**PREDIcting bacterial PATHogencity on plant: PREDIPATH**

Felipe LIRA¹, Gilles HUNAUT³, Martial BRIAND¹, Perrine PORTIER¹, Claudine LANDES¹, and Marion FISCHER-LE SAUX¹

¹IRHS, INRA, Université d’Angers, Agrocampus Ouest, SFRA 4207 Quadsav, 49071, Beaucouzé, France.
²Haodynamique, Interaction Fibrose et Invasivité Tumorales Laboratoire, Unité Propre de Recherche de l’Enseignement Supérieur 3693, Structure Fédérative de Recherche 4128, Bretagne Loire Université, Angers, France.

**Backgrounds**

Prediction of bacterial pathogenicity commonly relies on microbiological methods. **Comparative genomics** emerges as a efficient method for distinction and detection of genomics elements able to distinguish two or more classes of organisms (pathogenic vs. non-pathogenic; commensal vs. free-living organisms). Genes, genes clusters, and operons, are closely associated with the bacterial survival and spread. Most of them are exclusive and determinant to characterize bacterial groups. In order to facilitate the prediction of potential bacterial plant-pathogenicity of plant-associated bacteria, we propose the PREDIPATH workflow.

**Objective**

Creation of specific datasets of genes, clusters and sequence markers (kmers) to discriminate bacterial species based on their genomic sequences.

**Development**

The PREDIPATH methodology relies on the detection of genome-based markers and creation of specific datasets of markers enabling to identify potential pathogenic organisms based on their genomes. PREDIPATH pipeline was developed using Python programming language and external bioinformatics tools in the process. Our methodology for detection of markers is summarized in three major steps:

1. **Query genome**
   - Selection of bacterial group and download of genomes
   - Genome assessment² and phylogenomics
   - Manual curation and classification of genomes
   - Detection of genomic markers
     - Pathogenicity related genes
     - Secondary metabolites clusters
     - Kmers
2. **Detection of genomic markers**
   - PREDIPATH DB
   - Antismash²
   - DBGWAS²
   - Selection of markers and statistical analysis
   - Specific markers for each class of bacteria
     - Markers class 1
     - Markers Class 2
     - Markers Class n
3. **Statistical analysis**
   - Markers class 1
   - Markers Class 2
   - Markers Class n

**Bacterial classification based on their profile**

**Step I** prioritizes the correct assignation of genomes in their classes and the correction of their nomenclature when needed. **Step II** consisted in to create a customized database to detect potential genes to be used as markers - PREDIPATH Database (a priori approach); the detection of differential secondary metabolites clusters, and small DNA fragments, such as kmers exclusive for each class of organisms described in **Step I**. **Step III** gave the results for the obtained in **Step II**.

**Conclusions**

- **Phylogenetic distribution was not able to distinguish between pathogenic and non-pathogenic organisms in genus Erwinia.**
- **Our approach enable the compilation of a complete genomic datasets, composed by genes, clusters and kmers.**
- **Detection of exclusive markers by comparative genomics using the PREDIPATH workflow allowed the creation of exclusive datasets of predictors to diagnostic potential pathogenicity of plant-associated bacteria.**

**68 Genomes downloaded**

<table>
<thead>
<tr>
<th>Genomes</th>
<th>Total</th>
<th>Sec.Metabolites</th>
<th>Kmers</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDIPATH DB</td>
<td>14248</td>
<td>24</td>
<td>143,478</td>
</tr>
<tr>
<td>Non-pathogenic Pathogenic</td>
<td>213</td>
<td>9</td>
<td>512</td>
</tr>
<tr>
<td>Genomic elements</td>
<td>5</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Total detected markers</td>
<td>213</td>
<td>9</td>
<td>512</td>
</tr>
<tr>
<td>Predictors* of classes</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Simple binary logistic regression with PREDIPATH DB results were able to define a profile to predict the potential pathogenicity of plant-associated species:

- fur transcriptional repressor of iron-responsive genes
- htrp type III secretion lipoprotein
- htrp type III secretion protein
- htrp hypersensitivity response secretion protein
- part fluoroquinolones resistance gene

**Markers exclusive to NP class were present from 19 to 100% of genome; exclusive kmers in class P were distributed between 7 to 53% of the genome.**