DNA phosphorothioation is widespread and quantized in bacterial genomes

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.
**DNA phosphorothioation is widespread and quantized in bacterial genomes**

Lianrong Wang\(^{a,b,c}\), Shi Chen\(^{b,c,\dagger}\), Kevin L. Vergin\(^d\), Stephen J. Giovannoni\(^d\), Simon W. Chan\(^a\), Michael S. DeMott\(^a\), Koli Taghizadeh\(^h\), Otto X. Cordero\(^i\), Michael Cutler\(^i\), Sonia Timberlake\(^e\), Eric J. Alm\(^b,\dagger\), Martin F. Polz\(^j\), Jarone Pinhasi\(^i\), Zixin Deng\(^b,c\), and Peter C. Dedon\(^a,\dagger,\#\)

\(^a\)Department of Biological Engineering, \(^b\)Center for Environmental Health Sciences, and \(^c\)Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; \(^d\)Laboratory of Microbial Metabolism and School of Life Science and Biotechnology, Shanghai Jiaotong University, Shanghai 200030, China; \(^e\)College of Pharmacy, Wuhan University, Wuhan 430071, China; \(^h\)Department of Microbiology, Oregon State University, Corvallis, OR 97331; and \(^i\)Marine Microbiology, School of Natural Sciences, Linnaeus University, SE-39182 Kalmar, Sweden

Edited* by Gerald N. Wogan, Massachusetts Institute of Technology, Cambridge, MA, and approved January 7, 2011 (received for review November 17, 2010)

Phosphorothioate (PT) modification of DNA, with sulfur replacing a nonbridging phosphate oxygen, was recently discovered as a product of the *dnd* genes found in bacteria and archaea. Given our limited understanding of the biological function of PT modifications, including sequence context, genomic frequencies, and relationships to the diversity of *dnd* gene clusters, we undertook a quantitative study of PT modifications in prokaryotic genomes using a liquid chromatography-coupled tandem quadrupole mass spectrometry approach. The results revealed a diversity of unique PT sequence contexts and three discrete genomic frequencies in a wide range of bacteria. Metagenomic analyses of PT modifications revealed unique ecological distributions, and a phylogenetic comparison of *dnd* genes and PT sequence contexts strongly supports the horizontal transfer of *dnd* genes. These results are consistent with the involvement of PT modifications in a type of restriction-modification system with wide distribution in prokaryotes.

**Results and Discussion**

**Development of a Sensitive Method to Quantify PT Modifications in Bacterial Genomes.** We approached the problem of defining the biological function of PT modifications by quantifying them and defining their sequence context. Specifically, we developed a highly quantitative electrospray ionization LC-MS/MS technique that identifies the two-nucleotide sequence context of PT modification. By using synthetic dinucleotides containing PT in the *R*<sub>2</sub> configuration (the only stereochemistry observed to date), the HPLC retention times, collision-induced dissociation molecular transitions, and limits of detection were optimized for all 16 possible dinucleotide sequence contexts for PT (Fig. 1 and Table S1). We also optimized conditions for nucleotide P1, which is inhibited by PT in the *R*<sub>2</sub> configuration, and alkaline phosphatase hydrolysis of genomic DNA to PT-containing dinucleotides and canonical nucleosides (Fig. S1), with canonical nucleosides eluting well before the PT-modified dinucleotides. Quantification was achieved by using the *s*<sub>1</sub> stereoisomer of d(GPSA) as an internal standard with multiple reaction monitoring mode of the mass spectrometer (Fig. 1). We were thus able to detect PT modifications at levels as low as 1 per 10<sup>6</sup> nt for d(TFGS) to 2 per 10<sup>9</sup> nt for d(CGPS) in 20 μg of genomic DNA (Table S1). This quantitative bioanalytical approach is a rigorous means to screen prokaryotic genomes for the sequence context and quantity of PT modifications, with immediate implications for the understanding of biological function.

**Widespread Distribution and Diverse Sequence Contexts for PT Modifications in Bacteria and Archaea.** The LC-MS/MS method was first applied to define the PT sequence contexts and quantify PT modifications in bacteria known to harbor *dnd* gene clusters. These taxonomically unrelated bacteria, including *Salmonella enterica* serovar Cerru 87, *Escherichia coli* B7A, *P. fluorescens* Pf0-1, *S. lividans* 1326, *Geobacter uraniumreducens* Rf4, *Hahella chejuensis* KCTC2396, *Bermannella marisrubri* RED65, and *Shewanella pealeana* ATCC70345, represent genera of variable clusters (**Table S1**). This quantification of biological function.

---


The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.*

To whom correspondence may be addressed. E-mail: shichen@whu.edu.cn or pcedon@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1017261108/*DCSupplemental.

---

www.pnas.org/cgi/doi/10.1073/pnas.1017261108

PNAS  |  February 15, 2011  |  vol. 108  |  no. 7  |  2963–2968
originally observed to possess d(GPSA) (Table 1) (2). In the cases of \textit{S. enterica} serovar Cerro 87 and \textit{E. coli} B7A, d(GPSA) and d(GPSA) occurred at the same level of approximately 4 PT per 10^4 nt (Table 1). However, this pairing of PT sequence contexts was not universal, with marine bacteria \textit{B. marisrubri} RED65 and \textit{H. chejuensis} KCTC2396 possessing d(GPSA) accompanied by only barely detectable d(GPSA). Both d(GPSA) and d(GPSA) were simultaneously present in \textit{G. uraniumreducens} Rf4 and \textit{S. lividans} 1326, but at levels that differed by two orders of magnitude (Table 1). Distinct from the others, \textit{P. fluorescens} Pf0-1 possesses only a single PT context, d(GPSG), at a level close to that in \textit{S. lividans} 1326 and \textit{G. uraniumreducens} Rf4. These results in bacteria known to possess \textit{dnd} genes revealed the potential for a wide range of sequence contexts for PT modifications, as might be expected for a restriction-modification system.

Further insights into PT function were gained from surveys of bacteria not previously known to possess \textit{dnd} genes. To this end, we interrogated 63 \textit{Vibrio} strains derived from an approximate 1,000-strain library of ecologically differentiated coastal \textit{Vibrio}-

...
information for these isolates revealed dnd gene homologues in
three of the strains (1F267, ZS139, and 1F230; Materials and Methods). These results both expand the repertoire of sequence contexts and broaden the range of PT levels in bacterial genomes to cover three orders of magnitude.

**PT Modifications Are Quantized in Three Discrete Frequencies.** Analysis of the quantitative data revealed that the levels of PT modifications were quantized into three distinct levels: two to three per 10^3 nt, three to eight per 10^4 nt, and one to six per 10^5 nt (Tables 1 and 2). Along with defined sequence contexts, the first two frequency ranges are consistent with a restriction-modification system (9). The highest frequency of two to three PT modifications per 10^3 nt (one PT modification in 333–500 nt) was observed in *Vibrio* species as d(CpG) (Table 2), which is consistent with a 4-nt consensus sequence, such as GGCpG or CcppGG, with a statistical frequency of once every 256 nt. Analysis of available partial genome sequence data for these strains (Table S2) reveals that the CCGG motif occurs every approximately 905 ± 130 nt and the GGCC motif every approximately 518 ± 54 nt (averages for four species), whereas the CC dinucleotide motif occurs every 24 nt. With allowance for the TA richness of the *Vibrio* genomes (Table S2), the GGCC and CCGG motifs are thus reasonable candidates for palindromic PT consensus sequences.

The other PT frequency consistent with a restriction-modification function is three to eight PT per 10^4 nt (Tables 1 and 2), which is equivalent to one PT in 1,250 to 3,333 nt or a 5–6-nt consensus sequence (~1 modification every 1,024–4,906 nt). This frequency was observed for the d(GpG) motif in *S. lividans* (Table 1) (2), in which a loose consensus sequence of 5′-cG-GCCGc-3′ (GGCC strictly conserved) was determined based on cloning of dnd phenotype break sites (10, 11). The analysis of the limited genome sequence data available for *S. lividans* 1326 reveals that the strictly conserved 4 nt GpGpGCC core sequence occurs at too high a frequency (one site per ~110 nt; Table S3) to serve as the consensus sequence for a PT-based restriction-modification system. Such is also the case for *P. fluorescens* P90-1 and *G. uraniumreducens* Rf4, in which GGCC occurs, on average, every 114 nt and 177 nt, respectively (Table S3). Therefore, a 6-nt consensus, such as CcppGpGCCG, which occurs at a frequency of 6 to 11 sites per 10^4 nt in *S. lividans* 1326, *P. fluorescens* P90-1, and *G. uraniumreducens* Rf4 (Table S3), is more consistent with the observed levels of PT modifications (three to eight per 10^4 nt).

A similar situation holds for *E. coli* B7A and *S. enterica* 87, in which we observed a 1:1 ratio of d(GpGpA) and d(GpGpT) with each PT-containing site occurring every 2,500 to 2,800 nt on average (Table 1). This frequency is too low for a 4-nt consensus sequence such as GpGpAACpGpTTC, which occurs in the related *E. coli* DH10B once in every 258 nt (Table S4), but it is consistent with a 1- to 2-nt extension of this core sequence. The situation with *B. marisrubri* RED65 and *H. chejuensis* KCTC2396, in which d(GpSpA) was observed as the predominant dinucleotide context at frequencies consistent with a 5- to 6-nt consensus (Table 1), suggests a palindromic core sequence such as GpGSpAC.

One consequence of the high sensitivity of the LC-MS/MS method is that we were able to detect PT modifications at levels well below those expected for a restriction-modification system. As shown in Tables 1 and 2, we observed two to three PT modifications per 10^6 nt in the d(GpGpT) motif in *S. lividans* 1326 and *G. uraniumreducens* Rf4 for which d(GpGpG) was the high frequency modification site, whereas d(CpGpA), d(TpGpA), and d(ApGpA) occurred at low levels (two to six per 10^6 nt) in *E. coli* expressing the *S. enterica* dnd genes (Table 1) and d(ApGpG) and d(TpGpC) were minor sites observed in *Vibrio* isoles in Table 2.

There are two explanations for the low PT levels: a function for PT other than restriction-modification, as is the case for DNA methylation in many prokaryotes (9, 12); and biochemical nonspecificity for the Dnd protein responsible for target selection. The latter would be similar to a restriction enzyme cleaving at a thermodynamically or kinetically disfavored site (i.e., secondary cleavage sites), or cleaving with altered sequence specificity as a result of salt- or pH-induced alterations in protein–DNA interactions (i.e., “star activity”) (9). As one test of the latter hypothesis, we compared the levels of d(GpGpT) and d(GpGpA) in *E. coli* DH10B harboring low and high copy number plasmids containing the *S. enterica* 87 dnd gene cluster, pJTU1980 and pJTU1238, respectively. The high copy number vector produced a 16-fold increase in dndC transcription (Table S5). As shown in Table 1, the increased expression of dnd genes caused PT modifications in d(GpGpT) and d(GpGpA) to increase 1.5- and 2-fold, respectively, in comparison with *S. enterica* 87 (Table 1). This is consistent with relaxation of a strict modification consensus sequence or an increase in the proportion of a consensus sequence population that becomes modified with a PT. Support for the former model comes from the appearance of three low-frequency PT modifications in d(GpGpA), d(TpGpA), and d(ApGpA) in direct proportion with dnd gene expression (Table 1). The results support the hypothesis that low-frequency PT modifications result from a degree of relaxed DNA target recognition by Dnd proteins.

**Phylogenetic Analysis of dnd Genes and PT Sequence Contexts Is Consistent with Horizontal Gene Transfer.** In addition to the quantitative PT levels, support for the involvement of PT modifications in a restriction-modification system comes from an apparent association of PT sequence contexts with phylogenetic relationships drawn from Dnd protein sequences but not species phylogeny. There is strong evidence for the distribution of classical methylation-based restriction-modification systems (13) and for the distribution of dnd gene clusters (14) by horizontal gene transfer and mobile genetic elements, as opposed to the vertical gene transfer associated with species phylogeny. As shown in Fig. 2, there is a strong correlation between the Dnd protein sequence phylogeny and the distribution of PT sequence modifications, with the exception of outlier Candidatus Pelagibacter ubique. The fact that the Dnd phylogenies do not follow their corresponding species tree suggests that PT sequence context is dependent on Dnd protein sequence and not on the phylogenetic descent of the strains. This is most clearly seen in the three *Vibrio* isolates (ZS139, 1F230, and 1F267), which are phylogenetically incoherent in all four Dnd proteins. The phylogenetic differentiation of the *Vibrio* isolates also suggests horizontal gene transfer of the whole dnd cluster. The fact that this split is matched by a corresponding switch in PT sequence context is a strong indication of the dependency between dnd genes and PT contexts.

**Insights into Ecological Distributions of PT Modifications and dnd Genes by Metagenomic Analysis of Ocean Genomes.** The widespread distribution of dnd genes and PT modifications is further illustrated by a metagenomic analysis of ocean bacteria (15). To
initiate these studies, we identified a dnd gene cluster homologue in C. Pelagibacter ubique strain HTCC1002, one of the smallest known free-living bacteria and the first cultured member of the alphaproteobacterial SAR11 clade. SAR11 is a ubiquitous group of marine bacteria that can account for as much as 35% of bacterioplankton populations in the ocean surface (16). The SAR11 dnd gene cluster was located in a hypervariable region of the genome and was not found in closely related strain HTCC1062, which is consistent with our observation of PT only in strain HTCC1002 as d(GPSA) and d(GPSG) modifications.

Having established a PT modification in strain HTCC1002, we used tBLASTn with the HTCC1002 dndBCDE genes as queries to detect the distribution of dnd genes in oceanic metagenomes. We observed significant hits (E < 10^{-20}) in several oceanic metagenomes (Table S6): 136 reads contained dndB, 123 reads contained dndC, 80 reads contained dndD, and nine reads contained dndE. In particular, one of the most significant hits was detected in the metagenome of Sargasso Sea, a low-nutrient, low-productivity, subtropical ocean gyre. DNA samples collected from depths to 200 m in the Sargasso Sea revealed natural DNA PT contexts of d(CPSG) and d(GPSG) in all samples, with d(GPSG) and d(GPSG) occurring mainly at lower depths (Table S3). A similar analysis of DNA samples collected from 5 m, 20 m, and 40 m in the highly productive, temperate Oregon coastal waters revealed the presence of four major PT contexts: d(CPSG), d(GPSG), d(GPSG), and d(GPSG). Interestingly, d(CPSG) and d(GPSG) were detected throughout the water columns off the Oregon coast (5–40 m) and the Sargasso Sea (0–200 m), whereas d(GPSG) was detected only in deeper zones of the water column in both locations. Dramatic transitions in the composition of microbial communities as a function of depth have been documented in previous studies (17). These results suggest that d(GPSG) is associated with microbial taxa that occupy the dark mesopelagic ocean region beneath the euphotic zone, which may have implications for horizontal gene transfer among microbes occupying the various ocean communities.

In summary, these results provide insights into the function of PT modifications in bacterial genomes. Our LC-MS/MS approach to studying the only defined chemical modification of the DNA backbone, with potential application to the study of putative arsenic-modified microbial nucleic acids (18), provides a rich source of information that complements genetic and molecular approaches to defining biological function. The data reveal quantized levels of PT, which suggest involvement of PT in a new restriction-modification system. This is consistent with our recent observation that a dptF-H cluster adjacent to dnd genes in S. enterica serovar Cerro 87 restricts the uptake of plasmids without PT (19) and with studies showing inhibition of methylation-based restriction enzymes by PT modifications (20, 21). The observations of widespread ecological distribution and strong phylogenetic relationships of dnd genes and...
PT modifications demonstrate the importance of PT modifications in prokaryotic physiology.

Materials and Methods

Materials, Bacterial Strains, and Plasmids. Enantiomerically pure, PT-containing dinucleotides in $R_o$ or $S_o$ configuration were obtained from IBA Biotechnology. The bacteria strains harboring natural sets of $dnd$ clusters and DNA samples were gifts from different laboratories and institutions as noted in the Acknowledgments. Plasmid pJTU1238 was a derivative of high copy plasmid pBluescript II SK+ containing $dnd$ gene cluster from S. enterica 87 (2). A KpnI-XbaI fragment containing the $dnd$ gene cluster from pJTU1238 was cloned into pBl2925 to yield pJTU1976, from which the KpnI-BglII fragment harboring the $dnd$ genes was inserted into the KpnI- and BamHI-treated low-copy plasmid pACYC184 to yield pJTU1980. We have also characterized the PT modifications in a set of well characterized Vibrionaceae strains (8) for which partial genome sequence information has been obtained.

E. coli B7A is an enterotoxigenic strain isolated from an American soldier suffering from diarrhea during the Vietnam War (22) and represents a frequent etiologic agent for short-incubation travelers diarrhea and endemic infantile diarrhea (23). S. enterica 87 was isolated from a commercial egg-producing farm (24). Marine bacteria H. chejuensis KCTC2396, B. marisrubri RED6S, and S. pealeana ATCC700345 were isolated from marine sediment of Cheju Island in Korea, surface seawater from Gulf of Ellat in Red Sea, and the accessory nodulamental gland of the squid Loligo pealei, respectively (25–27).

DNA was extracted using the sucrose lysis method followed by phenol extraction and ethanol precipitation (31). DNA was further purified by treatment with RNase and extraction by using a Qiagen DNeasy Blood and Tissue Kit. Strain identities were verified by amplification and sequencing of the oxireductase gene as described previously (32).

Controlled Enzymatic Digestion of PT Modified DNA. The first step in the development of the LC-MS/MS technique to quantify PT modifications involved nuclelease P1-mediated hydrolysis of PT-containing DNA to a limit digest of nucleosides and PT-bridged dinucleotides, as observed in earlier studies (2). Optimal parameters and MS detection parameters for resolving and quantifying the PT-containing dinucleotides. To this end, we defined the HPLC retention times for the set of all possible 16 PT-linked dinucleotides by using a Thermo Hypersil GOLD aQ column (150 × 2.1 mm, 3 μm) with elution conducted at 35 °C and a flow rate of 0.3 mL/min, with a gradient of 97% buffer A (0.1% acetic acid in water) and 3% buffer B (0.1% acetic acid in acetonitrile) for 5 min, followed by 3% to 6% buffer B over a period of 30 min and 6% to 98% buffer B over a period of 1 min. Retention times are noted with other parameters later, with canonical nucleosides eluting as follows: dc, 2.2 min; da, 3.8 min; dg, 3.9 min; and dt, 5.0 min, the dDA(A) S5, ret ions eluted at 28.1 min. The HPLC column was coupled to an Agilent 6410 Triple Quad LC/MS mass spectrometer with an electrospray ionization source in positive mode with the following parameters: gas flow, 10 L/min; nebulizer pressure, 25 psi; drying gas temperature, 300 °C; and capillary voltage, 3,100 V. Multiple reaction monitoring mode was used for detection of product ions derived from the precursor ions, with all instrument parameters optimized for maximal sensitivity (retention time in min, precursor ion m/z; product ion m/z, fragmentor voltage, collision energy): d([G]G), 10.3, 573, 152, 23 V, 25 V; d([G]C), 13.3, 533, 112, 123 V, 25 V; d([G]G), 19.5, 613, 152, 123 V, 29 V; d([G]A), 16.1, 557, 136, 126 V, 29 V; d([G]T), 18.7, 548, 112, 110 V, 13 V; d([A]G), 18.8, 597, 136, 120 V, 40 V; d([A]G), 20.5, 597, 136, 120 V, 40 V; d([T]G), 20.9, 588, 152, 117 V, 17 V; d([A]G), 24.5, 573, 112, 129 V, 25 V; d([A]T), 20.9, 585, 152, 117 V, 17 V; d([A]G), 27.3, 581, 136, 117 V, 33 V; d([T]G), 29.5, 572, 136, 125 V, 20 V; d([A]C), 30.3, 557, 112, 117 V, 25 V; d([T]C), 30.8, 548, 112, 117 V, 13 V; d([T]A), 31.1, 572, 136, 125 V, 20 V; and d([T]T), 33.5, 563, 127, 110 V, 37 V.

RNA Preparation and Real-time Quantitative PCR. To assess the expression of dnd genes, total RNA was isolated by using a Qiagen RNeasy Protect Bacteria Mini Kit and 15 ng of RNA was used as template for real-time PCR performed with the Power SYBR Green RNA-to-CT Kit (Applied Biosystems) and an Applied Biosystems 7900HT Fast real-time PCR system. To measure the transcription of $dnd$ cluster in DH108(pJTU1238) and DH108(pJTU1980), primers were designed within dndC gene. The housekeeping gene gapA, which codes for α-GAPDH, was used as reference. Primers 5′-ATTGTTGGCCTGGTGGC3′ and 5′-CGGCGCGGAGCCGGCGAG-3′ were used to amplify dnd genes; primers 5′-CCGATGCGCGGCGCGCGG3′ and 5′-CTCGGTCCGCTTGGCGAGG3′ were used to amplify gapA gene. RT-PCR data analysis was performed according to the comparative threshold cycle method, also known as $2^{ΔΔC_T}$.

Partial Sequencing of Vibrionaceae Genomes. Each genome was sequenced utilizing the Illumina platform with an independent lane of sequence for each, yielding approximately 12 million 76-bp single-end reads per genome. We removed reads consisting containing Ns and trimmed poor-quality read termini with Euler qualityTrimmer (using default parameters) (33). The Velvet software package was used for de novo assembly of reads into contigs (34). Parameters were optimized to yield the most contiguous (i.e., highest N50) assembly, independently for each strain. N50 values ranged from 3 to 200 kb, with an average strain reaching an N50 of 40 kb and N80 of 20 kb.

Analysis of Bacterial Genomes. Bacterial genomes were analyzed for sequence motifs by using the DiProGB genome browser with sequences obtained from the National Center for Biotechnology Information (NCBI) Genome database. Complete or partial genome sequences were uploaded into the browser and the number of sequence motifs determined for both genomic DNA strands (Table S2). Dnd protein sequences for 12 strains analyzed here were retrieved from the NCBI Protein database and aligned by using Muscle (35) and phylogenetic trees were computed with phyML (36), with phylogenies displayed using iTol (37).

ACKNOWLEDGMENTS. The authors thank Dr. Evgenya Shoblombolina for providing G. uraniuromedicus Rf4, Dr. Jian-Shen Zhao for S. pealeana ATCC700345, Dr. Hong Kum Lee for H. chejuensis KCTC2396, Dr. Eduardo Robledo for P. fluorescens PM0-1, Dr. Toshiyuki Murase for S. enterica serovar Cerro 87, Joshua Kitner for culturing the SAR11 cells, Anthony Bertagnolli and the crew of the R.V. Elakha for collecting and processing Oregon Coast DNA samples, Rachel Parsons and the crew of the R.V. Weatherbird II for collecting and processing BATS samples, and Dr. H. James Tripp for assistance with the hypervariable region analysis. Chromatography and mass spectrometry were performed in the Bioanalytical Facilities Core of the Massachusetts Institute of Technology Center for Environmental Health Sciences, supported by National Institute of Environmental Health Sciences Grant ES002109. This work was supported by National Science Foundation Grant CHE-1019990.