Molecular surveillance and subtyping of *Listeria monocytogenes* with BioNumerics

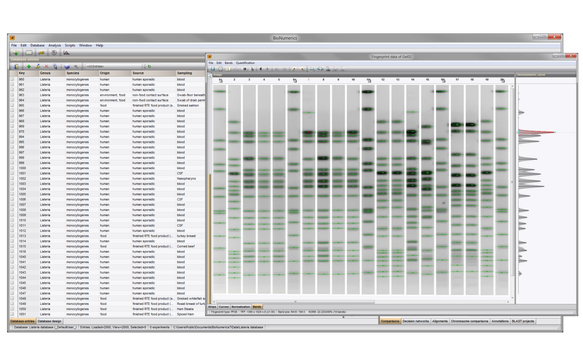
Johan Goris[1], Hannes Pouseele[1], Benjamin Felix[2], and Koen Janssens[1]

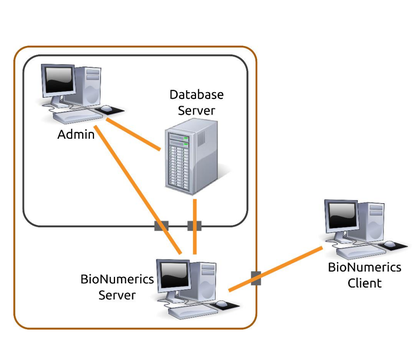
[1]: Applied Maths NV, Keistraat 120, B-9830 Sint-Martens-Latem, Belgium
E-mail: info@applied-maths.com – Phone: +32 9 2222 100
[2]: ANSES - Laboratoire de sécurité des aliments de Maisons-Alfort, France

Introduction: *Listeria monocytogenes* is a ubiquitous organism in the environment and a rare cause of human disease. Even though listeriosis occurs infrequently, it is characterized by a high case-fatality rate which can exceed 30% percent. The high morbidity and mortality of this infection make a strong case for the importance and priority of molecular surveillance platform for the pathogen.

MOLECULAR SURVEILLANCE TODAY: PFGE

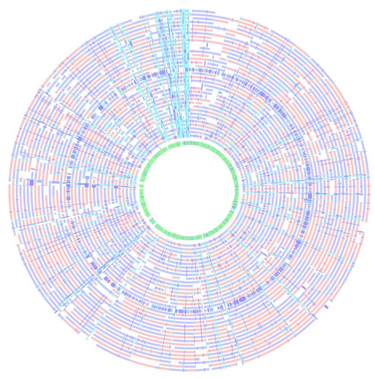
Pulsed Field Gel Electrophoresis: In 2006, ANSES has been designated as European Union Reference Lab (EURL) for *L. monocytogenes* and has developed a standardized method for Pulsed Field Gel Electrophoresis (PFGE), which was implemented as a first technique in the EURL *L. monocytogenes* database in 2011. The BioNumerics software suite acts here as a backbone for data preprocessing, submission, curation, identification and cluster detection. Until today, PFGE is considered the "gold standard" in most surveillance networks. Its major drawbacks are that the technique is time-consuming, labor-intensive and difficult to standardize across laboratories.





MOLECULAR SURVEILLANCE TOMORROW: WGS?

Whole genome sequencing: As next-generation sequencing (NGS) technology advances, sequencing errors decrease to rates equal to or lower than traditional Sanger sequencing. At the same time, prices drop rapidly so we can anticipate that whole genome sequencing (WGS) will become affordable for routine molecular surveillance in a few years. WGS obviously holds the promise to provide extremely detailed information about any strain analyzed. However, the technology and especially the data analysis still needs to be standardized. For subtyping in the context of molecular surveillance, there is no need for a fully closed and annotated genome. Instead, for the sake of interpretation, the complexity of the data needs to be reduced in a biologically meaningful way. Two different workflows are currently being implemented in the BioNumerics software package, i.e. genome-wide SNPs and gene-by-gene systems.



Circular representation of 35 whole genome sequences

GENOME-WIDE SNPS

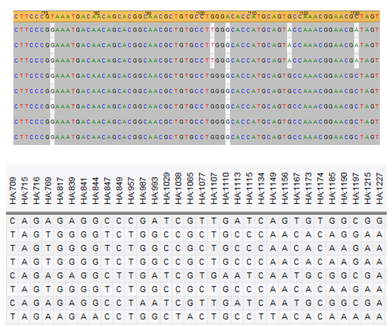
Principles: In a genome-wide SNP analysis, only Single Nucleotide Polymorphisms (SNPs) within a set of samples are taken into account. This excludes indels, inversions, and translocations. A consistent position numbering (i.e. against a reference genome) is maintained.

Workflow in BioNumerics:

- Map reads against a reference sequence
- Calculate consensus sequence
- Find mutating positions in samples
- Store as temporary character set

➔ **Use for clustering, identification,…**

…ccttgccaannnnnnnggtgggggtcgcgggttcgaatcccgtcgtt…

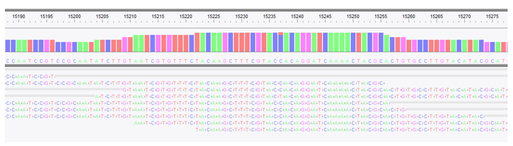


GENE-BY-GENE SYSTEMS

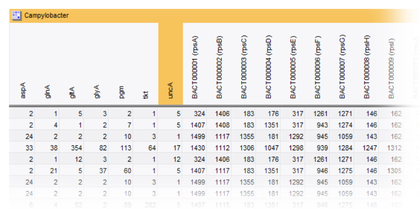
Principles: Gene-by-gene systems are the NGS variant of "classical" Multi Locus Sequence Typing (MLST). Instead of using targeted sequencing of specific loci, whole genomes are shotgun sequenced and alleles identified with respect to a reference set of loci. A type is defined by a specific list of allele numbers.

Workflow in BioNumerics:

- *De novo* assembly of reads
- Allele lookup on contigs using BLAST
- Store allele numbers in character experiment
- Select gene subset for analysis



Mega BLAST: for highly similar sequences



- MLST (7 housekeeping genes)
- rMLST (ribosomal genes)
- cgMLST (core genome)
- wgMLST (whole genome)
- …

Evaluation: The genome-wide SNP approach seems to provide a very detailed answer to the subtyping question. A fundamental question with the technique is how many SNPs would be allowed between isolates from the same outbreak? In fact, this question might prove particularly hard to answer since a single evolutionary event can introduce several SNPs. Another major drawback is that the data set in principle is unstable, because the SNP set (i.e. the number of polymorphisms found) depends on the set of isolates analyzed.

Evaluation: In gene-by-gene systems, the resolution can be chosen by selecting an appropriate subset of loci, resulting in detailed up to very detailed data. Some issues remain, e.g. with paralogous genes and multiple gene copies. Specifically for wgMLST, the situation arises where a certain gene is present in some but not in all samples. For this reason, it might be better to limit the set of loci to the core genome (cgMLST). The technique generates a stable data set that can be the basis for an international nomenclature, provided that allele IDs and sequence types are properly curated. In light of the large number of loci, automated curation pipelines should be developed to make this a feasible task.

APPLIED MATHS
www.applied-maths.com

Conclusions: Rapid and automatic processing of WGS data is needed to ensure a reliable and easy to follow workflow in routine molecular surveillance, reducing the time needed to detect and contain potential outbreaks, eventually reducing the cost on public health and food safety. Existing BioNumerics client-server infrastructure can be extended to accommodate WGS data, which can be used in parallel with current typing methods.