Genome analysis

DomainSieve: a protein domain-based screen that led to the identification of dam-associated genes with potential link to DNA maintenance

Pierre Brézellec², Mark Hoebbeke², Marie-Sophie Hiet¹,², Sophie Pasek² and Jean-Luc Ferat¹,*
¹Centre de Génétique Moléculaire du CNRS, 1 Avenue de la Terrasse, 91190 Gif sur Yvette, France and ²Laboratoire Statistique et Génome du CNRS, Tour Evry 2, 523, Place des Terrasses de l’Agora, 91034 Evry Cedex, France

Received on May 12, 2006; revised on June 9, 2006; accepted on June 13, 2006
Advance Access publication June 20, 2006
Associate Editor: Alex Bateman

1 INTRODUCTION

In Escherichia coli, Dam methyltransferase (DamMT)-dependent methylation of DNA is actively recognized and discriminated during the course of the cell cycle. E. coli cells defective in DamMT activity display defects in replication initiation, chromosome partition, cell division and mismatch repair [reviewed in Lobner-Olesen et al. (2005)]. In addition, DamMT has been shown to control gene expression (Oshima et al., 2002) and is known to be a pathogenicity factor in some organisms (Low et al., 2001).

In E. coli, the methylation of DNA is a post-replication step. The newly synthesized strand of DNA remains unmethylated for a short period of time, in the range of minutes. Two loci on the chromosome (the origin of replication of the chromosome, oriC, and the promoter region of dnaA), however, exhibit a much slower kinetics of remethylation, on the order of one-third of the cell cycle (Campbell and Kleckner, 1990). The delay of remethylation at oriC and dnaA results from a strong and specific interaction of the protein SeqA with clusters of hemimethylated GATC sequences (Slater et al., 1995), thought to insure that replication is initiated once and only once per cell cycle. Despite the critical function of SeqA in replication initiation control, orthologs of the gene specifying this protein have been identified only in a few orders of the gammaproteobacteria (the Enterobacteriales, the Pasteurellales, the Vibrionales and the Alteromonadales) but not in closely related genomes such as the ones belonging to the Pseudomonadales, the Legionellales and the Xanthomonadales.

In addition to its contribution to the control of replication initiation, the recognition of the methylation status of the DNA, could conceivably be a valuable sensor for the control of other steps of the cell cycle. Possible roles include, assessing the position of the polymerase during an on-going round of replication, participating in the control of the multiple steps associated with replication termination (chromosome dimer resolution, decatenation of the daughter chromosomes, etc.), and even coordinating replication with cell division. If such methylation sensing activities exist, i.e., are related to the partioning of the origin of replication of the chromosome to control the timing of replication initiation.

Interestingly, seqA is only present in a few DamMT-specifying proteobacteria. This observation led us to hypothesize that other genes, specifying related functions, might also be found in these organisms. To test this hypothesis, we implemented a large-scale comparative genomic screen meant to identify genes specifying DNA methylation sensing domains, probably involved in DNA maintenance functions.

We show here that the Dam proteins fall into two distinct phylogenetic groups. One group is monophyletic and congruent with the phylogeny of the 16S rRNA, indicating that the genes
were vertically inherited (the ‘resident’ dam genes) while the other group gathers Dam proteins whose distribution is that expected for mobile genetic elements. We established that seqA is systematically and strictly found in the one group of dam that is monophyletic.

In order to identify genes such as seqA that are systematically associated with ‘resident’ dam genomes, we developed a bioinformatic screen. This bioinformatic screen, available as a web resource called ‘DomainSieve’ (http://stat.genopole.cnrs.fr/ds), works with Pfam-A protein domains (Bateman et al., 2004). It collects protein domains that are systematically present in a set of ‘in-group’ organisms, and excludes those that are also identified in at least one organism of the ‘out-group’ set of organisms. Using a set of ‘resident’ dam genomes (in-group) and a set of genomes lacking a bona fide ‘resident’ dam gene (out-group)—both of which have been defined phylogenetically, our screen retained 18 candidates, including the genes seqA, mutH, mukB, mukE and mukF, which specify factors involved in DNA maintenance by way of recognizing the methylation status of the DNA. In addition, we identified 11 candidate genes of unknown function, among which some are likely to specify new activities involved in DNA maintenance.

We believe that this approach is generally applicable to a large variety of biological questions. In particular, DomainSieve might happen to be an additional and valuable tool to investigate issues as diverse as aerobies versus anaerobies, respiration versus fermentation, pathogenicity versus non-pathogenicity, etc.

2 METHODS

2.1 Datasets

Genomes used in this study are as follows. The date of release at the NCBI is under bracket. It helped us to constitute the sets of genomes that were analyzed to test the robustness of DomainSieve (the 13 dam and the 7 non-dam genomes used for the screen presented here are underlined).


2.2 Pfam annotations

The Pfam annotations pertaining to the above-mentioned bacteria were downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/database-files. The screen has been performed with the release 20.0 (corresponding to Swiss-Prot release 48.1 and SP-TrEMBL release 31.1). Pfam is a database of multiple alignments of protein domains or conserved protein regions. In our study, we used only Pfam-A domains, i.e. protein families whose alignments have been individually validated.

2.3 DomainSieve

Algorithm. Given a set of ‘in-group’ organisms (i.e. sharing a same feature; in the present case, organisms that contain a copy of the ‘resident’ dam gene) and a set of ‘out-group’ organisms (i.e. organisms, related or not, that do not contain the feature shared by the ‘in-group’ organisms), our screen is divided in five steps:

1. Associate to each organism belonging to the ‘in-group’ its own set of Pfam-A domains. Then, intersect all the computed sets and collect the intersected domains in one set (in-group domain set).

2. Associate to each organism of the ‘out-group’ its own set of Pfam-A domains. Then, regroup all domains in one set (out-group domain set).

3. Keep and display the domains that belong solely to the in-group organism, i.e. remove from the in-group domain set each domain that appears also within the out-group domain set.

4. Choose an organism from the in-group.

5. Display the proteins that contain at least one of the selected domains with the nomenclature of the chosen organism.

Web resource. Using DomainSieve first requires the selection of a set of in-group and a set of out-group organisms. For this purpose, a user friendly interface is available. A clickable phylogeny of the main orders of bacteria as well as a set of labels—which captions represent the main phyla of bacteria—are displayed [taxonomy according to Garrity et al. (2004)]. Clicking on the leaves of the phylogeny (or on the ‘phylum labels’) displays boxes allowing the user to add organisms (or complete orders) to the ‘in-group’ (green color) or ‘out-group’ (red color); unselected organisms (or orders) remain blue. Submitting the query, through an independent window, returns a list of domains, each of which is associated to a ‘checkbox’, allowing subsequent removal of irrelevant domains. The user may pick an organism belonging to the ‘in-group’ to display the set of proteins from this organism, containing at least one of the selected domains. To date, our web resource hosts 228 sets of Pfam-A domains corresponding to 228 bacteria.

2.4 Phylogenetic analysis

A crude alignment of the protein sequences, generated using the program ClustalW, was refined by hand before being fed to the SEQBOOT program in order to generate multiple datasets for bootstrap score calculation. From these datasets distance matrices were calculated by PROTDIST and trees were obtained by running the program CONSENSE and fed as input to PROML.

Using the nomenclature of the chosen organism.

Downloaded from http://bioinformatics.oxfordjournals.org by guest on May 31, 2013
their degree of similarity to the query, whereas sequences with a higher \( E \)-value are numerous and rather uniformly divergent from the query. The Sequences with \( E \)-value \(< 10^{-20} \) (dashed line) were retained for further analysis.

3 RESULTS

3.1 Setting up the input of the screen

While the distribution of seqA is restricted to a few orders of the proteobacteria, \textit{dam} is widely distributed within microbial genomes. Yet, the functions, in which DamMT activity is involved, are more diverse than those of SeqA. Thus, we investigated the phylogeny of DamMT in order to determine whether the profile of the protein involved in DNA maintenance could be clearly distinguished from others.

The phylogeny of Dam sequences (Fig. 1) reveals that \textit{dam} genes from the gammaproteobacteria form a separate subtree (together with two sequences belonging to the betaproteobacteria) and that within that subtree, two sets of sequences can be distinguished. The first set coincides with a distally lying clade with strong bootstrap support. Within that clade, sequences from Enterobacteriales, Vibrionales, the Pasteurellales, and the Enterobacteriales, the Alteromonadales, and the Aeromonadales is monophyletic and congruent with that of 16S rRNA. Inset, plot of the cumulative number of protein sequences as a function of their \( E \)-values (Methods). The arrow points to a breakpoint (~\( 10^{-20} \)) within the distribution (marked by a thick bar in the tree). The scale bar represents the number of substitutions per sequence position. When known, the genetic shuttle of \textit{dam} gene is indicated, \([\alpha]\) bona fide genome sequence, \([\phi]\) phage or prophage genome, \([\pi]\) plasmid-borne: \textit{Serretia enterica}.1 (Paratyphi); \textit{S. enterica}.2, (CT18, Ty2); \textit{S. enterica}.3, (CT18, Paratyphi, Ty2); \textit{S. enterica}.4, (CT18, Ty2); \textit{L. pneumophila}.1, (str. Philadelphia 1, str. Paris, str. Lens); \textit{L. pneumophila}.2, str. Philadelphia 1. (a), Archaea; (c), Cyanobacteria; (f), Firmicutes. A complete listing is provided in the Supplementary Materials section.

### Fig. 1

Unrooted phylogenetic tree showing the relationships of Dam methyltransferase sequences (Methods). AER, Aeromonadales; ALT, Alteromonadales; ENT, Enterobacteriales; PAS, Pasteurellales; VIB, Vibrionales. The solid vertical line separates Dam sequences into two groups: on the right side (gamma), Dam sequences that were identified in gammaproteobacteria (note the occurrence of two Dam proteins in betaproteobacteria; \( b \)) and on the other side the Dam sequences identified elsewhere (other). ‘resident’ Dam regroups the sequences whose distribution within the Enterobacteriales, the Vibrionales, the Pasteurellales, the Alteromonadales and the Aeromonadales is monophyletic and congruent with that of 16S rRNA. Inset, plot of the cumulative number of protein sequences as a function of their \( E \)-values (Methods). The arrow points to a breakpoint (~\( 10^{-20} \)) within the distribution (marked by a thick bar in the tree). The scale bar represents the number of substitutions per sequence position. When known, the genetic shuttle of dam gene is indicated, \([\alpha]\) bona fide genome sequence, \([\phi]\) phage or prophage genome, \([\pi]\) plasmid-borne: \textit{S. enterica}.1 (Paratyphi); \textit{S. enterica}.2, (CT18, Ty2); \textit{S. enterica}.3, (CT18, Paratyphi, Ty2); \textit{S. enterica}.4, (CT18, Ty2); \textit{L. pneumophila}.1, (str. Philadelphia 1, str. Paris, str. Lens); \textit{L. pneumophila}.2, str. Philadelphia 1. (a), Archaea; (c), Cyanobacteria; (f), Firmicutes. A complete listing is provided in the Supplementary Materials section.
Pfam domains were arranged into two categories depending on their distribution in bacterial genomes. Restricted identifies Pfam domains that were almost exclusively identified in dam\textsuperscript{res} genomes (see below for exceptions) whereas Broad gathers domains that were largely distributed outside this group of genomes (we excluded the distribution of these domains within the Alteromonadales as a valid criterion). Pfam domain-containing genes are indicated with their size in amino acids. The size of the domains (in units) or to the activity of the domain (function) is also provided.

Information relative to the number of Pfam domains per gene (units) or to the activity of the domain (function) is also provided.

The gene product recognizes, or is sensitive to, DNA methylation. Shares homology with CHY\_1718 of Carboxythermus hydrogenoformans Z-2901. ENSANGP000000027933 of Anopheles gambiae str. PEST. DBDB0188916 of Dictyostelium discoideum.\textsuperscript{3} with Noc\_2831 of Nitrosococcus oceanis ATCC19707, lpp1436 of L. pneumophila str. Paris, BL0563 of Bifidobacterium longum NCC2705, sau3A IR of Staphylococcus aureus, Sau3A I-like of Listeria monocytogenes\textsuperscript{4} BdB3688 of Bdellovibrio bacteriovorus HD100 and DDB0231295 of Dictyostelium discoideum.\textsuperscript{5} THERM\_00266730 of Tetrahymena thermophila SB120, FN0411 of Fusobacterium nucleatum subsp. nucleatum ATCC 25586.\textsuperscript{6} with ENSANGP00000001137 of Anopheles Gambiae str. PEST, F22D16.16 of Arabidopsis thaliana, LOC416802 of Gallus gallus, GSTEM\_00029441.G:001 of Tetrasomus nigroviridis, B2A19.050 of Neurospora crassa.\textsuperscript{7} corresponds to the actual size of the domain (in amino acid) within the gene that is underlined.

Table 1. pfam domains associated with ‘resident’ dam genomes

<table>
<thead>
<tr>
<th>Genes</th>
<th>Amino acid</th>
<th>Units</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restricted</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF03603</td>
<td>hsdD</td>
<td>137</td>
<td>DNA pol III loader</td>
</tr>
<tr>
<td>PF01340</td>
<td>metJ</td>
<td>105</td>
<td>Met Apo-repressor</td>
</tr>
<tr>
<td>PF04310\textsuperscript{a}</td>
<td>mukB\textsuperscript{a}</td>
<td>1486</td>
<td>Chromosome partitioning</td>
</tr>
<tr>
<td>PF04288\textsuperscript{b}</td>
<td>mukE\textsuperscript{b}</td>
<td>225</td>
<td>Chromosome partitioning</td>
</tr>
<tr>
<td>PF03882\textsuperscript{c}</td>
<td>mukF\textsuperscript{c}</td>
<td>440</td>
<td>Chromosome partitioning</td>
</tr>
<tr>
<td>PF02976\textsuperscript{d}</td>
<td>mumH\textsuperscript{d}</td>
<td>229</td>
<td>DNA mismatch repair</td>
</tr>
<tr>
<td>PF03925</td>
<td>seqA\textsuperscript{e}</td>
<td>181</td>
<td>Replication initiation, etc.</td>
</tr>
<tr>
<td>PF06062</td>
<td>yacL</td>
<td>136</td>
<td>Unknown</td>
</tr>
<tr>
<td>PF06303</td>
<td>ycbG</td>
<td>150</td>
<td>Unknown</td>
</tr>
<tr>
<td>PF06305</td>
<td>ycbS</td>
<td>102</td>
<td>Unknown</td>
</tr>
<tr>
<td>PF04269</td>
<td>yciU</td>
<td>135</td>
<td>Unknown</td>
</tr>
<tr>
<td>PF06185\textsuperscript{f}</td>
<td>yecM\textsuperscript{f}</td>
<td>196</td>
<td>Unknown</td>
</tr>
<tr>
<td>PF04393\textsuperscript{g}</td>
<td>yeeX\textsuperscript{g}</td>
<td>131</td>
<td>Unknown</td>
</tr>
<tr>
<td>PF07208\textsuperscript{h}</td>
<td>yejL\textsuperscript{h}</td>
<td>75</td>
<td>Unknown</td>
</tr>
<tr>
<td>PF04217</td>
<td>yfbV</td>
<td>151</td>
<td>Unknown</td>
</tr>
<tr>
<td>PF04219</td>
<td>yifE</td>
<td>112</td>
<td>Unknown</td>
</tr>
<tr>
<td>PF04220\textsuperscript{i}</td>
<td>yihI\textsuperscript{i}</td>
<td>169</td>
<td>Unknown</td>
</tr>
<tr>
<td>PF04222</td>
<td>yjuG</td>
<td>196</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Broad</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF01182</td>
</tr>
<tr>
<td>PF08211</td>
</tr>
<tr>
<td>PF00939</td>
</tr>
<tr>
<td>PF04626</td>
</tr>
<tr>
<td>PF03605</td>
</tr>
<tr>
<td>PF07840</td>
</tr>
<tr>
<td>PF02300</td>
</tr>
<tr>
<td>PF02313</td>
</tr>
<tr>
<td>PF01228</td>
</tr>
<tr>
<td>PF02664</td>
</tr>
<tr>
<td>PF03205</td>
</tr>
<tr>
<td>PF02901</td>
</tr>
<tr>
<td>PF0317</td>
</tr>
<tr>
<td>PF06005</td>
</tr>
<tr>
<td>PF05127</td>
</tr>
</tbody>
</table>

Odds with that of 16S rRNA, indicating that their evolution is independent of their host genomes: they constitute the set of non-resident dam genes. Although their behavior is typical of that of mobile elements, we believe that the close association of these dam genes with various shuttles, rather than the activity they specify, is responsible for their mobility. A single genome may harbor both types of dam genes and it is striking that all pathogenic strains contain a non-resident dam gene, consistent with previous data linking Dam activity to pathogenesis (Low et al., 2001).

The status of the dam genes from L. pneumophila is somewhat ambiguous. These sequences lie half-way between the sets of non-resident and resident genes. One copy (Lp.2) is only present in the strain Philadelphia 1 while the other copy (Lp.1), identified in all three strains sequenced to date, is of prophage origin. Moreover, a closely related dam gene was identified in strain BF13, but not in strain Z2491 of Neisseria meningitidis (a bacterium belonging to the beta proteobacteria), consistent with a recent horizontal transfer. Therefore, we chose to group this branch together within the set of non-resident dam genes. Finally, protein sequences more distantly related to E.coli Dam (Fig. 1, left part of the tree) are widespread in genomes of gram\textsuperscript{+}, gram\textsuperscript{-} and cyanobacteria, as well as archaea. Many of these dam genes are associated with a
restriction endonuclease gene as part of a defense system, and they may also be carried by plasmids and prophages.

From this phylogenetic analysis, we defined two sets of genomes based on the presence or the absence of a $\text{dam}^{ab}$ gene. The $\text{dam}^{ab}$ set groups the Enterobacteriales, the Vibrionales and the Pasturellales. We restricted the group of genomes in which $\text{dam}$ is not resident, to those belonging to gammaproteobacteria, i.e. members of the Pseudomonadales, Legionellales and Xanthomonadales (Methods). We excluded the Alteromonadales from our analysis since this order contains genomes with ($\text{dam}^{ab}$)-containing genomes within the Alteromonadales. The phylogenetic tree is based on Dam protein sequences (branch length is not scaled). (+) or (−) indicate the presence or the absence of the gene (top) in Shewanella oneidensis, in Idiomarina loihiensis, in Pasteurellales (PAS), in Enterobacteriales (ENT) or in Vibrionales (VIB).

3.2 Genes related to DamMT are picked up through the screen

The 13 $\text{dam}^{ab}$-containing (in-group) and the 7 $\text{dam}^{ab}$-lacking genomes (out-group) (Methods) were processed through DomainSieve in order to establish the list of domains strictly associated with the $\text{dam}^{ab}$-containing genomes. The domains systematically associated with the genomes of the $\text{dam}^{ab}$ set amounted to 867. During the second step, any domain present in at least one of the 7 $\text{dam}^{ab}$-lacking genomes (Methods) was discarded, which reduced the set of provisional $\text{dam}^{ab}$-associated Pfam domains to 33, or 42 genes, since a given domain may be shared by several genes (Table 1). Finally, each of the remaining domains was subjected to a BLAST search against the non-redundant database stored at the NCBI.

The 15 domains that were found to be distributed outside of the gammaproteobacteria (broad in Table 1) were discarded at that stage, which left just 18 $\text{dam}^{ab}$ domains and as many genes (restricted in Table 1).

It should be noted that the 18 $\text{dam}^{ab}$-associated domains are variably distributed within the Alteromonadales, an order which was excluded from our analysis because of the heterogeneous distribution of the $\text{dam}^{ab}$ gene (Fig. 2). Some domains were identified in all, or at least several, Alteromonadales (domains corresponding to $\text{metJ}$, $\text{mutH}$, $\text{seqA}$, $\text{yacL}$, $\text{yecM}$, $\text{yelL}$, $\text{yfbV}$, $\text{yifE}$, $\text{yihl}$ and $\text{yjaG}$), while the others were never found in those genomes (domains corresponding to $\text{holD}$, $\text{mukB}$, $\text{mukE}$, $\text{mukF}$, $\text{ycbG}$ and $\text{yciU}$). This situation may reflect either a late acquisition of the genes $\text{holD}$, $\text{mukB}$, $\text{mukE}$, $\text{mukF}$, $\text{ycbG}$, $\text{yciS}$ and $\text{yciU}$/ within the common ancestor of the Enterobacteriales, the Pasteurellales and the Vibrionales or result from secondary losses in the Alteromonadales. Yet, we did not consider the presence of these genes within the Alteromonadales as a valid criterion, and decided to retain the 18 $\text{dam}^{ab}$-associated genes together as a single set.

Seven genes ($\text{holD}$, $\text{metJ}$, $\text{mukB}$, $\text{mukE}$, $\text{mukF}$, $\text{mutH}$ and $\text{seqA}$) out of the 18 retained by our screen have already been characterized. In addition to $\text{seqA}$, no fewer than four of those were shown to recognize directly or indirectly the methylation status of DNA. $\text{MutH}$ is an endonuclease associated with the mismatch repair system that recognizes and cleaves specifically the unmethylated strand at GATC heteroduplexes (Welsh et al., 1987). The three proteins specified by the operon $\text{muk}$ form a complex involved in chromosome partitioning (Yamazoe et al., 1999). Although $\text{MukB}$-directed sister chromosome cohesion seems not to be affected in a $\text{dam}$-mutant (Sunako et al., 2001), it has been shown that a $\text{dam}$ null allele suppresses the thermosensitivity associated with a $\text{mukB}$ null mutant (Onogi et al., 2000), suggesting an indirect link between the Muk complex and the methylation status of DNA. No connection was established, hitherto, between $\text{dam}$ and either $\text{holD}$ or $\text{metJ}$.

The function of the remaining 11 genes ($\text{yacL}$, $\text{ycbG}$, $\text{yciS}$, $\text{yciU}$, $\text{yecM}$, $\text{yeyX}$, $\text{yelL}$, $\text{yfbV}$, $\text{yjaG}$, $\text{yifE}$ and $\text{yihl}$) is unknown. The genomic context of each gene was checked for the presence of potential transcription and translation initiation signals upstream of the coding sequence. The $\text{ycbG}$ and $\text{yifE}$ coding sequences are oriented oppositely with respect to surrounding genes. In each instance, there exists downstream of the coding sequence a stretch of nucleotides that could fold into a structure typical of rho-independent transcription terminators, suggesting that the genes are expressed. In addition, the expression of $\text{yjaG}$ has been authenticated by a direct expression assay (Guo et al., 1998).

Five among the seven already characterized genes that were identified by this screen specify factors with clear linkage to Dam-dependent DNA methylation. Such a large ratio strongly suggests that some among the 11 genes of unknown function should specify activities that modulate or recognize the methylation status of DNA, too. To test this prediction, we performed preliminary analyses on three genes of unknown function ($\text{ycbG}$, $\text{yifE}$ and $\text{yjaG}$). This study, which will be published elsewhere, reports that one gene among the three that were tested, $\text{yjaG}$, specifies a factor that affects the Dam-dependent remethylation of DNA.

As a preliminary characterization of these genes, we constructed null mutants of the three genes, $\text{ycbG}$, $\text{yifE}$ and $\text{yjaG}$, and analyzed their spontaneous mutation rate by using the assay developed by Garibyan et al. (2003). We estimated the frequency at which rifampicin resistance was acquired within each mutant. The spontaneous mutation rate was nearly identical in $\text{ycbG}$, $\text{yifE}$ and WT cells. In contrast, the spontaneous mutation rate of $\text{yjaG}$ cells was lower than that of WT cells (M.S. Hiet, H. Beuneu, P. Brezellec, J.L. Ferat, manuscript in preparation). The characterization of the mutations leading $\text{yjaG}$ cells to become resistant to rifampicin revealed a significant and specific deficit in transition mutations—mainly repaired by the Dam-dependent mismatch repair system (data not shown). We confirmed that transition mutations were more efficiently repaired in a $\text{yjaG}$ mutant strain, by establishing that the mutant cells were also

---

*Fig. 2.* Distribution of the genes associated with the $\text{dam}^{ab}$-containing genomes within the Alteromonadales. The phylogenetic tree is based on Dam protein sequences (branch length is not scaled). (+) or (−) indicate the presence or the absence of the gene (top) in Shewanella oneidensis, in Idiomarina loihiensis, in Pasteurellales (PAS), in Enterobacteriales (ENT) or in Vibrionales (VIB).
less sensitive to the mutagenic agent 2-aminopurine—a base analog that exerts its toxicity by increasing specifically the rate of transition mutation (Garibyan et al., 2003, data not shown). Finally, we showed by assaying directly the kinetics of remethylation at four different sites around the unique origin of replication of the chromosome, oriC, that transition mutations were more efficiently repaired in a yjaG mutant strain, owing to a significant delay of Dam-dependent remethylation of DNA (data not shown).

4 DISCUSSION

4.1 A general method to search accurately for co-evolving activities

It would seem at first that in order to identify genes that are systematically present in the dam<sup>−</sup> genomes, one should generate lists of orthologous genes. Identifying orthologs is usually achieved by looking for ‘bidirectional best hits’ (e.g. Overbeek et al., 1999). This approach, however, suffers from several drawbacks. The higher the number of organisms, the higher the probability of breaking orthologous links. The concept of orthology breaks down for genes specifying complex multidomain proteins (Koonin et al., 2000). Also, pairwise sequence alignments (e.g. Altschul et al., 1997) are known to be less sensitive than position-sensitive scoring matrices (Gribskov et al., 1987) or hidden Markov model profiles (Eddy, 1998) that are stored, for instance, in the Pfam database (Bateman et al., 2004) and which allow to decompose a protein into ‘domains’, i.e. conserved evolutionary units that often also correspond to functional units (Vogel et al., 2004). Last, but not least, a domain may be characteristic of a set of organisms, i.e. it belongs to at least one protein of each considered organism, and nevertheless be ‘hosted’ by non orthologous genes. For these reasons, we chose to focus on domains rather than on complete genes and developed an approach relying on Pfam domains.

As outlined previously, we chose to base our search strategy for dam<sup>−</sup>-associated activities on domains because of our concern that a mere search for similarity would result in systematically discarding complex proteins with multiple activities, only one of which had evolved to become sensitive to DNA methylation, the remainder of the protein being widely distributed among genomes. MukB offers an immediate illustration of this situation. Only a small section of this 1486 amino acid protein, corresponding to its PF04310 domain (227 amino acid), is specific to dam<sup>−</sup> genomes. A BLAST would have shown that MukB is related to various proteins distributed in bacterial and eukaryotic genomes, thus eliminating the protein from the set of genes strictly associated with dam<sup>−</sup> genomes, despite the fact that part of the mukB gene belongs unequivocally to this set. Thus, an alignment-based search for homology would require analyzing one by one the output items in order to decide whether or not to retain a gene. Hence, for the task to be automated, such an approach should be replaced with a domain-based strategy.

It is interesting to note that two Pfam domains that had been retained by our initial screen specify activities required for fermentation. They were eventually discarded because they are largely distributed outside the gamma proteobacteria (Table 1). These domains correspond to pyruvate formate lyase (PF02901) and glycine radical activity (PF01228). The presence of these domains in our provisional list of activities potentially associated with dam<sup>−</sup> revealed that there had been a second, unintentional component in our initial screen. The Pseudomonadales, the Legionellales and the Xanthomonadales that we used as ‘out-group’ happen to be strictly oxidative. Thus, by screening dam<sup>−</sup> (facultative fermentative) against dam<sup>−</sup> lacking (strictly oxidative) genomes, we accidentally collected two genes, pifE and grcA, that potentially specify activities critical for fermentative growth.

The positive outcome of our screening methodology led us to perform the opposite request in the hope of identifying genes involved in DNA maintenance in gammaproteobacteria that do not have a dam<sup>−</sup> gene. We collected the domains that were strictly confined to the genomes of the Pseudomonadales, the Xanthomonadales and the Legionellales; the Enterobacteriales, the Vibrionales and the Pasteurellales constituted the out-group. Interestingly, two domains (PF04079 and PF02616) among the three that were returned by DomainSieve specify activities similar to the chromosome condensation and segregation system, Muk, of the dam<sup>−</sup> organisms, indicating that different sets of genes were selected through evolution to perform a conserved function (evolutive considerations pertaining to this point are developed in Section 4.3). The third domain (PF03653) belongs to a family of hypothetical integral membrane proteins found exclusively in gram-negative bacteria.

4.2 Robustness of the screening methodology

Our screen extracted 33 domains, among which 18 correspond precisely to our query. Given the few genomes sequenced among each order selected for this screen, one may argue that the results are highly dependent on the particular genomes that have been analyzed. The results presented in Figure 3 infer, on the contrary, that the amount of relevant domains stabilizes when a critical number of genomes is considered for analysis.

We constituted nine sets based on the chronological release of the genomes at the NCBI. The first set contains the first five genomes that have been released, belonging to any of the six orders considered for the screen. The second set contains the genomes of the first set plus the next five genomes, and so on. In each set, the genomes were distributed into two groups: the dam<sup>−</sup>-containing and the dam<sup>−</sup>-lacking genomes.

The number of non-relevant domains retained through the screen decreases exponentially with the number of genomes analyzed, indicating that a critical number of genomes is required for this kind of analysis. Also, the fraction of relevant domains stabilizes at a value around 0.6 when 30 or more genomes are taken into account, indicating that the addition of extra genomes within a given set does not improve the sharpness of the screen above a certain threshold. As a matter of fact, the analysis that we carried out with >35 genomes resulted in the elimination of a few relevant domains, PF02976 (mutH) within the eighth set, PF03603 (holD) and PF04310 (included in mukB) within the ninth set. The distribution of mutH shows that the gene is present in Legionella, while the elimination of holD and mukB results from a frameshift (probably an annotation error) that interrupts the coding sequence of both genes in Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67. Yet, the activity of the proteins specified by mutH and mukB are clearly associated with the methylation status of the DNA.

This situation illustrates a limit of this screen as it is constructed now; it does not take into account rare situations where a given gene is seldomly distributed outside of, or exceptionally missing from, a genome belonging to its own group. Further development of DomainSieve should overcome this drawback.
rather narrow phylogenetic distribution of those damrdt assertion. The control of replication initiation, as a means to modulate the gene products that specify it. Two examples may illustrate this of universal and essential biological processes (i.e. DNA and damrdt from each request were split into two groups; domains strictly confined to based on their chronological release at the NCBI database. Output results because they specify factors involved in replication initiation control ment. These genes have not been selected during evolution merely case DNA maintenance—to modifications of the genomic environ-

This situation, which may seem paradoxical at first, reflects the fact that the pressure of selection is exerted on the function rather than on the gene products that specify it. Two examples may illustrate this assertion. The control of replication initiation, as a means to modulate cell proliferation, is actively maintained in living organisms. Only damrdt organisms, however, exploit the ability of SeqA to sense the methylation status of the DNA to prevent the initiation of extra rounds of replication. SeqA is appropriate in these organisms because it binds tightly to hemimethylated oriC, i.e. to DNA on which replication has just been initiated. Faithful replication of the chromosome provides another example. The correction of nucleotide mis-incorporation is critical for the species stability and is implemented by activities that act on newly synthesized DNA. For this post-replication process to be efficient, the newly synthesized strand must be identified accurately. In this case again, the evanescent trail of hemimethylated DNA that follows the polymerase during replication is exploited by taking advantage of MutH, an endonuclease that marks newly synthesized DNA by cleaving specifically the unmethylated strand. The emergence of activities such as the ones specified by seqA or mutH in damrdt genomes illustrates the adaptation of a biological process—in this case DNA maintenance—to modifications of the genomic environment. These genes have not been selected during evolution merely because they specify factors involved in replication initiation control and DNA repair, but also because SeqA and MutH recognize efficiently the DNA that has just been replicated. Knowing when the various sections of their chromosome have been replicated is critical for all cells and in damrdt genomes, DNA methylation became the signal that reports in real time the state of the replication process.

Because of the major impact of DNA methylation on DNA maintenance in general, it is safe to predict that the list of biological functions that depend on the methylation status of the DNA is destined to grow. Our screen provides a list of candidate genes for such functions, among which yjaG is thus far one of the most promising.

4.3 The origin of genes associated with damrdt might shed light on evolutionary constraints behind complex biological functions

Except for isolated instances of mukE, mukF, mutH, yecM, yeeX, yehL and yihl outside of the gammaproteobacteria (Table 1), the genes identified through our screen are strictly confined to genomes with a resident dam gene. Hence, it is striking to observe, given the rather narrow phylogenetic distribution of those damrdt-associated genes, that some of them specify activities that took over the control of universal and essential biological processes (i.e. DNA repair, replication initiation, daughter chromosome cohesion and segregation).

AKNOWLEDGEMENTS

The authors are indebted to François Michel for the fruitful discussions, critical reading of the manuscript and are specially grateful for his consistent support throughout this project. The authors thank David Bates, Rita Cha and Bénédicte Michel for critical reading of the manuscript and helpful comments, Danielle Bittencourt and Hugues Deletain for technical assistance.

Conflict of Interest: none declared.

REFERENCES


Slater,S. et al. (1995) E.coli SeqA protein binds oriC in two different methyl-modulated reactions appropriate to its roles in DNA replication initiation and origin sequest-

Fig. 3. DomainSieve was run with nine sets of genomes that were constituted based on their chronological release at the NCBI database. Output results from each request were split into two groups; domains strictly confined to damrdt genomes (grey) and broadly distributed (empty). The fraction of relevant domains is indicated for each set (crosses). The list of genomes used, as well as their time of release is provided in the Methods section. Sets 1–9 are constituted with genomes released till 7/26/00, 10/5/01, 7/10/02, 1/22/03, 8/21/03, 4/23/04, 9/11/04, 10/30/2004, 8/29/05, respectively. The results presented in this paper correspond to the seventh set (arrow).

REFERENCES