

# Genomic Distribution of Short Motifs Involved In DNA Repair in Pathogenic and Non Pathogenic *Escherichia coli*

Isabelle Bourgain<sup>(1)</sup>, H el ene Chiapello<sup>(1)</sup>, Christelle Hennequet-Antier<sup>(1)</sup>, St ephane Robin<sup>(2)</sup>, Sophie Schbath<sup>(1)</sup>, Alexandra Gruss<sup>(3)</sup> and Meriem El Karoui<sup>(3)</sup>

<sup>(1)</sup>Unit e Math ematique Informatique et G enome, INRA Domaine de Vilvert 78352 Jouy en Josas cedex

<sup>(2)</sup>D epartement Organisation et Mod elisation de l'Information et des Processus INA-PG 16, rue Claude Bernard 75231 Paris cedex 05

<sup>(3)</sup>Unit e de Recherche Laiti eres et G en etique Appliqu ee. INRA Domaine de Vilvert 78352 Jouy en Josas cedex. France. meriem@jouy.inra.fr

**Keywords** : genome comparison, Chi, homologous recombination, markov chain models

## Introduction

DNA double strand breaks are deleterious lesions which cause loss of genetic information and genomic rearrangements. They arise from normal cellular processes like chromosome replication [1] or because of exposure to DNA damaging agents such as UV light or oxygen. In *Escherichia coli* DNA molecules carrying a double-strand break are mainly dealt with via the RecBCD-Chi pathway [2]. RecBCD is a double strand exonuclease and a helicase (exo/hel). RecBCD exonuclease activity degrades DNA from a double strand end, but is attenuated when it encounters a short DNA motif called Chi (5'-GCTGGTGG-3', Chi is active only in this orientation, [3]). After a Chi encounter, the helicase activity remains, leading to formation of a single strand extremity (the preferential substrate for RecA-mediated homologous pairing reaction) and subsequent repair of the molecule by homologous recombination. This process allows faithful repair of the broken molecule by copying the missing genetic information from an intact copy of the chromosome. Thus Chi determines the fate of a DNA molecule with a double strand end: it will be repaired or degraded depending on whether it carries a Chi motif or not. It has been shown that Chi sites are very frequent on the K12 MG1655 (referred to as "K12" in this text) laboratory strain chromosome suggesting that they might be necessary for the conservation of genome integrity in this organism. We have previously used a statistical method (see below) to show that indeed Chi is over-represented on the K12 genome suggesting that its presence has been selected [4].

To check whether this property is conserved in the *E. coli* species, we made a comparative analysis of Chi distribution on the complete genome of K12 and a pathogenic (enterohemorrhagic) strain called 0157:H7 Sakai (referred to as "Sakai" in this text). The Sakai chromosome is 5,5Mb in length which is 0,85Mb larger than K12. Comparison of both genomes has revealed a "mosaic" structure : it is possible to define a common almost identical "backbone"(of around 4,1Mb) which is interrupted by strain specific sequences called "loops" [5]. There are 296 loops in Sakai and 325 in K12 with a size varying from 20 bp up to 85 kb. The total length of the K12 loops (K-loops) and the Sakai loops (S-loops) is respectively 537 kb and 1393 kb. Some of these loops arose by lateral gene transfer and are associated with virulence factors, others correspond to bacteriophages (bacterial virus) integrated on the genome [6].

## Methods

### Genome Sequences

We used the complete genome sequences from GenBank (Accession no : U00096 for K12 and BA000007 for Sakai). The origin and terminus of replication are annotated in K12 and were determined by sequence similarity search in Sakai (these regions are almost identical in both strains). The coordinates of the loops were kindly provided by Dr Kurokawa and Pr. Hayashi (Dpt of microbiology, Miyazaki Medical College, 5200 Kyotake, Miyazaki 889-1692 Japan).

### Statistical method

To study the statistical significance of the Chi frequency in the different genomes we compared the observed count of Chi to the estimated expected counts under several stationary Markov models. Depending on the order  $m$  of the model, it allows to take into account the base, dimer, ...  $(m+1)$ -mer composition of the sequence. The difference between both counts is then normalized such that the obtained score is asymptotically Gaussian with mean 0 and variance 1 [7,8]. The probability that the count is greater than the observed one, the so-called  $p$ -value, is the probability for a standard Gaussian variable to be greater than the observed score. We used the R'MES software [9] that also provides the rank of Chi's score with respect to the scores of all possible 65536 octamers sorted in decreasing order.

## Results

### General Properties of Chi Distribution :

In both strains 75% of Chi motifs are on the leading strand. This orientation allows efficient repair of the chromosome if it is broken at the replication fork. The average frequency of Chi is 1 every 4.6 kb in K12 and 1 every 4.9kb in Sakai. To check whether this difference can be explained by the mosaic structure of these genomes we analysed separately backbone and loops. The average frequency of Chi in the backbone is 1 every 4.35 kb whereas it is much lower in the loops (1 every 8.6 kb in K-loops and 1 every 8 kb in S-loops). Thus it seems that the somewhat lower frequency of Chi in Sakai genome is explained by its lower frequency in S-loops (note that the loops are much longer in Sakai than in K12).

### Over-representation of Chi

The lower frequency of Chi in the loops could be simply due to the difference in base composition in these loops (the bacteriophages for example have a lower GC%). To test whether this lower frequency is significant we used the statistical method described above to take into account the nucleotides composition. Here we present the results using the Markov models of order 1 and 6 where we take into account the mono- to di-nucleotides frequency (M1) and the mono- to hepta-nucleotides frequency (M6, note that this is the highest possible model as Chi is an octamer).

**Table1 : Chi is the most over-represented octamer on the backbone but is not over-represented in the loops.**

	K12		Sakai	
	backbone	K-loops	backbone	S-loops
Observed count	714	46	714	127
expected count	77	8	77	28
<b>M1 model</b> p-value	$<10^{-316}$	$10^{-33}$	$<10^{-316}$	$10^{-72}$
rank	1	51	1	67
expected count	569	38	567	118
<b>M6 model</b> p-value	$10^{-24}$	0.02	$10^{-24}$	0.15
rank	1	3919	1	11246

Chi is the most over-represented octamer in both backbones and in both models (Table1) indicating that the over-representation is very strong and is not due to the over-representation of sub-motifs. This result suggests that Chi frequent occurrence has very probably been positively selected on the backbone during evolution. In contrast, in S- and K- loops Chi is only slightly over-represented in the M1 model and is not over-represented in M6 model (Table1). This suggests that Chi occurrence has not been selected on the loops.

## Discussion

Taken together these results indicate that the selective pressure that led to the over-representation of Chi on the backbone (probably the chromosomal part inherited from the common ancestor) is of a different nature, and probably on a much longer time scale, than the one that led to the acquisition of the strain specific loops. These results also raise the question of whether one can distinguish between different types of loops : Chi should be over-represented on the loops that were acquired a long time ago.

## Present Work

We are now expanding our analyse to several complete genomes of other *E. coli* strains that have recently been published [10,11]. We are testing two softwares (MUMmer 2 [12]; and MGA [13]) to define the common backbone and the loops from a multiple comparison. We will define more precisely the nature of the loops to determine whether they are the result of an acquisition or a loss from the common ancestor and to distinguish the loops that have been acquired recently from the “old” ones. Beside the analysis of the Chi frequency in *E. coli* genomes (loops/backbone), we also want to study its distribution along the genome by comparing the observed r-scans (distances between occurrences) of Chi with the expected ones under a Markov model or a compound Poisson model [14,15]. The later has the advantage to take into account the exceptional frequency of Chi and to take a prior information on the sequence heterogeneity.

## Acknowledgements

This work is supported in part by the “programme Inter-EPST de Bio-informatique 2001” and the « ACI Microbiologie fondamentale et maladies infectieuses et parasitaires 2000» from the French Research Ministry.

## References

- [1] Gruss A and B. Michel, The replication-recombination connection: insights from genomics. *Curr Opin Microbiol.* 45:595-601, 2001
- [2] Kuzminov A. Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage lambda. *Microbiol Mol Biol Rev.* 63:751-813, 1999
- [3] Kowalczykowski SC, Dixon DA, Eggleston AK, Lauder SD, Rehrauer WM. Biochemistry of homologous recombination in *Escherichia coli*. *Microbiol Rev.* 58:401-65, 1994
- [4] El Karoui M, Biaudet V, Schbath S, et Gruss A Characteristics of Chi distribution on different bacterial genomes. *Res Microbiol.* 150:579-87, 1999
- [5] Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* 8:11-22, 2001
- [6] Ohnishi M, Kurokawa K, Hayashi T. Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors?. *Trends Microbiol.* 9:481-5, 2001
- [7] Prum, B., Rodolphe, F. and Turckheim, d. Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B.* 57 205-220, 1995
- [8] Schbath, S. An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Comp. Biol.* 4 189-192, 1997
- [9] Bouvier, A., Gélis, F. and Schbath, S. R'MES : Recherche de Mots Exceptionnels dans les Séquences d'ADN - version 2. Technical report, Guide de l'utilisateur. INRA, Biométrie, F78352 Jouy-en-Josas, 1999
- [10] Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A.* 99:17020-4, 2002
- [11] Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR. (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature.* 409:529-33
- [12] Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. 30:2478-83, 2002
- [13] Hohl M, Kurtz S, Ohlebusch E. Efficient multiple genome alignment. *Bioinformatics.* 2002 18 Suppl 1:S312-20, 2002
- [14] Robin et Daudin, Exact distribution of word occurrences in a random sequence of letters. *Journal of Applied Probability*, 36:179-193, 1999
- [15] Robin, S. A compound Poisson model for words occurrences in DNA sequences. *J. Royal Statist. Soc., C series.* 0. In Press , 2002