Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Biology and Chemistry 28 (2004) 109–118

Computational Biology and Chemistry

www.elsevier.com/locate/compbiolchem

The difficult interpretation of transcriptome data: the case of the GATC regulatory network

Alessandra Riva[a,*], Marie-Odile Delorme[a], Tony Chevalier[b], Nicolas Guilhot[b], Corinne Hénaut[b], Alain Hénaut[a]

[a] *CNRS, Laboratoire Génome et Informatique, Tour Evry 2, 523 Place des Terrasses, 91034 Evry cedex, France*
[b] *METabolic EXplorer S.A., Biopôle Clermont-Limagne, 63 360 Saint-Beauzire, France*

Received 22 November 2003; received in revised form 12 December 2003; accepted 12 December 2003

Abstract

Genomic analyses on part of *Escherichia coli*'s chromosome had suggested the existence of a GATC regulated network. This has recently been confirmed through a transcriptome analysis. Two hypotheses about the molecular control mechanism have been proposed—(i) the GATC network regulation is caused by the presence of GATC clusters within the coding sequences; the regulation is the direct consequence of the clusters' hemi-methylation and therefore their elevated melting temperature, (ii) the regulation is caused by the presence of GATCs in the non-coding 500 bp upstream regions of the affected genes; it is the consequence of an interaction with a regulatory protein like Fnr or CAP. An analysis of the transcriptome data has not allowed us to decide between the two hypotheses. We have therefore taken a classic genomic approach, analyzing the statistical distribution of GATC along the chromosome, using a realistic model of the chromosome as theoretical reference. We observe no particular distribution of GATC in the non-coding upstream regions; however, we confirm the presence of GATC clusters within the genes. In order to verify that the particular distribution observed in *E. coli* is not a statistical artefact, but has a physiological role, we have carried out the same analysis on *Salmonella*, making the hypothesis that the genes containing a GATC clusters should be largely the same in the two bacteria. This has been indeed observed, showing that the genes containing a GATC cluster are part of a regulation network. The present is a case study, which demonstrates that the analysis of transcriptome data does not always permit to identify the primary cause of a phenomenon observed; on the other hand, a classic genomic approach linked with a comparative study of related genomes may allow this identification.

Keywords: GATC; Statistics; Transcriptome analysis; Comparative genomics; Fnr

1. Introduction

1.1. The facts known about GATC and Dam

The tetranucleotide GATC is methylated in *E. coli* by the DNA methyltransferase (Dam); this enzyme methylates the adenine residue within 5′-GATC-3′ sequences in double stranded DNA. The GATC tetranucleotide has an important physical property: the melting temperature of a 10 bp oligonucleotide containing methylated GATC is by 13°C lower compared to the hemi-methylated oligonucleotide (see Fazakerley et al., 1985). In fast growing bacterial cells (for example in the intestine of their warm blooded host) the Dam is a limiting factor and the DNA will be under-methylated (Boye et al., 1992; Plumbridge and Söll, 1987). It is known that GATC motifs and their methylation by Dam play an important role in *E. coli*. They are involved in mismatch repair (see Marti et al., 2002) for a review on the subject of mismatch repair and Bhagwat and Lieb (2002) for a review concerning *E. coli* only) and the control of chromosome replication (see Donachie, 2001) for a concise overview on the subject). The methylation state of GATC is also involved in the expression of the pap operon; this operon codes of the Pap pili, which are of great importance in the pathogenicity of uropathogenic *E. coli* (Hale et al., 1998).

1.2. A GATC regulated network implied in the stress response

Mutants that lack Dam are characterized by a pleiotropic phenotype; they show for example an increased sensitivity

[*] Corresponding author. Tel.: +33-1-60-87-38-63; fax: +33-1-60-87-38-97.
E-mail address: gucki@genopole.cnrs.fr (A. Riva).

1476-9271/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2003.12.004

to DNA-damaging agents, have a higher mutability and increased hyper-recombination (Marinus, 2000).

In *E. coli* an additional function for GATC, when present in clusters, was proposed in the work by Hénaut et al. (1996); an analysis on part of *E. coli*'s genome of the distribution of GATC clusters suggested the existence of a network of genes, containing these clusters, involved in the control of the cell's metabolism when undergoing cold shock and oxygen shift (the environmental temperature is suddenly lowered, for example when the bacteria pass from the intestine of their warm blooded host to the external environment; this process also involves a change from an anaerobic to an aerobic environment). The existence of this network has recently been confirmed by a transcriptome analysis carried out by Oshima et al. (2002).

1.3. The mechanisms proposed for the GATC regulated network

Hénaut et al. (1996) propose the following mechanism: When the bacteria undergo coldshock (and an oxygen shift), caused by the passage from the intestine to the external environment, the transcription of genes containing a GATC cluster will be blocked at the level of the cluster because of the high stability of the hemi-methylated DNA.

Oshima et al. (2002), on the other hand, propose the following: "The promoters of most of these Dam controlled genes were also found to contain GATC sequences that overlap with recognition sites for two global regulators, fumarate nitrate reduction (Fnr) and catabolite activator protein (CRP). We propose that Dam-mediated methylation plays an important role in the global regulation of genes, particularly those with Fnr and CRP binding sites".

The two hypotheses differ considerably from each other:

- Hénaut et al. (1996) propose that the GATC network regulation is caused by the presence of GATC clusters within the coding sequences; the regulation is the direct consequence of the clusters' hemi-methylation and therefore their elevated melting temperature.
- Oshima et al. (2002) propose that the GATC network regulation takes place upstream of the coding sequences and that it is the consequence of an interaction with a regulatory protein like Fnr or CAP.

1.4. The present paper

In order to investigate the two hypotheses, we first looked at the relationship between the position, frequency and distribution of GATC and the differences in gene expression observed during the transcriptome analysis carried out by Oshima et al. (2002). We could not find any relationship. Thus the analysis of the transcriptome data does not allow us to confirm or reject either of the two hypotheses.

We find ourselves with the following problem: firstly, the transcriptome data show that almost ten percent of the total genes analyzed are influenced by the *dam* genotype. Secondly, we know that the target of Dam is GATC. It is therefore surprising that there seems to be no relationship between the two.

One possible explanation could be that for a majority of the genes the observed changes in transcription levels is not caused by a direct interaction with the Dam protein but is the result of a secondary effect.

As one cannot obtain any further information from the transcriptome analysis, we have decided to take a classical genomic approach to the question. We have created a model chromosome for *E. coli* and have compared the GATC motif distribution in the model chromosome with the real chromosome. We then have looked for the existence of GATC patterns particular only to the real chromosome of *E. coli*.

We have made the following hypothesis: if a pattern found has a physiological role, it should have been conserved in the course of evolution: genes containing the GATC pattern should be the same—or affect equivalent functions—in *E. coli* and a closely related bacterium like *Salmonella*. This has been indeed observed; we suggest that the genes containing the GATC clusters are indeed the key-elements of a regulation network. We propose also that these genes cause indirectly the changes in transcription levels observed by Oshima et al. (2002).

2. Materials and methods

2.1. Experimental data

The genome sequences and feature tables for *E. coli K12* and *Salmonella enterica* serovar Typhi CT18 were obtained from the EMBL Nucleotide Sequence database (available at ftp://ftp.ebi.ac.uk/pub/databases/embl/genomes/Bacteria/ecoli_K12 and ftp://ftp.ebi.ac.uk/pub/databases/embl/genomes/Bacteria/styphiCT18, respectively).

Oshima et al. (2002) studied the expression of 4019 genes in a dam^{+}/dam^{-} background under different experimental conditions. The transcriptome data by Oshima et al. (2002) were obtained from their article as well as the following web site: http://ecoli.aistnara.ac.jp/xp_analysis/dam/all.html. Oshima et al. (2002) consider 349 genes as relevant; these are listed in Table SI, column D in the Supplementary Data. We will refer to this list as "Column D". Using the complete data of Oshima et al. (2002) (http://ecoli.aistnara.ac.jp/xp_analysis/dam/all.html), we have obtained a list of 389 genes (see Table SI, column E in the supplementary data) whose expression varies by a factor three in at least one of the four experimental conditions tested by Oshima et al. (2002). We will refer to this list as "Column E". The two lists obtained have 60% of the genes in common.

2.2. Operons

In order to analyse the data by Oshima et al. (2002), we needed to identify the operons in *E. coli* to a reasonable degree. A presumed operon is defined as a group of contiguous genes that are on the same reading strand delimited either by a promoter and a terminator (predicted or not) or a gene, which lies on the other DNA strand (see Table SI, columns F and G in the Supplementary data).

2.3. Construction of the virtual chromosome

To build a virtual chromosome it is necessary to define the nature of the question to be examined. In this study we want to investigate the distribution of GATC along the chromosome. We therefore need to respect the GATC frequency at a local scale (in our case within a gene or within an intergenic region). Fig. 1 illustrates well that the distance between two successive GATCs in the real chromosome is not monotonous and characterized by a strong periodicity of three. This reflects the chromosomal organization, which is based on codons (85% of *E. coli*'s chromosome is made up of coding sequences). It is therefore imperative to construct the coding sequences of the virtual chromosome gene by gene. In order to generate a virtual gene, we start by compiling the frequency table for the di-codons found in the real gene. The first codon of the virtual gene is the same as the real gene (normally ATG). The second codon will be chosen at random from the di-codons "ATG-XXX", taking into account their respective frequencies in the table. The procedure is repeated until the virtual gene has the same length as the real one. Fig. 1 shows that the virtual chromosome follows well the GATC frequency at the local scale. The virtual chromosome thus obtained is more realistic than the one constructed by Hénaut et al. (1996), as it based on a gene-by-gene procedure, rather than on a global simulation, gene class by gene class.

The consequence of this procedure is that the frequencies of the di-, tri- and tetra-nucleotides (like GATC) are the same as in the real gene. However, the distance between two successive GATC is not imposed on the virtual chromosome. Therefore, if differences are observed in the GATC distribution, between the real and the virtual chromosomes, these differences must be meaningful and related to the question studied (Hénaut et al., 1985; Hénaut and Vigier, 1985).

To construct the intergenic sequences we proceed as above, utilizing the di-nucleotide frequencies rather than the di-codon frequencies.

In order to obtain a reasonable precision in the statistical analyses we compare the real chromosome with ten virtual chromosomes.

2.4. Data mining tools

In order to gain information about the regions thus identified, we have used various databases:

- EcoCyc (http://biocyc.org/ecocyc/),
- KEGG (http://www.genome.ad.jp/kegg/kegg2.html),
- METAVISTA® (proprietary data base of the Metabolic Explorer society http://www.metabolic-explorer.com),
- PubMed http://www.ncbi.nlm.nih.gov/,
- Swiss-Prot (http://www.ebi.ac.uk/swissprot/index.html) using the SRS search tool to interrogate the SWALL (SPTR) database (accessible as SWall on the SRS server).

2.5. Analysis procedure

2.5.1. Analysis of the transcriptome data

We examine the relationship between the presence and the position of GATC and the differences in gene expression observed during the transcriptome analysis.

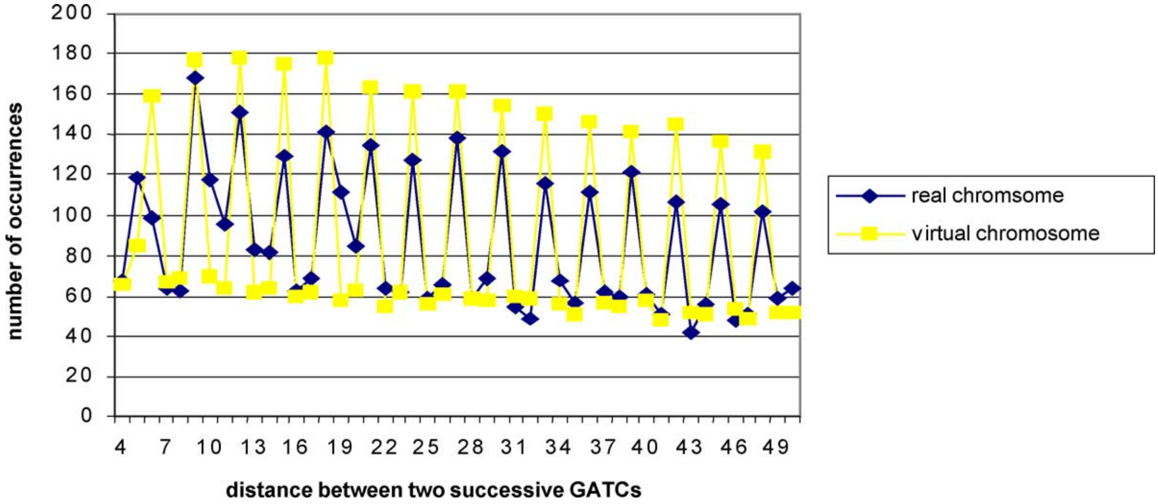

Fig. 1. Comparison of the GATC distribution in the real and the virtual chromosomes. The distance of successive GATCS is displayed in the real chromosome and the virtual chromosome. The distribution in the real chromosome is characterized by a strong periodicity of three, reflecting the fact that 85% of the chromosome is made up of coding sequences. The virtual chromosome follows this distribution closely.

1. We analyze the **direction of the expression change** in the two columns (Columns D and E).
2. As Oshima et al. (2002) propose that the GATC network regulation takes place upstream of the coding regions, we take a closer look at the distribution of GATC in the 500 bp **upstream region** of each gene. Listed in their results (http://ecoli.aistnara.ac.jp/xp_analysis/dam/all.html) are the numbers of GATCs found up to 500 bp upstream of each of the 4019 genes analyzed (see also Table SI, column C). We proceed to examine whether there is a relationship between the number of upstream GATCs and the change of expression levels in function of the *dam* genotype (see Table 1 and Fig. 2).
3. Oshima et al. (2002) also propose that the GATC network regulation (apart from taking place upstream of the coding sequences) is the consequence of an interaction with a regulatory protein like Fnr or CAP. In order to test this hypothesis, we have identified all the Fnr consensus sequences present in *E. coli*. In accordance with Melville and Gunsalus (1996) we have utilized the following consensus sequence: TTGATnnnnATCAA. We have identified a total of 22 sites, 15 in upstream regions and 7 within coding sequences. In order to study the relationship between Fnr and the *dam* genotype, we introduce the concept of "genetic structure": If the Fnr site controls an isolated gene, the genetic structure corresponds to this gene. If, however, the Fnr site is upstream of an operon, or within a gene belonging to an operon, the genetic structure corresponds to the operon. We consider a genetic structure to be sensitive to the *dam* genotype if at least one gene belonging to the structure is present in Column D or Column E. The results are displayed in Table 2.

2.5.2. Genomic analysis on the GATC frequency

The hypothesis of Oshima et al. (2002) predicts the existence of regions, 500 bp long, that possess an unusual frequency of GATC. This hypothesis may be examined with the help of a binomial distribution model, as the region under examination is sufficiently large:

1. We examine the distribution of GATC in the 500 bp **upstream region** of each gene. The total number of GATCs present in these 4019 regions is 7174. The theoretical frequencies of the GATCs in the upstream regions are calculated with the hypothesis of a binomial distribution with a probability of $7174/(500 \times 4019)$ and compared with the actual distribution. The results are displayed in Table 3 and Fig. 3.
2. We then examine the results obtained above in more detail, by distinguishing three kinds of upstream regions. Regions, which are upstream of an isolated gene, upstream of the first gene of an operon or upstream of a gene within an operon. In the first two cases, the upstream regions are non-coding sequences, whilst the upstream region of a gene within an operon will correspond to a coding sequence. We therefore calculate three theoretical

Table 1
The GATC distribution in the 500 bp upstream regions of the genes in function of their sensitivity to the *dam* genotype

No of GATC	Sensitivity to the *dam* genotype	
	Yes (genes in column D)	No
1a		
0	54	728
1	114	1068
2	91	919
3	53	525
4	17	237
5	13	114
6	2	39
7	5	28
8		8
9		2
10		8
11		
12		
13		
14		
15		
16		
17		
18		1
19		
Total genes	349	3670
1b	Yes (genes in column E)	
0	60	722
1	121	1061
2	99	911
3	57	521
4	27	227
5	15	112
6	2	39
7	6	27
8		8
9	1	1
10		1
11		
12		
13		
14		
15		
16		
17		
18	1	
19		
Total genes	389	3630

We have analyzed the GATC distribution in the 500 bp region of each gene. In order to establish whether there is a relationship between the number of upstream GATCs and the change of expression levels in function of the *dam* genotype, the genes have been divided in function of their sensitivity to the *dam* genotype, in Table 1a according to Oshima et al. (2002) and in Table 1b according to our criterion (see Methods for details). In both cases no relationship is present (p-value = 21.5 and 30.4%, respectively). Note: the upstream region containing 18 GATCs corresponds to the region of the origin of replication. For an easier interpretation, the data of Tables 1a and 1b are displayed as histograms in Fig. 2a and b, respectively.

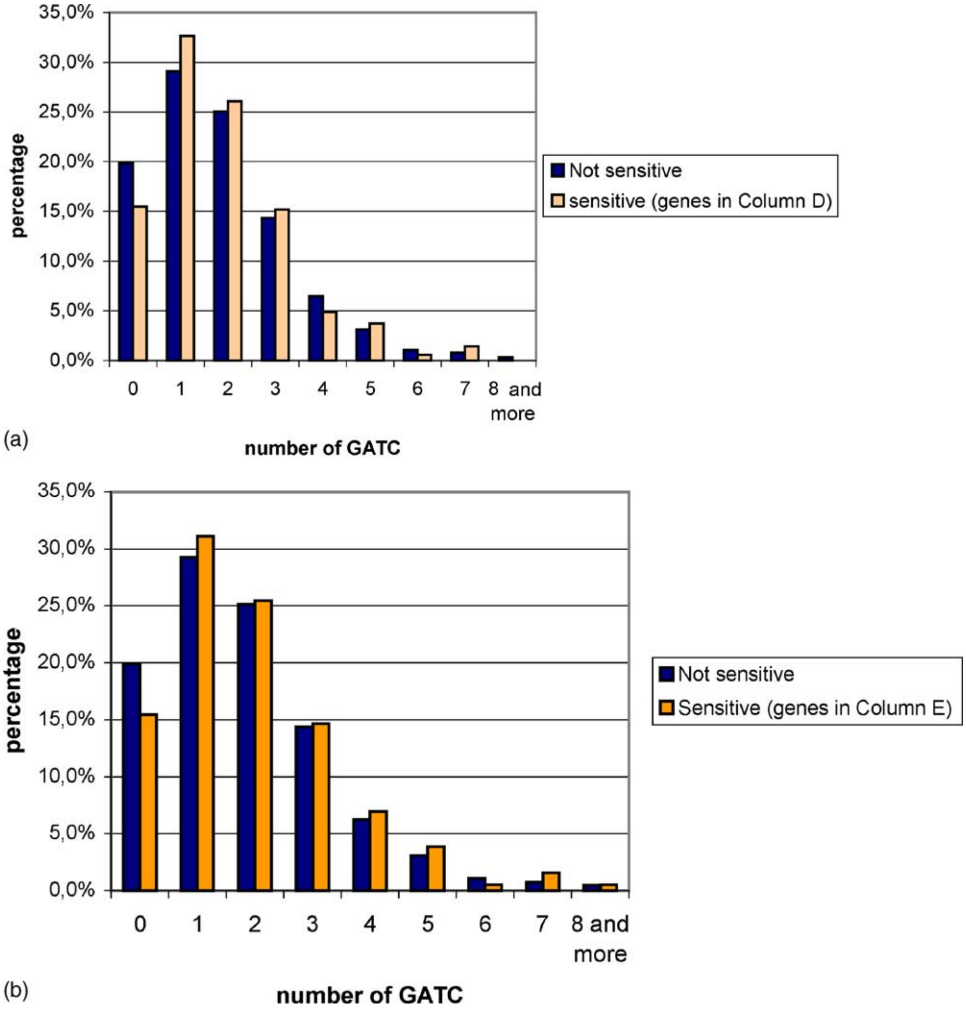

Fig. 2. The data contained in Table 2 are displayed in form of histograms. See the legend for Table 2 for details. (a) The GATC distribution in the 500 bp upstream regions of the genes in function of their sensitivity to the *dam* genotype (column D). (b) The GATC distribution in the 500 bp upstream regions of the genes in function of their sensitivity to the *dam* genotype (column E).

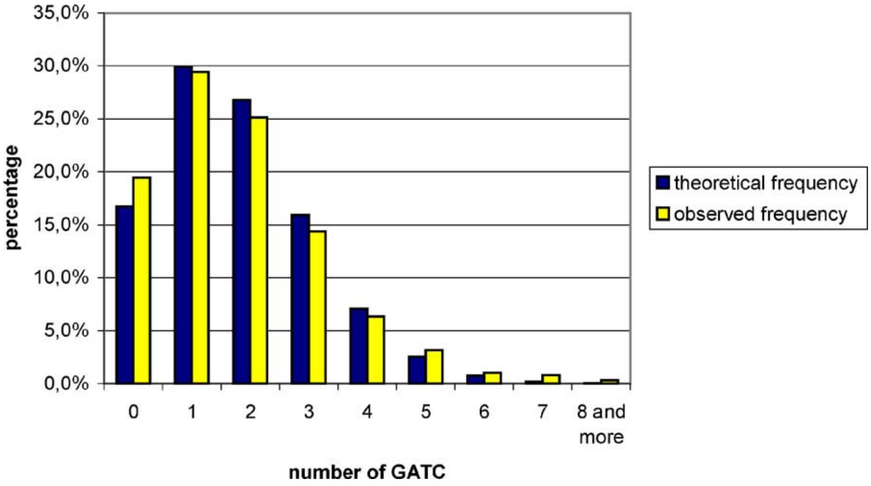

Fig. 3. The theoretical and observed GATC distribution in the 500 bp upstream regions of the genes. The data contained in Table 3 are displayed in form of a histogram. See the legend for Table 3 for details.

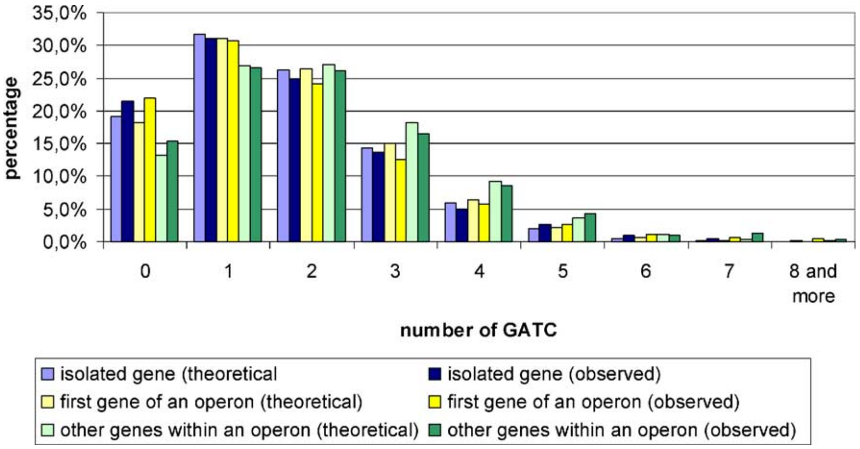

Fig. 4. The theoretical and observed GATC distribution in the 500 bp upstream regions of the genes, in function of the kind of gene. The data contained in Table 4 are displayed in form of a histogram. See the legend for Table 4 for details.

frequencies of the GATCs in the upstream regions, each with the hypothesis of a binomial distribution—(a) for non-coding regions upstream of an isolated gene, with a probability of $2986/(1813 \times 500)$, (b) for non-coding regions upstream of the first gene of an operon, with a probability of $1414/(831 \times 500)$ and (c) for coding regions upstream of a gene within an operon, with a probability of $2774/(1375 \times 500)$ (see Table 4 and Fig. 4).

Table 2
The relationship between Fnr and the sensitivity to the *dam* genotype

	Fnr consensus sequence		
	Not present	Present, upstream of a coding sequence	Present, within a coding sequence
(a) Sensitivity to the *dam* genotype (column D)			
No	2511	12	6
Yes	230	3	1
(b) Sensitivity to the *dam* genotype (column E)			
No	2492	10	7
Yes	249	5	0

We have identified all the Fnr consensus sequences present in *E. coli* and found a total of 22. Fifteen lie in upstream regions whilst seven are found within coding sequences. We introduce the concept of "genetic structure" controlled by Fnr: if an Fnr consensus sequence lies upstream or within an isolated gene, we consider only this gene to be regulated by Fnr and therefore to be the "genetic structure". If, however an Fnr consensus sequence lies upstream of an operon or in a coding sequence within an operon, we consider the entire operon to be one genetic structure, regulated by Fnr. We consider a genetic structure to be sensitive to the *dam* genotype if at least one gene belonging to the structure is sensitive. For the sensitivity to the *dam* genotype we refer to Column D and Column E; the results are displayed in Tables 2a and 2b, respectively. Table 2a: There exists no correlation between the presence of a Fnr consensus sequence and the sensitivity to the dam genotype (P-value = 10.7%). Table 2b: A slight correlation can be detected (P-value = 0.1%). However, it should be noted that the numbers dealt with are very small and that even in Table 2b, only 1/3 of the genes possessing an Fnr consensus sequence upstream are dam-sensitive.

Table 3
The theoretical and observed GATC distribution in the 500 bp upstream regions of the genes

No of GATC	No of genes expected	No of genes observed
0	672.2	782
1	1204.2	1182
2	1076.5	1010
3	640.2	578
4	285.0	254
5	101.3	127
6	29.9	41
7	7.6	33
8	1.7	8
9	0.3	2
10	0.1	1
11		
12		
13		
14		
15		
16		
17		
18		1
19		
Total	4019	4019

The theoretical frequencies of the GATCs in the upstream regions have been calculated with the hypothesis of a binomial distribution with a probability of $7174/(4019 \times 500)$ (7174 GATCs, 4019 regions of 500 bp). The theoretical frequency is compared with the actual frequency observed. For an easier interpretation, the data of Table 3 are displayed as a histogram in Fig. 3. The GATC distribution is not binomial (P-value $= 1.4 \times 10^{-06}$). There is a slight excess of genes with no upstream GATC and a slight excess of genes containing three to five upstream GATCs.

2.5.3. *Genomic analysis on local GATC enrichments*

The hypothesis of Hénaut et al. (1996) predicts the existence of small regions, a few dozens of bp long that are enriched in GATC and present in the coding sequences. This hypothesis may be examined with the help of the virtual chromosome constructed (see Section 2.3).

Table 4
The theoretical and observed GATC distribution in the 500 bp upstream regions of the genes, in function of the kind of gene

No of GATC	Isolated gene		First gene of an operon		Other genes within an operon	
	Theoretical	Observed	Theoretical	Observed	Theoretical	Observed
0	348.3	399	151.1	187	182.1	215
1	575.5	573	258.0	257	368.9	371
2	474.6	464	219.8	203	372.9	367
3	260.3	250	124.6	106	250.8	231
4	106.9	90	52.9	48	126.2	120
5	35.0	47	17.9	22	50.7	59
6	9.6	17	5.0	10	17.0	14
7	2.2	9	1.2	6	4.8	18
8	0.5	2	0.3	2	1.2	4
9	0.1	1	0.0	1	0.3	
10					0.1	1
11						
12						
13						
14						
15						
16						
17						
18				1		
19						

We distinguish three kinds of upstream regions: regions, which are upstream of an isolated gene, upstream of the first gene of an operon or upstream of a gene within an operon. In the first two cases, the upstream regions are non-coding sequences, whilst the upstream region of a gene within an operon will correspond to a coding sequence. Three theoretical frequencies of the GATCs in the upstream regions have been calculated, each with the hypothesis of a binomial distribution—(a) for non-coding regions upstream of an isolated gene, with a probability of 2986/(1813 × 500) (2986 GATCs, 1813 region of 500 bp), (b) for non-coding regions upstream of the first gene of an operon, with a probability of 1414/(831 × 500) and (c) for coding regions upstream of a gene within an operon, with a probability of 2774/(1375 × 500). For an easier interpretation, the data of Table 4 are displayed as a histogram in Fig. 4. In all three cases the GATC distribution does not differ significantly from a binomial distribution (*P*-value equal to 4.3, 4.8 and 4.7%, respectively).

(1) Based on the approach by Hénaut et al. (1996), we look for configuration(s) of GATC, which are particular to the real chromosome of *E. coli* by comparing it with a virtual chromosome. We take up the idea of “local enrichment in GATCs”, trying to find a rule, which maximizes the differences between the real and the virtual chromosome; this leads us to consider quadruplets of GATC rather than triplets, as done by Hénaut et al. (1996). With this stricter rule (see below) we find 2.5% of local enrichments (or clusters) in the real chromosome and 1.3% in the virtual chromosome (compared to 6 and 4%, respectively, in Hénaut et al. (1996)). The criterion is the following:
   (a) GATC pairs separated by less than 8 bp and triplets in regions spanning less than 62 bp are kept in a preliminary screening.
   (b) In a second step we retain only those regions where there are at least four GATC motifs and where the average distance between pairs is shorter than 18 bp.

(2) We search *E. coli*’s genome for regions containing GATC clusters by applying this criterion. We find 76 genes and the region of the origin of replication that contain a GATC cluster (see Table SII in the Supplementary Data). No correlation exists between the list of genes contained in Table SII and those contained in Column E (*P*-value = 59%); a slight correlation exists between Table SII and Column D (*P*-value = 3%).

(3) Applying the same criterion in *Salmonella*, we find 60 genes and the region of the origin of replication that contain a GATC cluster (see Table SII in the Supplementary Data). We obtain a total of 113 different genes affected in *E. coli* or *Salmonella*.

(4) With the help of the data mining tools (see Section 2) we try to group the genes into sets that exhibit self-consistent biological properties and compare the functions affected in the two organisms (see Table SII for details). The results are discussed in Section 3.2.

3. Results and discussion

3.1. The relationship between GATC and the transcriptome data

We have analyzed the transcriptome data by Oshima et al. (2002) in order to find a relationship between GATC and the differences in gene expression observed.

The prediction by Oshima et al. (2002) is that the genes, whose transcription varies in function of the *dam* genotype, have an elevated number of GATC in the 500 bp upstream. We could not find any relationship between the number of GATC in the genes’ upstream regions and their susceptibility to the *dam* genotype, neither when examining the genes regarded as relevant by Oshima et al. (2002) (“Column D” genes, see Table 1 and Fig. 2a) nor when examining the genes relevant according to our criterion (“Column E” genes, see Table 1 and Fig. 2b).

A second prediction by Oshima et al. (2002) is the existence of a link between the Fnr consensus sequence and the regulation by Dam. Fifteen genes contain a Fnr consensus sequence in their upstream region; only three or five of these are susceptible to the *dam* genotype (depending on whether one works on the genes of Column D or Column E, see Table 2). When working with Column E, a slight correlation exists between the sensitivity to the dam genotype and the presence of an Fnr consensus sequence (P-value $= 0.1\%$). However, it should be noted that the majority of the 15 genes containing a Fnr consensus sequence are not sensitive to the dam genotype (4/5 when working with column D and 2/3 when working with Column E). Furthermore, when we take into consideration that out of 349 (389) *dam*-sensitive genes only 3 (5) genes contain an upstream Fnr consensus sequence, we come to the conclusion that we cannot confirm the affirmation made by Oshima et al. (2002) that “The promoters of most of these Dam controlled genes were also found to contain GATC sequences that overlap with recognition sites for two global regulators, fumarate nitrate reduction (Fnr) and catabolite activator protein (CRP)”.

The prediction by Hénaut et al. (1996) is that the genes, which are regulated by Dam, contain a GATC cluster in their coding sequences. We have taken up this idea and identified a list of genes that contain a GATC cluster (using a stricter definition of “cluster”, see Section 2 for details). These genes are displayed in Table SII. We have looked for a correlation between the genes contained in Table SII and those contained in Column E and in Column D, that is to say, we have examined whether there is a relationship between the presence of a GATC cluster in a gene and its sensitivity to the *dam*-genotype. Again, we could not find any relationship (data not shown).

To resume, the analysis of the transcriptome data shows no relationship between the position, frequency and distribution of GATC and the differences in gene expression observed during the transcriptome analysis carried out by Oshima et al. (2002). Thus the analysis of the transcriptome data does not allow us to confirm or reject either of the two hypotheses.

3.2. *The genomic approach*

It is known that the target of Dam is GATC. This must have repercussions on the distribution of GATC along the chromosome.

We have therefore proceeded to make a genomic analysis of the GATC frequency. Oshima et al. (2002) affirm that a particular distribution of GATC is to be found within the non-coding 500 bp upstream of the *dam*-susceptible genes. We have analyzed the abundance of GATC upstream of each of the 4019 genes (Table 3 and Fig. 3); as observed by Oshima et al. (2002) the distribution is not binomial (P-value $= 1.4 \times 10^{-6}$). We have then proceeded to look in more detail at the different kinds of upstream regions: the upstream region of an isolated gene or the first gene of an operon will coincide with a non-coding sequence, whilst the upstream region of a gene within an operon will coincide with the coding sequence of its upstream neighbour. In all three cases, the GATC distribution does not differ significantly from a binomial distribution (the three P-values are all around 5%, see Table 4 and Fig. 4). This observation does not speak in favour of the hypothesis made by Oshima et al. (2002).

We have then concentrated on the local enrichment of GATC in coding sequences in order to test the hypothesis made by Hénaut et al. (1996). We have searched the genome of *E. coli* for GATC clusters and have found them to be exclusively in coding sequences (76) and the origin of replication. If this statistical property has a physiological reason for existing, we should find that the homologous genes in a bacterium closely related to *E. coli* also contain GATC clusters. We have therefore also searched the genome of *Salmonella* for GATC clusters. Again, only coding sequences (60) and the origin of replication are affected. We can distinguish four categories:

- Twenty three genes that are in common to the two bacteria.
- Forty one genes that, though not homologues of each other, play a similar or equivalent physiological role in the two bacteria.
- Twenty nine genes found in *E. coli* only and belonging to sets not represented in *Salmonella*’s list.
- Twenty genes found in *Salmonella* only and belonging to sets not represented in *E. coli*’s list.

The genes can be grouped into sets that exhibit self-consistent biological properties. For both organisms we find again the classification observed by Hénaut et al. (1996) and Oshima et al. (2002). This is a **supplementary** argument that strengthens the hypothesis of a GATC regulated network. 64 genes belong to groups present in both organisms, namely chromosome replication, respiration, the metabolism of succinate, the metabolism of propionate, purine synthesis, interaction with RNA or ribosomes, cell envelope, ions, coenzymes and cofactors, the phosphotransferase system and “Nitrogen Source”.

The presence of the 41 genes shows that the divergence between *E. coli* and *Salmonella* is sufficiently large for the comparison to be instructive. The same functional groups may be affected by GATC in the two organisms through different genes. A particularly interesting case is given when the same pathway is affected, but through different genes. We will concentrate on these latter, representing a particularly strong argument for the evolutionary conservation of the GATC regulated network, beyond the mere conservation of the sequences themselves. The following serve as examples:

- A number of genes involved in respiration is affected; eight out of at total of 14 genes are in common to both organisms. However, at a closer look one can see that

narL (on the list of *E. coli*) plays an equivalent role to narP (present in *Salmonella*): both enzymes are involved in the co- regulation of a number of genes encoding oxidoreductases and dehydrogenases.
- A characteristic of the formate dehydrogenases (involved in respiration) is that they contain selenocysteine. The insertion of selenocysteine is blocked in *E. coli* through *selB*, which codes for a selenocysteine-specific translation factor. In *Salmonella* we find *cysN*; CysN is required for the formation of selenocysteine tRNA (KEGG). Through *selB* (*E. coli*) and *cysN* (*Salmonella*) the formation of the formate dehydrogenases is thus directly influenced.
- A total of three genes is affected in the propionate catabolism. *prpE* is on the lists of both organisms and codes for a propionate-CoA ligase. *prpB* is present in *E. coli* only and involved in the same metabolic pathway. *prpR* is on the list of *Salmonella*; recent work on *Salmonella enterica* serovar typhimurium strain LT2 shows that PrpR is required for the expression of the *prpBCDE* operon (Palacios and Escalante-Semerena, 2000).

These cases reinforce the hypothesis that the particular distribution of the GATC clusters within the coding sequences is not a statistical artefact. We propose that the genes containing the clusters are the key-elements of the GATC regulatory network; it is at this level that Dam protein acts.

3.3. An apparent contradiction between the genomic and the transcriptome approach

One important point still needs to be addressed—no relationship has been found between the sensitivity to the *dam* genotype (genes listed in Column D and Column E) and the GATC motif; there also has not been found a relationship between genes containing GATC clusters (Table SII) and the sensitivity to the *dam* genotype.

It should be noted that between nine and ten percent of the 4019 genes analyzed by Oshima et al. (2002) vary in function of the *dam* genotype; between one half and two thirds are over-expressed, whilst one half to one third are under-expressed in the *dam*$^-$ genotype (referring to genes in Column D and E, respectively, data not shown). This means it is highly unlikely that we are observing the primary effect of the Dam protein on the GATC motif. The following example shows that such a phenomenon can have a simple explanation. If we take the case of the *lac* operon we can observe that the addition of lactose induces a change of expression of the genes belonging to the operon, but not a change of expression of the *lac*-repressor, who is the primary target of lactose. Thus, the primary effect (the interaction lactose–repressor) would go unnoticed in a transcriptome analysis, whilst only the secondary effect (the change of transcription levels in the *lac* operon) would be observed.

Another explanation for this apparent contradiction between the transcriptome data and the genomic analysis may be found in the work of Sekowska et al. (2001). They conducted steady state and transcriptome experiments on the sulphur metabolism of Bacillus subtilis. The experiments were repeated on different days (day A and B) under "identical" conditions. However, the authors observed that the culture of day A expressed operons involved in competence whilst the culture of day B expressed operons involved in sporulation (these operons are unrelated to the metabolism of sulphur). The authors came to the conclusion that the differences observed in the culture was due to subtle differences in the handling of the pre-cultures (differences in the room temperature when inoculating the culture with the pre-culture).

In other words, Sekowska et al. (2001) conclude that a transitional phenomenon at the beginning of an experiment may give an imprinting to a culture for the entire duration of the experiment; the transcriptome analyses will reveal this imprinting without permitting to identify the genes, which were the primary target of the transitional phenomenon.

Perhaps we can explain the apparent contradiction in our case in this light. During the experiments of Oshima et al. (2002), the transcription of the genes containing GATC clusters may have been changed during the handling of the pre-cultures, as it is very likely that the pre-cultures underwent temperature and oxygen shifts at that moment. What is observed in the transcriptome analysis, are the repercussions on the expression of genes belonging to the same physiological groups as the genes containing the clusters.

At the current state of the technique, the transcriptome analysis is ill suited for the study of such transitional phenomena.

3.4. In conclusion

In the present paper we show that the analysis of transcriptome data does not always permit to identify the primary cause of a phenomenon observed. We also show, on the other hand, that a classic genomic approach coupled with a comparative study of related genomes may permit this identification.

Acknowledgements

The authors wish to thank J.-L. Risler for critical reading of the manuscript. The authors also gratefully acknowledge the constructive criticism made by the referee. This work was supported in part by the French Ministry of the Economy, Finance and Industry (contract ASG no. 01 4 90 6093).

References

Bhagwat, A.S., Lieb, M., 2002. Cooperation and competition in mismatch repair: very short-patch repair and methyl-directed mismatch repair in *Escherichia coli*. Mol. Microbiol. 44, 1421–1428.

Boye, E., Marinus, M.G., Lobner-Olesen, A., 1992. Quantitation of dam methyltransferase in *Escherichia coli*. J. Bacteriol. 174, 1682–1685.

Donachie, W.D., 2001. Co-ordinate regulation of the *Escherichia coli* cell cycle or the cloud of unknowing. Mol. Microbiol. 40, 779–785.

Fazakerley, G.V., Teoule, R., Guy, A., Fritzsche, H., Guschlbauer, W., 1985. NMR studies on oligodeoxyribonucleotides containing the dam methylation site GATC. Comparison between d(GGATCC) and d(GGm6ATCC). Biochemistry 24, 4540–4548.

Hale, W.B., van der Woude, M.W., Braaten, B.A., Low, D.A., 1998. Regulation of uropathogenic *Escherichia coli* adhesin expression by DNA methylation. Mol. Genet. Metab. 65, 191–196.

Hénaut, A., Rouxel, T., Gleizes, A., Moszer, I., Danchin, A., 1996. Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. J. Mol. Biol. 257, 574–585.

Hénaut, A., Vigier, P., 1985. Study of constraints that act upon polynucleotidic sequences. I. Significance of the code degeneracy. C. R. Acad. Sci. Paris 301, 277–282.

Marinus, M.G., 2000. Recombination is essential for viability of an *Escherichia coli* dam (DNA adenine methyltransferase) mutant. J. Bacteriol. 182, 463–468.

Marti, T.M., Kunz, C., Fleck, O., 2002. DNA mismatch repair and mutation avoidance pathways. J. Cell. Physiol. 191, 28–41.

Melville, S.B., Gunsalus, R.P., 1996. Isolation of an oxygen-sensitive FNR protein of Escherichia coli: Interaction at activator and repressor sits of FNR-controlled genes. Proc. Natl. Acad. Sci. U.S.A. 93, 1226–1231.

Oshima, T., Wade, C., Kawagoe, Y., Ara, T., Maeda, M., Masuda, Y., Hiraga, S., Mori, H., 2002. Genome-wide analysis of deoxyadenosine methyltransferase-mediated control of gene expression in *Escherichia coli*. Mol. Microbiol. 45, 673–695.

Palacios, S., Escalante-Semerena, J.C., 2000. prpR, ntrA, and ihf functions are required for expression of the prpBCDE operon, encoding enzymes that catabolize propionate in Salmonella enterica serovar Typhimurium LT2. J. Bacteriol. 182, 905–910.

Plumbridge, J., Söll, D., 1987. The effect of dam methylation on the expression of glnS in *E. coli*. Biochimie 69, 539–541.

Sekowska, A., Robin, S., Daudin, J.J., Hénaut A., Danchin, A., 2001. Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in Bacillus subtilis. Genome Biol. 2, Research0019, e-pub.