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

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

To cite this version:


HAL Id: inria-00000142
https://inria.hal.science/inria-00000142
Submitted on 3 Jul 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Data Mining using Hidden Markov Models (HMM2) to detect heterogeneities into bacterial genomes

ENG Catherine³, THIBESSARD Annabelle¹, HERGALANT Sébastien³, MARI Jean-François², LEBLOND Pierre¹

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 thermophila*). 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)⁴ 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**)

- **Create HMM2**
  - Using 5 states model handling structure observation
- **Learning**
  - Studing *Streptococcus thermophilus* CNRZ1066
  - Involving a dump programme
- **Compute and display a posteriori probabilities**
  - Using a R environment, computing SD then select each position having a SD higher than a threshold (6.1)
  - Using a density function on a R environment
- **Visualise their distribution along the sequence**

Results

By this approach, 86 segments displaying an important density of high SD positions on a minimum 2.2 kb window, are detected. 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](image)

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).

![Figure 2: Density function of positions having SD > 6.1 around the 49th region](image)

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

Accession no. CP000024 (CNRZ1061) and CP000023 (LMG18311).