved domains were identified using the CDD database.

The GinD super-family exhibits the lowest diversity (Fig. 1A). Almost all GinD sequences are associated with sequences belonging to HigA antitoxin super-family (88%). The GinC super-family shows a more complex pattern (Fig. 1B). Interestingly, a significant proportion (32%) of GinC sequences are associated with FizC sequences. We experimentally validated one of the FizC sequence as a novel antitoxin in the original paper. FizC sequences are predicted thioredoxin-like domains, which are also found within Cas proteins from CRISPR-Cas systems, highlighting again a possible connection between TA and CRISPR-Cas systems. Moreover, 13% of the GinC sequences are associated to HigA sequences. The remaining sequences are associated with sequences encoding predicted enzymes (phosphatases, decarboxylases, dehydrogenases, reductases, etc). A very small proportion of GinC sequences are associated with ORFs for which no
conserved domain has been detected. The GinA super-family shows the most complex pattern (Fig. 1C). A significant number of GinA sequences are associated with ORFs without conserved domains. Among these, 5.4% are FizF sequences (another experimentally validated antitoxin super-family). About 16% of the GinA sequences are associated with HigA and/or FizA sequences. As for FizC sequences, we experimentally validated one of the FizA sequence as a novel antitoxin in the original paper. FizA sequences are predicted to be NTPases. The remaining sequences are associated with sequences encoding predicted enzymes such as recombinases, nucleases, transferases, sortases, etc. Interestingly, a small proportion are linked to flagellar proteins. Flagellar systems can be converted into secretion systems (partially homologous to type III secretion systems). This could be an indication that some secreted toxins are homologous to toxins from TA systems.

It appears that there is a great variation in the genetic environments of the three super-families, although a significant proportion of the sequences are associated with HigA in the three cases. Moreover, GinC and GinA sequences are associated with a significant proportion of unknown proteins, thus constituting a pool of putative antitoxins that should be experimentally tested. Another possibility is that GinA and GinC sequences are not bona fide type II toxins and/or are associated with type I or type III RNA antitoxins. We cannot exclude that some GinA and GinC toxins might be translation regulators involved in specific physiological processes, which could be determined by the genomic context. Indeed, although both the GinA and GinC sequences are associated with sequences encoding predicted enzymes, the nature of these enzymes is different. GinA sequences appear to be associated with sequences encoding predicted enzymes mostly involved in DNA metabolism while GinC sequences seem to be linked preferentially to sequences encoding predicted enzymes involved in general metabolism. This could represent a possible evolution of toxins and explain the fact that gene neighborhoods seem to be toxin-specific. Of course, we cannot rule out the possibility that some of these ORFs containing an enzyme domain are acting as antitoxins, as FizA and FizF antitoxins.

Conclusion

Since their discovery, type II TA systems have been classified several times using the ever-growing amount of data that is collected, increasing the number of families at each new classification. Successive classifications were mainly based on primary sequence similarity and on the specificity of interactions between antitoxin and toxin families. It now appears that type II TA systems are far more flexible than expected. The gene order is not fixed: the antitoxin gene can be located upstream or downstream of the toxin gene and three-components systems are also found in which the antitoxin and repression activities are encoded by different genes. Moreover, shuffling occurs between the three types of TA systems since the ToxN toxin of the type III toxIN system belongs to the CcdB/MazF super-family. This ‘mix and match’ phenomenon is further exemplified by our observation that the Gin ‘solitary’ toxins share common three-dimensional structures with the classical type II HicA toxin. In addition, the group of Storz showed that the type I SymE toxin has structural similarities with MazE/AbrB super-family member to which belongs the MazE antitoxin, indicating that a toxin can evolve from a antitoxin if the genetic neighborhood allows it.

The concept of type II TA system families has also been challenged, since the same toxin family can be associated with different antitoxin families. In addition, three-dimensional structures of toxins and antitoxins revealed relationships that were not evidenced by looking at the primary sequences. As an example, MazF and CcdB toxins have a common ancestor considering their three-dimensional structure although they have different activities. The same observation is made for the large ParE/RelE super-family. However, obtaining experimental three-dimensional structures might be fastidious and tricky since overexpression of these toxins causes cell growth inhibition and/or death. Co-expression with the cognate antitoxin is often used to overcome this problem since toxin resistant mutants are not available except for the CcdB toxin (for reviews, see refs. 41 and 42). Therefore, three-dimensional structure prediction tools might provide useful information prior to obtaining the experimental data.

Based on the data obtained in our original work and in this work, we propose an approach that combines the ‘guilt by association’ principle, MCL clustering and three-dimensional structure predictions to allow discovery and putative definition of novel toxin and antitoxin super-families. Table 2 shows the 13 super-families of type II toxins and the four super-families of ‘solitary’ toxins known at the time of writing.

Three-dimensional structure predictions, albeit speculative, can also give insight into the evolution of TA systems. Some of them are structurally related to systems that mediate defense against invading plagues and plasmids as seen with the VapD sequences, which might pinpoint a common origin to TA and CRISPR-cas systems. Interestingly, each of the super-families except Zeta contains translation inhibitors. The ParE and CcdB that convert DNA-gyrase into cellular
poisons are not isolated families, but are clearly related to translation inhibitors on sequence and structure basis. What if toxins from TA systems were initially translation inhibitors that evolved in different directions such as replication inhibitors, Cas genes or secreted toxins? Nevertheless, TA systems in which the toxin is a translation inhibitor or RNase have been evolutionary selected as compared with other activities that are certainly more ‘dangerous’ for the cell. Less dangerous, or ‘slowly killing’, toxins might have a wider evolutionary landscape, which could be explained by the fact that selection has the time to act on them, leading by chance to pseudo-genes or to new functions.

Acknowledgments

We thank Marie Deghorain, Nathalie Goeders, Damien Geerearts and Thomas Jové for critical reading of the manuscript and Eduardo Rocha for useful discussions. Work in the laboratory of LVM is supported by the Fonds de la Recherche (FRSM-3.4530.04), Fondation Van Buuren and Fonds Jean Brachet.
48. Brown BL, Wood TK, Peti W, Page R. Structure of the Escherichia coli antitoxin MqsA (YgiT/b3021) bound to its gene promoter reveals extensive domain rearrangements and the specificity of transcriptional regulation. J Biol Chem 2011; 286:2285-96; PMID:21068382; http://dx.doi.org/10.1074/jbc.M110.172643

49. Mutschler H, Gebhardt M, Shoeman RL, Meinhart A. A novel mechanism of programmed cell death in bacteria by toxin-antitoxin systems corrupts peptidoglycan synthesis. PLoS Biol 2011; 9:e1001033; PMID:21445328; http://dx.doi.org/10.1371/journal.pbio.1001033

50. Garcia-Pino A, Christensen-Dalsgaard M, Wyns L, Yarmolinsky M, Magnuson RD, Gerdes K, et al. Doc of prophage P1 is inhibited by its antitoxin partner Phd through fold complementation. J Biol Chem 2008; 283:30821-7; PMID:18757857; http://dx.doi.org/10.1074/jbc.M805654200

51. Schumacher MA, Piro KM, Xu W, Hansen S, Lewis K, Brennan RG. Molecular mechanisms of HipB-mediated multidrug tolerance and its neutralization by HipB. Science 2009; 323:396-401; PMID:19150849; http://dx.doi.org/10.1126/science.1163806

52. Bunker RD, McKenzie JL, Baker EN, Arcus VL. Crystal structure of PAE0151 from Pyrobaculum aerophilum, a PIN-domain (VapC) protein from a toxin-antitoxin operon. Proteins 2008; 72:510-8; PMID:18398909; http://dx.doi.org/10.1002/prot.22048