Data Mining Using Hidden Markov Models (HMM2) to Detect Heterogeneities into Bacteria Genomes

Catherine Eng, Annabelle Thibessard, Sébastien Hergalant, Jean-François Mari, Pierre Leblond

Data Mining using Hidden Markov Models (HMM2) to detect heterogeneities into bacterial genomes

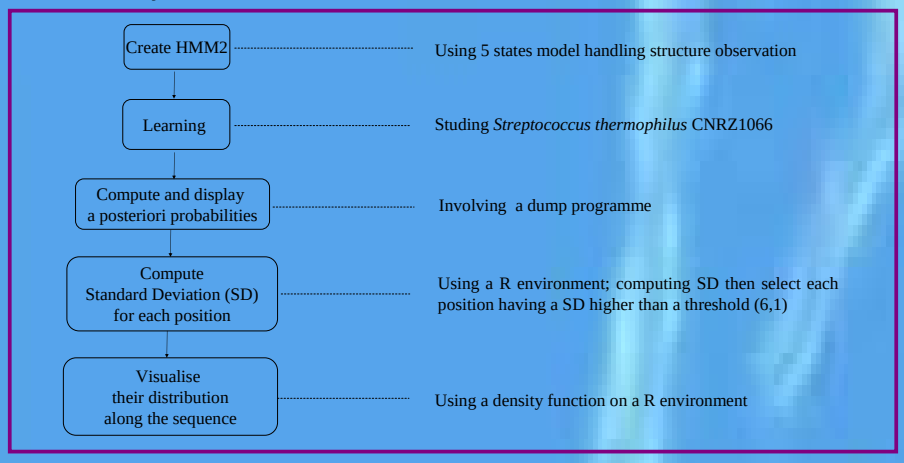
ENG Catherine[3], THIBESSARD Annabelle[1], HERGALANT Sébastien[3],
MARI Jean-François[2], LEBLOND Pierre[1]

[1] LGM, Laboratoire de Génétique et Microbiologie, UMR UHP-INRA 1128, IRF 110.
[2] LORIA, Laboratoire Lorrain en Recherche Informatique et ses Applications, équipe Orpailleur, UMR 7503.
[3] LGM et LORIA.

Introduction

The *Streptococcus* genus contains both pathogenic bacteria and bacteria used in the food-processing industry (such as *Streptococcus thermophilus*). The Horizontal Gene Transfer (HGT) involving streptococci could then have incidences on food safety and/or public health. We are developing a statistical segmentation method to identify heterogeneous sequences into the genome sequence of *S. thermophilus* CNRZ1066 (1,8 Mb)* and to identify the reason of their atypical reaction. The method based on Hidden Markov Models (HMM), without *a priori*, aims to detect unidentified HGT disregarding known features (GC%, sequence homology with frequently transfer sequences, and so on).

*Process (based on previous work**)*



Results

By this approach, 86 segments displaying an important density of high SD positions on a minimum 2,2 kb window, are dectected. The figure 1 represents the location of these 86 segments along the genome displayed on ARTEMIS software.

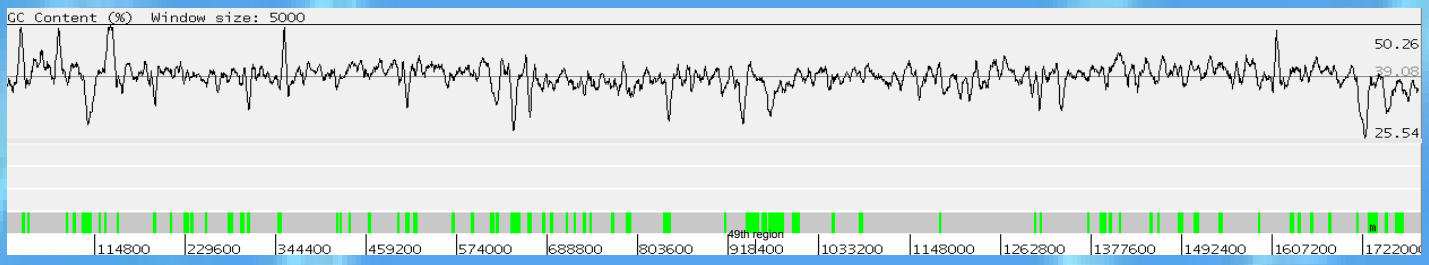


figure 1 : Distribution of the atypical segments along *S. thermophilus* CNRZ 1066 genome

The plot on the top of the figure 1 gives GC% along *S. thermophilus* CNRZ 1066 genome (39,1% GC on average). On the bottom of the figure, the grey strip symbolizes the genome and the green strips delimit the 86 atypical regions. The region overlapping the 49th segment is shown below (figure 2).

The top of the figure 2 is the result of the analysis of a part (about 46 kb) surrounding the 49th segment and displays the density function for all positions having a SD over 6.1. This density function is defined as follow :

$$d(x) = (P_x/\sum P_x)*100/L$$

L : Number of n nucleotides interval
P_x : Number of deviating positions into interval L_x

It appears that this density is higher from position 933 615 to 940 186, suggesting that this segment has a potential exogenous origin. This atypical area contains 5 genes flanked by two genes encoding transposases annotated tnp1239 and tnp1193 (bottom of the figure 2).

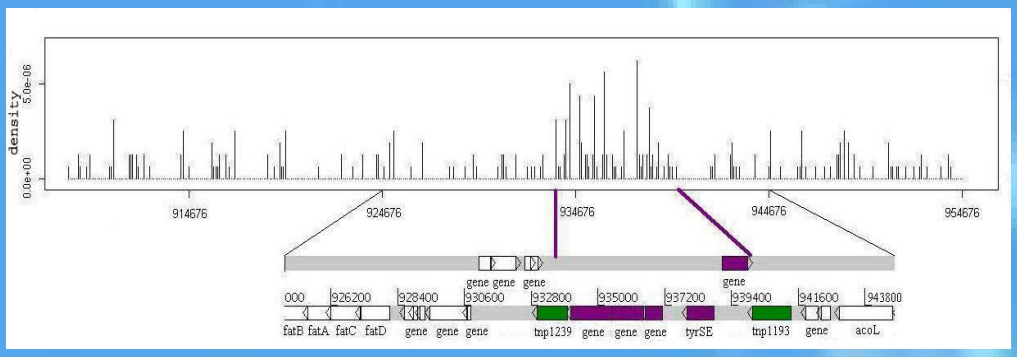


figure 2 : Density function of positions having SD > 6.1 around the 49th region

Moreover, a part of this segment is not detected in any other organism sequenced so far, including an other strain of *S. thermophilus* (LMG18311)*.

Conclusion and prospects

This method allows to detect 86 atypical segments and the extraction of the corresponding sequences reveals that:

- Some of them correspond to genomic islands already defined and supposed to be acquired by horizontal transfer. Among them, some contain IS (Insertion Sequence) and/or genes whose function is known to be frequently transferred (antibiotic resistance, restriction/modification system).
- The others are not labeled as exogenous DNA yet and will be the focus of the coming work, since they constitute potential uncharacterized HGT events.

Further work includes to go into detail in the sequence analyses of the unlabeled sequences to determine if their atypical feature is due to a foreign origin or to another particularity (genes encoding RNA, genes on the lagging strand). Moreover, this method could be performed using smaller window size limits. This method could be generalized to pathogenic *Streptococcus* genomes.

References

*Bolotin A. and *al.* **Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*.** *Nat Biotechnol.* 2004 Dec;22(12):1523-4. Accession no. CP000024 (CNRZ1061) and CP000023 (LMG18311).

S. Hergalant, B. Aigle, B. Decaris, J-F. Mari, P. Leblond. **Classification non supervisée par HMM de sites de fixation de facteurs de transcription chez les bactéries. Poster JOBIM 2004, Montreal.

JOBIM 2005 LYON, 6-8 juillet