GENOFRAG: a software to design primers optimized for whole genome scanning by long-range PCR amplification. Application to the study of *Staphylococcus aureus* genome plasticity.

Nouri Ben Zakour[1], Michel Gautier[1], Rumen Andonov[2], Dominique Lavenier[2], Philippe Veber[2] , Alexeï Sorokin[3] and Yves Le Loir[1]

[1]Laboratoire d'Hygiène Alimentaire, UMR STLO, Institut National de la Recherche Agronomique, Ecole Nationale Supérieure Agronomique, 65, rue de Saint Brieuc, CS84215, 35042 Rennes cedex, France
Email: benzakou@epi.roazhon.inra.fr leloir@roazhon.inra.fr
[2]Institut de Recherche en Informatique et Systèmes Aléatoires, Campus de Beaulieu, 35042 Rennes cedex, France
Email: lavenier@irisa.fr
[3]Unité de Génétique Microbienne, Institut National de la Recherche Agronomique, Domaine de Vilvert, 78352 Jouy en Josas cedex, France.

Abstract
Bacterial genome plasticity can vary tremendously among strains of a given species –notably in *Staphylococcus aureus*. We analyze *S. aureus* genome plasticity by means of Whole Genome PCR Scanning (WGPS). Short chromosomes can indeed be entirely amplified by Long Range-PCR using a set of primers generated from a reference strain and used on several other *S. aureus* strains. Amplification profiles analysis can then reveal genome plasticity. For WGPS, we developed GenoFrag, a software for the design of optimized primers for LR-PCR on whole genomes. GenoFrag initially seeks all the potential primers on a chromosome. Then it calculates the best distribution of the primer pairs, thanks to combinatorial optimization algorithms. A set of primers was designed for the amplification of the chromosome of the reference strain *S. aureus* N315, in fragments of ~10 Kbp overlapping on ~1Kbp. Part of these primers was used for the amplification of 2 portions of N315 chromosome. This work shows that i) GenoFrag is a fast, reliable and versatile tool for primer design and ii) LR-PCR can be considered for the analysis of bacterial genomic plasticity.

Keywords. Long Range PCR, Primer Design, Genomic Plasticity, Biodiversity, *Staphylococcus aureus*, Combinatorial Optimization.

Introduction

Genomes can be highly heterogeneous in conspecific bacterial strains. Reasons of this plasticity are chromosomal rearrangements, spontaneous deletions and or mutations, acquisition of mobile genetic elements, plasmids… Until recently, genome variability analysis in bacteria relied on approaches taking into account the whole DNA content e.g. the analysis of restriction enzyme digestion patterns. These techniques rely on uncharacterized genomic differences between strains of a bacterial species and were proved highly discriminatory. The development of sequencing projects allowed approaches focused on a few sequences e.g. the multilocus sequence typing that uses housekeeping gene sequences or methods using mobile genetic elements as molecular markers. These approaches can efficiently discriminate closely related strains and, to some extent, can highlight genomic differences from a strain to another. Nevertheless, they only partly reflect the genomic diversity and do not allow an identification of the genetic changes (e.g. chromosomal rearrangement, horizontal transfer…). Increasing numbers of complete genome sequences of prokaryotic organisms allows whole genome comparisons, a powerful and accurate approach to genome diversity. Strains of *Staphylococcus aureus*, a Gram-positive pathogenic bacterium, are genomically and phenotypically highly heterogeneous. *S. aureus* is the causative agent of a wide range of diseases in warm-blooded animals. It is notably a major cause of nosocomial disease worldwide and is also often involved in food poisoning outbreaks [2]. As the sequences of seven *S. aureus* strains are now available (three are complete and four are under annotation), we have started a whole genome scanning of *S. aureus* strains by long range polymerase chain reaction (LR-PCR).

Whole Genome PCR Scanning Problematic

This approach is based on a comparative analysis of the whole genome structures of different conspecific strains, as determined by whole genome amplification using LR-PCR technique [3]. The amplification profiles of these strains are compared to a reference profile in order to highlight plasticity phenomenon as chromosomal rearrangements, deletion or acquisition of mobile genetic elements. Although many efficient tools are available to design and to test the robustness of PCR primers [4], there is no software that can process a whole genome

sequence to design primers taking into account the specific requirements of a whole genome PCR scanning project. Here, we describe GenoFrag, a software that automatically designs primers optimized for this purpose.

GenoFrag Software

GenoFrag is composed of two parts, the generation of primers and the segmentation of the genome.

1) Generation of primers

This software identifies all the primers suitable for a LR-PCR. It acts as a suite of 6 filters. All the potential K-mers (default value = 25) are first considered and input to filter #1. Each filter output only the oligonucleotides satisfying specific constraints set by the user, as G+C content, thermodynamic stability, excluded regions, secondary structures and secondary binding sites to be avoided. In order to limit the computation time, the filters having the highest selectivity are the first activated. All the oligonucleotides, which successfully pass the 6 filters are promoted as primers for LR-PCR. In addition, they are labeled with their position in the genome and their ability to start or end an amplicon.

2) Segmentation of the genome

This second software aims to provide a list of amplicons for optimal covering of the whole genome, or a part of it, using the set of primers previously generated. Constraints are the minimal and maximal length of the amplicons together with the minimal and maximal overlap allowed. If we assume, for the sake of simplicity, that a solution is made of a list of N amplicons, and that each amplicons can take only P different positions, then the number of possibilities is equal to P^N. When N is large, to find the best possibility is clearly a combinatorial problem: enumeration of all the possibilities is impossible, even with the fastest computers. Computational optimization methods are required to solve this problem in a reasonable time.

Two solutions of this problem have been implemented. The first one finds an optimal list of amplicons whose size are as close as possible of an ideal length. For example, if the minimal and maximal length of the amplicons are respectively of 9Kilobase pairs (Kbp) and 11Kbp, then the ideal length is calculated to be the average (9+ 11)/2 = 10Kbp. The second solution finds an optimal list of amplicons, which minimizes the spread between the shortest and the longest one. For both solution (i) we formulate a suitable combinatorial optimization model and (ii) we program a dedicated graph algorithm for solving these models [1]. For both programs, the output is a list of primers pairs.

Validation of GenoFrag

The computation time for processing 1Mbp ranges from one to two minutes on a 1.5 GHz PC depending on the number of primers selected during the first step. Of course, the larger the number of primers, the longer the computation time. To validate GenoFrag, a set of primers was designed for the amplification of the chromosome of the reference strain *S. aureus N315*, in fragments of ~10 Kbp overlapping on ~1Kbp. Part of these primers was successfully used to amplify 2 portions of chromosome of 110 Kbp each.

Conclusion

In this work, we developed GenoFrag, a software dedicated to the design of primers optimized for whole genome PCR scanning. We demonstrate here that GenoFrag can be successfully used to generate primer pairs for the amplification of portions of *S. aureus* chromosome. GenoFrag is thus a powerful and robust tool for WGPS applied to the evaluation of *S. aureus* genome plasticity. It is versatile as well since GenoFrag can be used for other bacterial or viral species.

References

[1] Andonov, R., Yanev, N., Lavenier, D. and Veber, P. (2003) Combinatorial approach for segmenting bacterium genome. *IRISA research report*, PI1536.

[2] Le Loir, Y., Baron, F. and Gautier, M. (2003) *Staphylococcus aureus* and food poisoning. *Genet. Mol. Res.*, **2**(1), 63-76. http://www.funpecrp.com.br/gmr/

[3] Ohnishi, M., Terajima, J., Kurokawa, K., Nakayama, K., Murata, T., Tamura, K., Ogura, Y., Watanabe, H. and Hayashi, T. (2002) Genomic diversity of enterohemorrhagic *Escherichia coli* O157 revealed by whole genome PCR scanning. *Proc. Natl. Acad. Sci. USA*, **99**, 17043-17048.

[4] Rozen, S., Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365-386.