BIOINFORMATICS DISCOVERY NOTE Vol. 22 no. 21 2006, pages 2581–2584
doi:10.1093/bioinformatics/btl418

Genome analysis

The HicAB cassette, a putative novel, RNA-targeting toxin-antitoxin system in archaea and bacteria

Kira S. Makarova[1], Nick V. Grishin[2] and Eugene V. Koonin[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and [2]Howard Hughes Medical Institute and Department of Biochemistry, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050, USA

Received on July 18, 2006; accepted on July 26, 2006
Advance Access publication August 8, 2006
Associate Editor: Nikolaus Rajewsky

ABSTRACT

Toxin-antitoxin systems (TAS) are abundant, diverse, horizontally mobile gene modules that encode powerful resistance mechanisms in prokaryotes. We use the comparative-genomic approach to predict a new TAS that consists of a two-gene cassette encoding uncharacterized HicA and HicB proteins. Numerous bacterial and archaeal genomes encode from one to eight HicAB modules which appear to be highly prone to horizontal gene transfer. The HicB protein (COG1598/COG4226) has a partially degraded RNAse H fold, whereas HicA (COG1724) contains a double-stranded RNA-binding domain. The stable combination of these two domains suggests a link to RNA metabolism, possibly, via an RNA interference-type mechanism. In most HicB proteins, the RNAse H-like domain is fused to a DNA-binding domain, either of the ribbon-helix-helix or of the helix-turn-helix class; in other TAS, proteins containing these DNA-binding domains function as antitoxins. Thus, the HicAB module is predicted to be a novel TAS whose mechanism involves RNA-binding and, possibly, cleavage.

Contact: koonin@ncbi.nlm.nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Since the discovery of the first plasmid addiction module, numerous plamsid- and chromosome-encoded bacterial toxin-antitoxin systems (TAS) have been characterized (Buts *et al*., 2005; Gerdes *et al*., 2005; Hayes, 2003). The TAS are classified into two types on the basis of the nature of the antitoxin (Gerdes *et al*., 2005; Hayes, 2003). Type I TAS encompass an antisense RNA antitoxin that inactivates the toxin mRNA, whereas type II TAS employ a protein antitoxin to inactivate the toxin through the protein–protein interaction. Type I TAS typically mediate post-segregational killing of plasmid-free bacteria, with the respective toxins damaging the bacterial membrane. By contrast, in the majority of the type II TAS, the protein toxins are nucleases that target mRNA, although several type II TAS employ different mechanisms, such as inhibition of DNA gyrase (Buts *et al*., 2005; Gerdes *et al*., 2005).

*To whom correspondence should be addressed.

Recent comparative-genomic analyses have revealed unexpected abundance of (predicted) TAS in most prokaryotic genomes (with the exception of small genomes of parasitic bacteria, especially, intracellular parasites) and remarkable horizontal mobility of TAS operons (Anantharaman and Aravind, 2003; Pandey and Gerdes, 2005). Furthermore, it has been proposed that TAS are stress-response systems beneficial to free-living prokaryotes rather than simple cell-killing machines (Buts *et al*., 2005; Gerdes *et al*., 2005; Hayes, 2003; Pandey and Gerdes, 2005). The emerging diversity of TAS suggests that additional, functionally similar but structurally unrelated modules await discovery.

Here we present a detailed computational analysis of a two-gene cassette encoding previously uncharacterized HicA and HicB proteins. The *hicAB* locus has been described first as an insertion into the major pilus gene cluster in several strains of *Haemophilus influenzae* (Mhlanga-Mutangadura *et al*., 1998). We report several lines of evidence suggesting that the HicAB module is a novel TAS that functions via RNA-binding and, possibly, cleavage.

2 RESULTS

2.1 HicA/COG1724 protein family belongs to the double-stranded RNA-binding fold

We used exhaustive PSI-BLAST searches (inclusion cutoff *E*-value of 0.05; non-redundant protein database; starting query HicA protein, GI: 3603335), initiated from all the sequences retrieved in any search until no new sequences could be detected, to achieve the maximum coverage of the HicA family (COG1724; this COG is annotated as 'A predicted periplasmic or secreted lipoprotein' but we are unaware of any evidence supporting this prediction). Altogether, ~230 HicA family proteins were detected in most major clades of bacteria and archaea. The greatest number of HicA sequences (12) was found in the draft genome of cyanobacterium *Crocosphaera watsonii* (Supplementary material). Among the retrieved HicA family proteins was the TTHA1913 protein from *Thermus thermophilus* HB8 for which the crystal structure has been solved as part of one of the structural-genomics projects (pdb: 1WHZ). The VAST program for structure comparison and alignment aligned the 70 residue TTHA1913 protein with a dsRBD domain of PI-PfuI intein [pdb: 1DQ3 (Ichiyanagi *et al*., 2000),

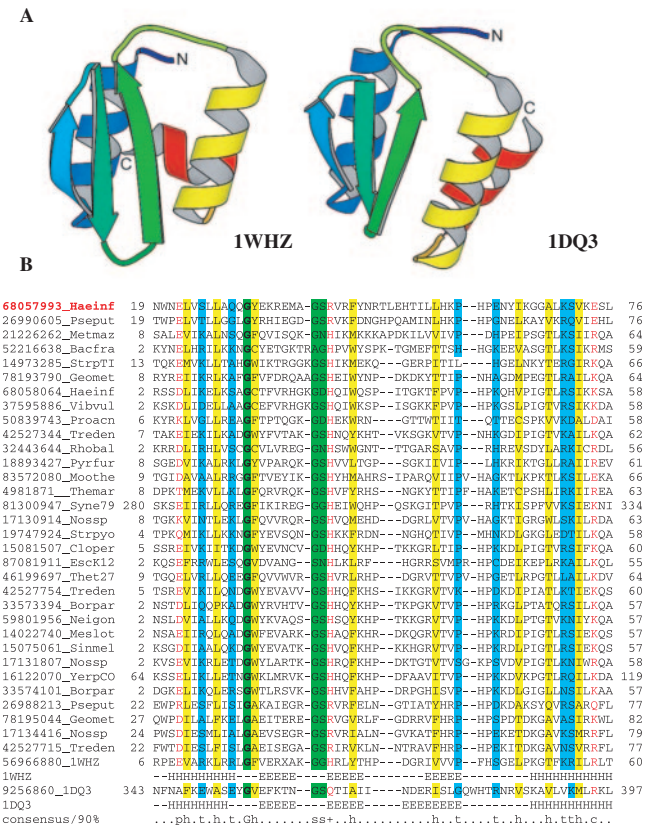

Fig. 1. (**A**) Ribbon diagrams of TTHA1913 from *Thermus thermophilus* HB8 compared with the dsRBD domain of PI-PfuI intein from *Pyrococcus furiosus* (residues A339–A412). Two dsRBD-like domains are rainbow-colored from N- to C-terminus such that the corresponding secondary structure elements in both domains have the same color. The figure was made using the program BOBSCRIPT (Esnouf, 1999). (**B**) Multiple alignment of the conserved core of the HicA family constructed using Muscle program followed by manual correction on the basis of predicted secondary structure (Cuff *et al*., 1998) and PSI-BLAST-based local alignments. Sequences are denoted by their GI numbers and abbreviated species names; the HicA protein sequence colored red. The species abbreviations are in the Supplementary Table 1S. The dsRBD sequence from the PI-PfuI intein (9256860_1DQ3) is shown in the bottom for comparison. The positions of the first and the last residues of the aligned region in the corresponding protein are indicated for each sequence. The numbers within the alignment represent poorly conserved inserts that are not shown. Positions with identical amino acids are in bold face. The coloring is based on the consensus shown underneath the alignment; 'h' indicates hydrophobic residues (ACFILMVWYHRK), 't' indicates turn-forming residues (ASTDNVGPERK), 'p' indicates polar residues (STEDKRNQH), 's' indicates small residues (AGCVDS), 'c' indicates charged residues (EDKRQ), '+' indicates positively charged residues (RKH). Secondary structure is shown for pdb: 1WHZ; 'H' indicates α-helix and 'E' indicates extended conformation (β-strand).

residues 336–414 as per the SCOP database] as the first hit. The VAST alignment spans the entire length of the domain, with a *P*-value of 0.0039 and RMSD of 2.2Å over 52 residues. This particular version of dsRBD in the intein has an additional α-helix at the C-terminus, a feature shared with the TTHA1913 domain (Fig. 1A). A ~50 amino acid core of the HicA family proteins precisely corresponds to the characteristic $\alpha-\beta(3)-\alpha$ structural arrangement of the dsRBD fold (Fig. 1B). With a few exceptions, the *hicA* gene is located in a predicted operon with the *hicB* gene.

2.2 HicB/COG1598/COG4226 adopts a degraded RNAse H fold which is often fused to a DNA-binding domain

Exhaustive PSI-BLAST search was also used for identification of the HicB family members (with the *H.influenzae* HicB protein [GI: 3603336] as the starting query). The majority of these proteins are included in the UPF0150 domain family in the Pfam database (Bateman *et al*., 2004) and in COG1598 or COG4226 in the COG database (Tatusov *et al*., 2003). Surprisingly, we detected almost twice as many (~450 sequences) HicB proteins than HicA family proteins, with the largest number, again, found in *C.watsonii* (33, Supplementary material). In part, this is due to the poor detection of the small *hicA* genes during genome sequence annotation. In several cases, we identified a *hicA*-like ORF in the untranslated region upstream of the *hicB* gene. However, in many other genomes, we were unable to detect the missing *hicA* gene, including several genomes in which no *hicA* genes were found whereas at least one *hicB* gene was present. This observation, together with the considerable overall excess of *hicB* genes, suggest that HicB might have functions independent of HicA. For one HicB-family protein, TTHA1013 from *T.thermophilus* HB8, the crystal structure has been solved (pdb: 1WV8) but no significant similarity to any proteins with known structures has been detected leading to the conclusion that TTHA1013 had a new fold (Hattori *et al*., 2005). The VAST program aligned the 70 residue TTHA1013 protein with domains of the RNAse H-like fold as the top hits. The RNase HII from *Methanocaldococcus jannaschii* was found as the longest hit with RMSD of 2.1 Å over 49 residues, and *P*-values for some RNAse H domains were as low as 0.0007. The alignment completely covers the TTHA1013 protein but typical RNAse H domains are longer at the C-terminus. Analysis of representatives of this fold in the SCOP database showed that RNAse H-like domains in certain alkyltransferases, e.g. O6-alkylguanine-DNA alkyltransferase from *Pyrococcus kodakaraensis* (pdb: 1MGT, residues 1–88), lack the C-terminal α-β unit and thus appear to have deteriorated in the same way as the TTHA1013 protein (Fig. 2A)—both domains have a $\beta(3)-\alpha-\beta$ core. The structure of the TTHA1013 dimer (Hattori *et al*., 2005) offers some clues as to the structural consequences of this deterioration because the C-terminal β-strands form the dimerization site.

Multiple alignment of selected HicB family representatives is shown in Figure 2B. Despite the significant structural similarity with the N-terminal portion of the RNAse H fold, none of the catalytic residues that are conserved in most nucleases of this superfamily is present in HicB. One of these catalytic residues is located in strand 5, which is missing in the HicB proteins, and two are located in strands 1 and 4 but are not retained in the HicB family (Fig. 2B). The only catalytic residue that is, mostly, conserved in the HicB family is Asp or Glu located in the α-helix (Fig. 2B). This residue is essential for catalysis and is conserved in several nuclease superfamilies of the RNAse H fold, including the bona fide RNAse H, endonuclease V, and PIWI domain (Katayanagi *et al*., 1990; Rand *et al*., 2004). In sum, the HicB proteins contain a

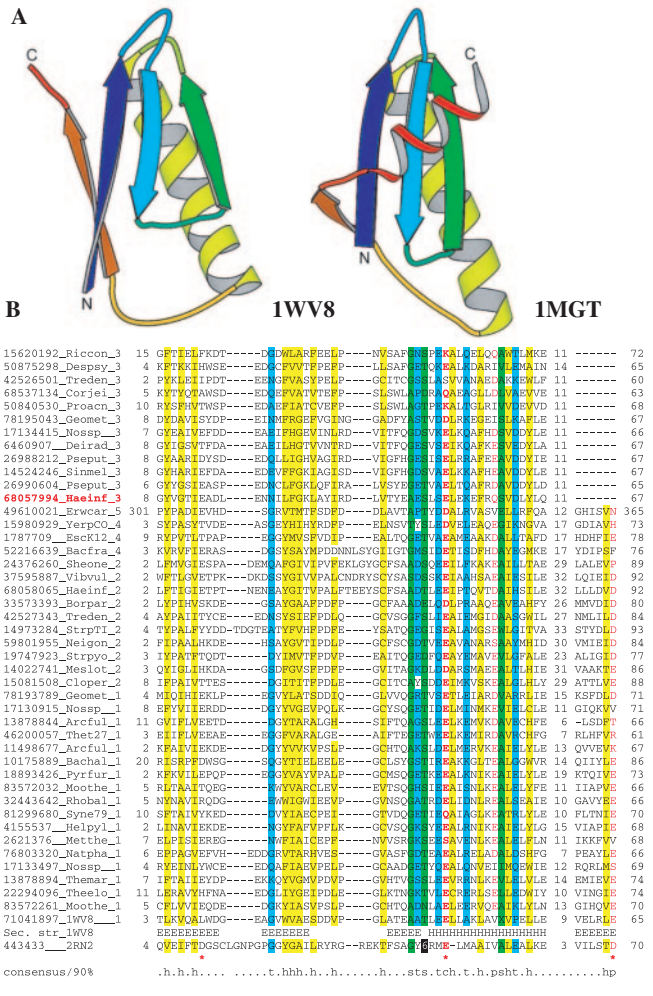

Fig. 2. (**A**) Ribbon diagram of TTHA1013 from *Thermus thermophilus* HB8 compared with the degraded RNAse H-like domain of O6-alkylguanine-DNA alkyltransferase from *Pyrococcus kodakaraensis* (residues A1–A72). Designations are the same as in Figure 1A. (**B**) Multiple alignment of the conserved core of the HicB family. Designations are the same as in Figure 1B. The HicB protein is highlighted in red. The HicB subfamilies are indicated by numbers (1–5; for details see text). Secondary structure is shown for pdb: 1WV8. The sequence of *Escherichia coli* RNAse H (443433_2RN2) is shown for comparison, with the catalytic residues indicated by asterisks.

partially degraded RNAse H fold, and it remains unclear whether or not these proteins have nuclease activity.

Further analysis of HicB-domain-containing proteins revealed at least five families with the following, distinct domain architectures (Fig. 2B). (1) Short (~70 amino acids) proteins—solo HicB domains. (2) HicB domains fused with CopG-like, ribbon-helix-helix (RHH) DNA-binding domains. In addition to the RHH domain, these proteins contain an insertion between the α-helix and β-strand 4 of the RNAse H-like domain (Fig. 2B) such that their average length is ~130 amino acids. (3) Approximately 120 aa long proteins with a distinct, C-terminal RHH domain that is only distantly related to the RHH domains of family 2 (Supplementary Fig. 1S). A specific feature of this family is that β-strand 4 of the RNAse H fold is either missing or very short. (4) HicB domains fused to helix-turn-helix (HTH) DNA-binding domains of the Xre family. (5) HicB domains fused to Shufflon-specific DNA recombinase (Rci). These proteins are encoded, mostly, in plasmids of several Enterobacteria. No specific function for this domain in the Rci recombinases has been reported but it has been proposed that it plays a role in the recognition of asymmetric sequences in recombination sites (Gyohda *et al*., 2004). The genes encoding HicB-family proteins of families 1–4 are, most often, linked to *hicA* genes; by contrast, family 5 *hicB* genes are not associated with hicA.

2.3 Occurrence of the HicAB module in prokaryotic genomes and its apparent horizontal mobility

The *hicA* and *hicB* genes are abundant in free-living archaea and bacteria (Fig. 2S in Supplementary material), with many genomes containing multiple copies of each, but are absent from the genomes of most obligate parasites and symbionts, in a pattern that is typical of TAS [(Pandey and Gerdes, 2005); Supplementary Table 1S]. Table 2S shows the numbers of identified HicA and HicB proteins encoded in selected genomes. As there is little, if any, correlation between the distribution of the five HicB families and taxonomy, and representatives of different HicB families are often found in the same genome, it appears likely that both *hicAB* cassettes and stand-alone *hicB* genes are transferred horizontally. In accord with this mode of transmission, several cassettes and stand-alone *hicB* genes are encoded in phages and plasmids that could serve as vehicles for horizontal gene transfer (HGT) (Supplementary material). Despite the abundance of *hicAB* modules in several genomes, we detected few demonstrable intragenomic duplications (e.g. MM0023 and MM0032 in *Methanosarcina mazei*). Thus, some genomes seem to have accumulated *hicAB* modules via HGT from different sources. In agreement with this notion, we observed virtually no colinearity in the localization of *hicAB* cassettes in the genomes of closely related species and even strains, and insertions of these genes into the conserved operons were commonly detected (Supplementary Fig. 3S). Thus, *hicAB* modules appear to be mobile elements prone to frequent gene rearrangement, movement within a genome, and HGT between prokaryotic species.

3 DISCUSSION

The characteristics of the *hicAB* module are similar to those of known TAS (Anantharaman and Aravind, 2003; Buts *et al*., 2005; Gerdes *et al*., 2005; Hayes, 2003; Pandey and Gerdes, 2005). Indeed, (1) HicAB is a two-component system of small proteins, (2) RHH and HTH DNA-binding domains, which are contained in most TAS, are often fused to HicB, (3) most type II TAS target RNA, and the structures of both HicA domain (dsRBD) and HicB domain (RNAse H fold) are compatible with their involvement in RNA metabolism, (4) like other TAS, the *hicAB* cassette appears to be highly mobile and prone to HGT; (5) the distribution of the *hicAB* systems in prokaryotic genomes resembles the distribution of TAS in that most free-living archaea and bacteria encode these systems, often, in multiple copies, whereas most obligate intracellular parasites and symbionts lack them, (6) similarly to other TAS, the *hicAB* module was detected in several phages and plasmids; thus, some of the HicAB systems could perform the originally discovered function of TAS, i.e. killing

plasmid- or phage-free cells. More specifically, given that the HicAB module consists of two protein components, one can predicted that it is a Type II TAS. This is compatible with the fact that all Type I TAS protein components are small membrane proteins whereas neither HicA nor HicB is. Furthermore, search of the vicinities of *hicAB* genes for potential stable, small RNAs that could function as RNA antitoxins revealed no plausible candidates (data not shown).

The mechanism of HicAB functioning remains obscure. It is not even quite clear which protein is the toxin and which is the antitoxin. Some of the available evidence suggests that HicB is the toxin. First, despite the lack of conservation of most of the canonical catalytic residues of the RNAse H fold nucleases, the possibility remains that at least some versions of HicB are RNAses, like many other toxins. Second, *hicB* genes are found in some genomes without *hicA*, which is also the case for known toxins (Mittenhuber, 1999; Pandey and Gerdes, 2005). In contrast, solitary antitoxins so far have not been detected. Furthermore, most often, in TAS operons, the antitoxin gene is located upstream of the corresponding toxin gene, and in the majority of the *hicAB* modules, *hicA* precedes *hicB*. Other observations, however, seem to contradict this hypothesis. In particular, in all known TAS, DNA-binding regulatory domains are fused to antitoxins. The mechanism of action of HicAB might be distinct from those of known TAS and could involve a complex interplay between HicA, HicB, DNA-binding domains, and a specific target.

A parallel exists between the domain architecture of the HicAB module and that of the eukaryotic RNA interference (RNAi) machinery. The essential components of the latter include at least two versions of dsRBD (in R2D2-like proteins and the dicer dsRNA nuclease) and a distinct RNAse H domain of the PIWI family in the Ago-like slicer nucleases ((Filipowicz, 2005; Tang, 2005) and references therein). These observations suggest that functional similarities between HicAB and RNAi systems might exist as well. Although we cannot currently predict the specific mechanism of HicAB toxicity, it appears most likely that it involves RNA-binding and, possibly, cleavage.

ACKNOWLEDGEMENTS

This work was supported, in part, by the Intramural Research Program of the National Library of Medicine at the US National Institutes of Health. Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Anantharaman,V. and Aravind,L. (2003) New connections in the prokaryotic toxin-antitoxin network: relationship with the eukaryotic nonsense-mediated RNA decay system. *Genome Biol.*, **4**, R81.

Bateman,A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–141.

Buts,L. *et al.* (2005) Toxin-antitoxin modules as bacterial metabolic stress managers. *Trends Biochem. Sci.*, **30**, 672–679.

Filipowicz,W. (2005) RNAi: the nuts and bolts of the RISC machine. *Cell*, **122**, 17–20.

Gerdes,K. *et al.* (2005) Prokaryotic toxin-antitoxin stress response loci. *Nat. Rev. Microbiol.*, **3**, 371–382.

Gyohda,A. *et al.* (2004) Structure and function of the shufflon in plasmid R64. *Adv. Biophys.*, **38**, 183–213.

Hattori,M. *et al.* (2005) Crystal structure of the hypothetical protein TTHA1013 from *Thermus thermophilus* HB8. *Proteins*, **61**, 1117–1120.

Hayes,F. (2003) Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. *Science*, **301**, 1496–1499.

Ichiyanagi,K. *et al.* (2000) Crystal structure of an archaeal intein-encoded homing endonuclease PI-PfuI. *J. Mol. Biol.*, **300**, 889–901.

Katayanagi,K. *et al.* (1990) Three-dimensional structure of ribonuclease H from *E.coli*. *Nature*, **347**, 306–309.

Mhlanga-Mutangadura,T. *et al.* (1998) Evolution of the major pilus gene cluster of Haemophilus influenzae. *J. Bacteriol.*, **180**, 4693–4703.

Mittenhuber,G. (1999) Occurrence of mazEF-like antitoxin/toxin systems in bacteria. *J. Mol. Microbiol. Biotechnol.*, **1**, 295–302.

Pandey,D.P. and Gerdes,K. (2005) Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res.*, **33**, 966–976.

Rand,T.A. *et al.* (2004) Biochemical identification of Argonaute 2 as the sole protein required for RNA-induced silencing complex activity. *Proc. Natl Acad. Sci. USA*, **101**, 14385–14389.

Tang,G. (2005) siRNA and miRNA: an insight into RISCs. *Trends Biochem. Sci.*, **30**, 106–114.

Tatusov,R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.