How to measure the robustness of bacterial genome comparisons?

Hugo Devillers, Helene Chiapello, Meriem El Karoui, Sophie Schbath

How to measure the robustness of bacterial genome comparisons?

Hugo Devillers[1], Hélène Chiapello[1], Meriem El Karoui[2], Sophie Schbath[1]

[1] Unité Mathématique, Informatique & Génome, UR1077 INRA,
Domaine de Vilvert, F-78350, Jouy-en-Josas France
[2] Unité Bactéries Lactiques et Pathogènes Opportunistes, UR888 INRA,
Domaine de Vilvert, F-78350, Jouy-en-Josas France
{hugo.devillers, helene.chiapello, meriem.el_karoui,
sophie.schbath}@jouy.inra.fr

Abstract: *The number of studies dealing with complete bacterial genome comparisons steadily increases. They allow us to gain insight into the molecular mechanisms involved in the evolution of bacterial genomes such as DNA exchanges. There exist several software tools and methods to align complete genomes and to determine conserved and variable regions. However, statistical methods to evaluate these tools are lacking. To fill this gap, two local scores for measuring the robustness of the comparisons of bacterial genomes are proposed. The calculation procedures of these scores are first presented and their interest is then discussed from two illustrative examples.*

Keywords: Comparative genomics, bacteria, conserved/variable segments, robustness, complete genome.

1 Introduction

The number of complete bacterial genome sequences available in public databases has considerably increased since the publication, in 1995, of the genome of *Haemophilus influenzae* that was the first bacterium to be completely sequenced [1]. There are currently more than 700 bacterial genomes entirely sequenced, representing about 250 distinct genera, and more than 1,200 other genomes will be available soon (see: http://www.ncbi.nlm.nih.gov/Genomes/, December 2008). Comparison of these genomes allows us to address new questions about their structure and their evolution [2]. Moreover, since the publication of a second strain of *Helicobacter pylori* in 1999 [3], the availability of genomes of closely related bacterial strains has rapidly increased. This offers new opportunities to gain insight into the understanding of short-term evolutionary processes, especially at the molecular level.

A comparison of two closely related bacterial genome sequences was performed by Hayashi *et al.* in 2001 [4]. An alignment of the two complete genomes of the enterohemorrhagic *Escherichia coli* O157:H7 Sakai strain and the *E. coli* K-12 MG1655 laboratory strain was performed. It allowed the determination of a highly conserved sequence between the two genomes, called the conserved backbone of the *E. coli* chromosome, which was interrupted by several DNA segments that were variable from one strain to the other. The backbone/variable segment structure is named segmentation. Its analysis is of great interest to study the molecular mechanisms involved in the

dynamics of bacterial genome evolution. Thus, for example, segments from the conserved backbone, which may correspond in large part to the common ancestral strain, have been shown to be enriched in functional DNA motifs [5]. Variable segments that may be associated to strain-specificities, are particularly relevant to study horizontal transfers, as they are probably associated to mobile elements such as prophages [6]. Consequently, the segmentation (backbone/variable segments) must be accurately determined. There exist various software tools to compare and to align bacterial genomes [2] and several databases store pre-computed comparisons such as xBASE [7] and MOSAIC [8].

The success of sequence alignment methods, such as BLAST or FASTA, lies, in part, in the evaluation of the statistical significance of the alignment score they provide. The genome comparison tools cited above generally suffer from a lack of statistical methods to evaluate their results [9]. To fill this gap, we propose two local scores measuring the robustness of the segmentations of bacterial genomes. In this paper, the calculation procedures of these two scores are first presented and their interest is then stressed from two illustrative examples.

2 Measuring the Segmentation Robustness

Here we present a method to measure the robustness of a segmentation (backbone/variable segments) obtained from the comparison of two genomes. Our method is based on a simulation process that aims at randomly perturb the original genomes.

2.1 Simulation Process

The determination of bacterial genome segmentation is generally based on the detection of the common elements between the compared sequences. Thus, to measure the robustness of such a procedure, it is relevant to perturb only conserved regions rather than random sequences chosen from the whole genomes. We therefore focus on maximal exact matches (MEMs), which correspond to common sequences between the compared genomes that cannot be extended (whose length is maximal). It is noteworthy that MEMs are frequently used as anchors to align complete genomes [10]. The nucleotides corresponding to a user defined proportion of these MEMs are randomly perturbed. Three types of perturbations are defined: 1) Deletions, MEM's positions are simply deleted; 2) Inversions, a MEM sequence is reverse-complemented and reinserted at the same position; 3) Translocations, two MEM sequences are switched. Perturbations are applied separately in each compared genome. The segmentation of the perturbed genomes is then computed and stored in a database. The process is repeated a sufficient number of times to ensure the statistical reliability of the scores defined below.

2.2 Score Definition

The measurement of robustness is based on the comparison of the segmentations of the perturbed genomes with the original segmentation. Two scores are derived, one focusing on the nucleotide robustness, the other one on the robustness of the segments. Considering the nucleotide *i* from one genome of the comparison, the nucleotide score is defined as follows:

$$S_{nuc}(i) = \frac{\#\{simulations \mid i \in variable\ segment\}}{\#\{total\ simulations\}}.$$

It is equal to the proportion of simulations in which the nucleotide *i* is assigned in a variable segment. Thus, S_{nuc} varies between 0 and 1. Its interpretation is the following: if $S_{nuc}(i)$ is near 1 then

i is likely to belong to a variable segment.

Considering the segment *g* of the original segmentation (*i.e.*, the non-perturbed segmentation), the segment score is defined by:

$$S_{seg}(g) = \frac{1}{|g|}\sum_{i \in g} S_{nuc}(i),$$

where |*g*| denotes the number of nucleotides in segment *g*. It is equal to the average of the nucleotide scores of the nucleotides belonging to segment *g*. Thus, if $S_{seg}(g)$ is close to 1 then the segment g is likely to be a robust variable segment.

3 Application to Two Segmentations in the *Escherichia coli* Species

3.1 Dataset Selection

We first compared the *E. coli* enterohemorrhagic O157:H7 Sakai strain and the *E. coli* K-12 MG1655 laboratory strain. The corresponding segmentation is available in the MOSAIC database (http://genome.jouy.inra.fr/mosaic/). This choice relies on the fact that this segmentation has been intensively studied and compares well to a manually curated dataset [4]. We also used a second segmentation based on the comparison of two *E. coli* K-12 strains: K-12 MG1655 and K-12 W3110. The segmentation was performed using the strategy developed for the MOSAIC database. Because these two genomes are almost identical, this segmentation is expected to be roughly constituted by a unique backbone segment. Surprisingly, it is not the case as 40% of the genomes appear in variable segments. This suggests that the segmentation strategy might need to be modified for such closely related genomes (see below). These two *E. coli* segmentations were used here to illustrate the interest of the two scores.

3.2 Nucleotide Score

For each selected segmentation, S_{nuc} was computed. After a preliminary investigation, it was decided to perturb 33% of the MEMs using a combination of the three types of perturbations described in section 2.1 and to perform 100 simulations. S_{nuc} values were then plotted for all the nucleotides of each genome. Three examples representative of the different score profiles are shown in Fig. 1.

Fig. 1A shows a first example for the K-12/Sakai strain segmentation, which is focused on a 5,000 bp variable segment. Along this region, S_{nuc} is equal to 1 at the variable segment and sharply decreases at the surrounding backbone segments. This strongly suggests that the nucleotides of the focused segment really belong to a variable segment. Fig. 1B displays another variable segment from the K-12/Sakai strain segmentation. Values of S_{nuc} indicate that although the assignment of the nucleotides of this variable segment is globally robust, the assignment for those located at the boundaries of the segment are less robust than the others.

Fig. 1C depicts S_{nuc} results for a variable segment of the comparison between the two *E. coli* K-12 strains. The very low S_{nuc} values along this segment reveal that the later is not robust and lead to suppose that it cannot be considered as a variable segment.

These three above examples of S_{nuc} profiles indicate that the nucleotide score allows us to precisely analyze the robustness of a segmentation along each nucleotide of a genome. It facilitates the detection of non or partially robust segments. Similar analyses were done along backbone

segments (not shown) and indicate that this score is also useful to analyze backbone segments.

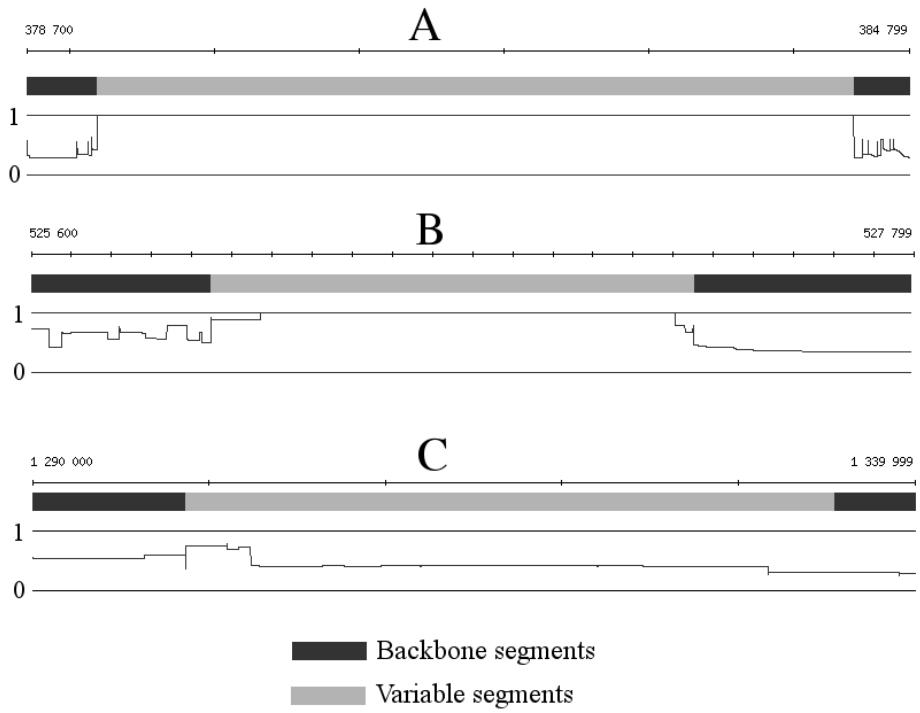

Figure 1. Nucleotide scores for three variable segments along the *E. coli* K-12 MG1655 genome from the segmentation of K-12/Sakai strains (A and B) and from the segmentation of the two K-12 strains (C). The axis at the top gives the nucleotide positions, the black and gray line shows the computed segmentation, and the curve (varying from 0 to 1) displays the nucleotide scores.

3.3 Segment Score

Computation of the segment scores (S_{seg}) was also performed on the two selected segmentations of the *E. coli* species. Fig. 2A displays the histogram of S_{seg} values for all the segments of K-12 MG1655 from the comparison of K-12/Sakai strains. This segmentation contains 617 variable segments and 618 backbone segments. The score distribution presents two peaks, one for the variable segments and the other for the backbone segments. Most of the variable segments (in gray in Fig. 2A) have a score between 0.99 and 1, indicating that they are robust. The backbone segments (in black in Fig. 2A) most often have a score ranging between 0.3 and 0.4. They are also probably robust. Indeed, the backbone being mainly constituted of MEMs, their percentage of perturbation will determine the expected value of a robust score for a backbone segment. Because in this study 33% of the MEMs were perturbed, robust backbone segment scores are expected to be around 0.33. Thus, from a rapid inspection of Fig. 2A, we can easily conclude that the whole segmentation of K-12/Sakai strains is robust.

Conversely, it is not the case for the segmentation of the two substrains of *E. coli* K-12 strains (Fig. 2B). This figure clearly shows that for most of the variable and backbone segments, the score values correspond to a low robustness. This is in agreement with the fact that the predicted segmentation contains unexpected variable segments while a unique backbone segment was expected. As a result we can conclude that the whole segmentation of the two substrains of *E. coli* K-12 strains is not robust.

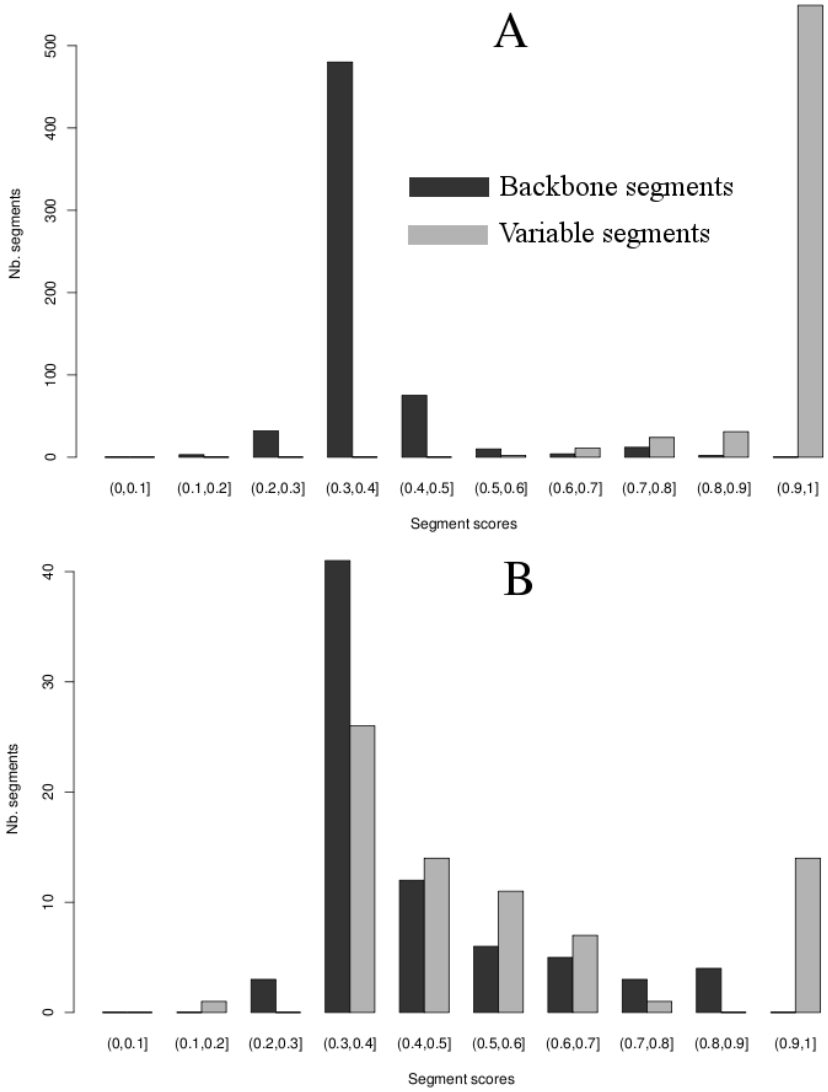

Figure 2. Segment scores for the segmentations of the *E. coli* K-12 MG1655 from the comparison of K-12/Sakai strains (A) and from the comparison of the two K-12 strains (B).

4 Concluding Remarks

To our knowledge, this study is the first attempt to statistically determine the robustness of bacterial genome segmentations. The two proposed scores, routed on classical statistics are simple to compute and easy to interpret. The examples presented here show that the proposed scores are able to distinguish robust and non robust segmentations. A statistical test will then be designed for this purpose. The nucleotide score (S_{nuc}) also allows to detect short non robust regions among a generally robust segmentation.

Such encouraging results have been also obtained from the analysis of several other segmentations from the MOSAIC database (data not shown). This suggests that the scores developed here could be used at a larger scale, for example on all comparisons stored in the MOSAIC database. To further validate our approach, we are also performing simulation studies on

artificial genomes for which the segmentation is known.

Comparison of multiple strains of a single species has also yielded the concept of species pan-genome as a measure of the whole gene repertoire that can pertain to a given bacterium [11]. Briefly, genes of the pan-genome are divided into three categories. The core-genome groups genes shared by all the strains, the dispensable genes correspond to those that are not present in each strain and last, the specific genes are observed in only one strain. In this context, it should be interesting to see whether genes of the core-genome belong to robust backbone segments as determined by the score calculations. This will be investigated in future works.

Acknowledgements

We are grateful to the INRA MIGALE platform (http://migale.jouy.inra.fr) for providing computational resources. We thank Annie Gendrault for her valuable help in database management. This work was supported by the French ANR (Agence Nationale de la Recherche) project CoCoGen (BLAN07-1_185484).

References

[1] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, K.S. McKenney, G. Sutton, W. Fitzhugh, C. Fields, J.D. Gocyne, J. Scott, R. Shirley, L. Liu, A. Glodek, J.M. Kelley, J.F. Weidman, C.A. Phillips, T. Spriggs, E. Hedblom, M.D. Cotton, T.R. Utterback, M.C. Hanna, D.T. Nguyen, D.M. Saudek, R.C. Brandon, L.D. Fine, J.L. Fritchman, J.L. Fuhrmann, N.S.M. Geoghagen, C.L. Gnehm, L.A. McDonald, K.V. Small, C.M. Fraser, H.O. Smith and J.C. Venter, Whole-genome random sequencing and assembly of *Haemophilus influenzae*. Science, 269:496-512, 1995.

[2] D. Field, G. Wilson and C. van der Gast, How do we compare hundreds of bacterial genomes? Curr. Opin. Microbiol., 9:499-504, 2006.

[3] R.A. Alm, L.S. Ling, D.T. Moir, B.L. King, E.D. Brown, P.C. Doig, D.R. Smith, B. Noonan, B.C. Guild, B.L. deJonge, G. Carmel, P.J. Tummino, A. Caruso, M. Uria-Nickelsen, D.M. Mills, C. Ives, R. Gibson, D. Merberg, S.D. Mills, Q. Jiang, D.E. Taylor, G.F. Vovis, T.J. Trust, Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature, 397:176-180, 1999.

[4] T. Hayashi, K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C.G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori and H. Shinagawa, Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res., 8:11-22, 2001.

[5] D. Halpern, H. Chiapello, S. Schbath, S. Robin, C. Hennequet-Antier, A. Gruss and M. El Karoui, Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling. PLoS genet., 9:153-160, 2007.

[6] H. Chiapello, I. Bourgait, F. Sourivong, G. Heuclin, A. Gendrault-Jacquemard, M.A. Petit and M. El Karoui, Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. BMC Bioinformatics, 6:171-180, 2005.

[7] R.R. Chaudhuri and M.J. Pallen, xBASE, a collection of online databases for bacterial comparative genomics. Nucleic Acids Res., 34:335-337, 2006.

[8] H. Chiapello, A. Gendrault, C. Caron, J. Blum, M.A. Petit and M. El Karoui, MOSAIC: an online database dedicated to the comparative genomics of bacterial strains at the intra-species level. BMC Bioinformatics, 9:498-506, 2008.

[9] W. Miller, Comparison of genomic DNA sequences: solved and unsolved problems. Bioinformatics, 17:391-397, 2001.

[10] M. Höhl, S. Kurtz and E. Ohlebusch, Efficient multiple genome alignment. Bioinformatics, 18:S312-S320, 2002.

[11] A. Muzzi, V. Massignani and R. Rappuoli, The pan-genome: towards a knowlegde-based discovery of novel targets for vaccines and antibacterials. Drug Discov. Today, 12:429:439, 2007.