Scientific Report of Luisa Filippin on the STSM

(ref. code: COST-STSM-FA0807-05285)

at INRA, UMR 1090 Génomique Diversité et Pouvoir Pathogène,

in Villenave d'Ornon, France

29/11/2009 – 13/12/2009

The topics of my mission at the INRA lab were the evaluation and comparison of the genetic variability, incidence and geographic distribution of FD related phytoplasma strains inhabiting grapevines and wild plants in France and Italy, and the characterization of genes coding for putative surface membrane proteins recently identified in FD phytoplasmas. My main objective was to acquire competences in the use of bioinformatic tools for sequence analysis and comparison.

During my stay, first I focused on the analysis of 4 Italian isolates, representative of different phytoplasmas as determined by RFLP or sequence analysis of 16S-23SrRNA, *rplV-rpsC* and *secY* genes. **Vv-AO262** was isolated from grapevine and is representative of FD-Lomb/Piem strain; also **Vv-SI257** was isolated from grapevine and is representative of FD-CCI; **Cl-AL31** was isolated from clematis and is representative of FD-PS; **Cl-UD147** was isolated from clematis and is an FD-C insertion mutant (Filippin *et al.* 2009).
Five genes were chosen as markers for a wider Multi-Locus Sequence Analysis: *map*, *urvB-degV* (Arnaud *et al.* 2007), *rplV-rpsC* (Lee *et al.* 2004, Martini *et al.* 2007), *tuf* and *rplF*. All of them are housekeeping genes; only *degV* is a hypothetical protein gene. For typing, nested PCRs followed by sequencing were performed.
The resulting chromatograms were used to obtain a consensus sequence of each gene. To do this, different softwares were used. First, the chromatograms were analysed by **Phred** program; it applies statistical methods in order to examine the trends of the 4 nucleotides in the region around each base. Specifically, it calculates the mean distance between 2 peaks in the considered region and then the theoretical position of each peak in the case that all are exactly at the same distance; in the following phase, it analyses the 4 traces independently, considering the real positions and the areas of each peak relating to the neighbouring ones; finally, an algorithm allows to relate the real peaks to inferred values obtained in the first phase. To calculate the quality of each base Phred values different parameters obtained from the profiles of the peaks in the surrounding region, for example the distance between the peaks and the ratio among the peaks' highness. The score calculated in this way may be empirically related to the probability that each base was attributed in a wrong way:

$$\text{Phred-score} = -10 \cdot \log_{10} (\text{probability of error})$$

At the end of the process, Phred generates a file with a final base sequence (or read), in which a reliability value is assigned to each base.
Then the different reads obtained from each gene are to be assembled. An assembling program has to be able to value all the possible sequence overlaps in both directions, and to determine the best alignment solution; then it has to generate a consensus sequence and attribute a reliability value for each base. In my stage I used mainly the **Phrap** (Phragment Assembly Program) software.

Phrap is incorporated in a series of automatic procedures defined in a script that launched the programs for the sequence elaboration, i.e. Phred, Phrap and others for completing the assembly (for example with information on the chemistry of sequencing an on the plasmid sequence; these programs were originally developed for the analyses of genomic libraries). The result is a final alignment corresponding to the best hypothesis taking into account the quality assigned by Phred to each read; each base has a score that is calculated from the original sequences. What Phrap is not able to do is visualising the assembly results. To do this other programs are to be used, for example **Consed**. These software are interactive and use advanced graphic interfaces for a user friendly visualisation of the consensus sequences. For each sequence it is possible to access to the list of the reads composing the consensus for comparing the quality. The base quality can be valued by eye according to its grey intensity; besides the reliable bases are in capital characters, the other in lower case. A lot of other possibilities are available for the editing of the final consensus sequence. When the result is satisfactory, a text file is exported and saved in FASTA format.
In my experience the sequencing of 2 amplicons gave chromatograms of bad quality that were not considered by the Phred-Phrap script; in these cases they were analysed with the **preGap4-Gap4** softwares, which are programs similar but simpler then Phred-Phrap, and less stringent in sequence analysis. Gap4 is a software for visualization of the final consensus, like Consed. At the end the result is again a text file in FASTA format.
The data obtained from the 4 Italian samples were used to complete the work made by the French group on a lot of samples isolated from grapevine, alders, insects collected manly in France, with contribution of a few isolates from other countries. This Multi-Locus Sequence Typing was designed to achieve the description of grapevine, alder and clematis phytoplasmas from the 16SrV group as a *Candidatus* species. The preliminary data were presented at the XVI ICVG that took place last August in Dijon (Malembic-Maher *et al.* 2009).
For each gene a unique text file containing the FASTA sequences of all the isolates was generated and submitted to **MEGA4**, a software used to reconstruct phylogenetic relationships among isolates applying different methods. This program was first used for alignment of the sequences with ClustalW application; the resulting data were checked for each mismatch, in order to compare and evaluate the differences - SNP, insertions or deletions. For example when unique SNPs present in only one sample or length mutations in coding regions were found, the corresponding consensus sequences were reanalysed with Consed to see if the differences are real or artefacts. The further step was the launch of the phylogenetic analysis. The Maximum Parsimony method was chosen; the program calculates all the possible trees resulting from the given sequences, and chose the simplest and shorten way that reflect that set of mutation; it's based on 2 principles: the sites evolved independently and the mutation rate was slow and constant in time. The reliability of the trees is evaluated by bootstrap method, based on performing a big number of tree replicates and taking into account the percentage of the times each branch is repeated in the different trees. Analysing the resulting cladograms for each gene we found different grouping for 2 of the Italian samples. This may be explained by events of recombination among genomes of different phytoplasma strains; but the data need to be reconfirmed, so I had them reamplified and resequenced; nowadays I'm waiting for the new results.
At the end a paper on the genetic diversity of European phytoplasmas of the 16SrV taxonomic group and the proposal of *'Candidatus* Phytoplasma caudwellii' and *'Candidatus* Phytoplasma rubi' as new taxa will be produced.

The *imp* gene of phytoplasma encodes for an immunodominant membrane protein which is thought to be involved in host-phytoplasma interactions; the gene was characterized in some phytoplasma species of the 16SrI, II, III, X, XI and XII groups (Kakizawa et al., 2006; 2009). In my lab several

primers were designed and tested for the amplification of this gene on isolates of the 16SrV group. The best results were obtained with 2 couples of primers which amplify about 800-1000 nt. The amplicons of about 20 different FD and related phytoplasma isolates from grapevine, clematis and alder were amplified and sequenced. The nested PCR amplified a region comprehensive of *imp* gene and of part of the flanking genes at both sides.
For the *in silico* characterization of *imp* gene at the INRA lab, the sequence obtained from FD70 phytoplasma strain was considered. The consensus sequence, obtained submitting 3 chromatograms to Phred-Phrap-Consed package, was first analysed with a program for identification of coding sequences (CDS). In the INRA lab **FrameD** program is used. The working matrix of FrameD for the codon usage and the identification of the start codon was composed of 40 genes already identified from the sequencing of the *Flavescence dorée* phytoplasma chromosome. In this way the start codon are more reliably found; moreover, the program can search upstream the putative star codon for RBS sequence, that is a sequence complementary to ribosomal RNA typical of prokaryotes; if it's present, the start codon is identified as a "high probability" one. The program revealed the presence of the final part of a CDS, then an entire CDS of 462nt and then again the final part of a third CDS, as expected. Then with **ExPasy-Translate Tool** software the complete CDS of *imp* gene was converted in aminoacid sequence; the program gave also information on the predicted protein as for example the number of residues (153), the Molecular Weight (17396 Da) and the Isoelectric Point (9.32). From here, the protein was compared with those present in public databases by means of **BLAST tool.**
In parallel the study on genetic variability of this protein was implemented with French samples; during the STSM, I made PCR on *imp* gene of 12 different isolates from grapevine and alders collected in different parts of France, then I sent them for sequencing; I'm waiting for the results.
All these data will converged in a paper focused on the genetic variability of *imp* gene of phytoplasma 16SrV group, together with description of *in silico* analysis on imp protein, compared also with the predicted structure of the same proteins from phytoplasma of other groups.
Finally, besides the methodological competences I acquired during my STSM, I also brought in Italy scions of periwinkle plants infected with 3 different French phytoplasma: FD92, FD-CAM and FD-PEY, belonging to 16SrV-C or –D group. In my lab they were grafted on healthy periwinkles, in order to maintain sources of phytoplasma available for further studies on phytoplasma genomics.

References

Arnaud, G., Malembic-Maher, S., Salar, P., Bonnet, P., Maixner, M., Marcone, C., Boudon-Padieu, E. & Foissac, X. (2007). Multilocus sequence typing confirms the close genetic inter-relatedness between three distinct Flavescence dorée phytoplasma strain clusters and group 16SrV phytoplasmas infecting grapevine and alder in Europe. *Applied and Environmental Microbiology* **73**, 4001-4010.

Filippin, L., Jovic, J., Cvrkovic, T., Forte, V., Clair, D., Tosevski, I., Boudon-Padieu, E., Borgo, M. & Angelini, E. (2009). Molecular characteristics of phytoplasmas associated with Flavescence dorée in clematis and grapevine and preliminary results on the role of Dictyophara europaea as a vector. *Plant Pathol* **58**, 826-837.

Kakizawa, S., Kenro, O., Namba, S. (2006). Diversity and functional importance of phytoplasma membrane proteins. *Trend Microbiol*., **14**, 254-256

Kakizawa, S., Oshima, O., Ishii, Y., Hoshi, A., Maejima, K., Jung, H.-Y., Yamaji, Y., Namba, S. (2009). Cloning of immunodominant membrane protein genes of phytoplasmas and their *in planta* expression. *FEMS Microbiol. Lett*., **293**, 92-101

Lee, I.-M., Martini, M., Marcone, C. & Zhu, S. F. (2004). Classification of phytoplasma strains in the elm yellows group (16SrV) and proposal of '*Candidatus* Phytoplasma ulmi' for the phytoplasma associated with elm yellows. *International Journal of Systematic and Evolutionary Microbiology* **54**, 337-347.

Malembic-Maher, S., Salar, P., Carle, P., Foissac, X. (2009). Ecology and taxonomy of Flavecence dorée phytoplasmas: the contribution of genetic diversity studies. *Proceedings of the XVI International Congress of Virus and virus-like diseases of Grapevine, Dijon, France, 31 Aug-4 Sept 2009*, pp 132-134

Martini, M., Lee, I.-M., Bottner, K.D., Zhao, Y., Botti, S., Bertaccini, A., Harrison, N.N., Carraro, L., Marcone, C., Khan, A.J., Osler, R. (2007). Ribosomal protein gene-based phylogeny for finer differentiation and classification of phytoplasma. *International Journal of Systematic and Evolutionary Microbiology* **57**, 2037-2051.