Análisis de *Pseudomonas* Fitopatógenas Usando Métodos Inteligentes de Aprendizaje: Un Enfoque General Sobre Taxonomía y Análisis de Ácidos Grasos Dentro del Género *Pseudomonas*

Analysis of Plant-Pathogenic Pseudomonas Species Using Intelligent Learning Methods: A General Focus on Taxonomy and Fatty Acid Analysis Within the Genus Pseudomonas

Bram Slabbinck, **Bernard De Baets**, Research Group Knowledge-based Systems, Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, 9000 Ghent, Belgium, **Peter Dawyndt,** Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, 9000 Ghent, Belgium y **Paul De Vos**, Research Group Knowledge-based Systems, Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, 9000 Ghent, Belgium.

(Recibido: Septiembre 20, 2009 Aceptado Enero 12, 2010)

Slabbinck, B., De Baets, B., Dawyndt, P., y De Vos, P. 2010. Análisis de especies de *Pseudomonas* fitopatógenas usando métodos inteligentes de aprendizaje: un enfoque general sobre taxonomía y análisis de ácidos grasos dentro del género *Pseudomonas*. Revista Mexicana de Fitopatología 28:1-16.

Resumen. La identificación de bacterias fitopatógenas es de alta relavancia. En este trabajo se evaluó la identificación de especies fitopatógenas dentro del género *Pseudomonas* mediante análisis de esteres metílicos de ácidos grasos (FAME). A partir de una base de datos de FAME, se han generado conjuntos de conjuntos de datos de alta calidad. Dos aspectos fueron investigados: la separación de especies fitopatógenas de *Pseudomonas*, y la diferenciación del grupo de espcies fitopatógenas de *Pseudomonas* de las no fitopatógenas. En la primera fase se realizó un análisis de componentes principales para evaluar la variabilidad de los datos. Posteriormente el método de aprendizaje árboles aleatorios fue evaluado para propósitos de identificación. El método inteligente permite aprender de la variabilidad y los patrones de los datos y mejorar la identificación de especies. El análisis de componente principal de especies fitopatógenas mostró claramente sobreposición de grupos de datos. Se desarrolló un modelo de árboles aleatorios que permitió alcanzar una eficiencia de identificación de especies del 71.1%. Discriminar el grupo de especies fitopatógenas del

Slabbinck, B., De Baets, B., Dawyndt, P., y De Vos, P. 2010. Análysis of plant-pathogenic *pseudomonas* species using intelligent learning methods: A general focus on taxonomy and fatty acid analysis within the genus *Pseudomonas*. Revista Mexicana de Fitopatología 28:1-16.

Abstract. The identification of plant-pathogenic bacteria is often of high importance. In this paper, we evaluate the identification of plant-pathogenic species within the genus *Pseudomonas* by fatty acid methyl ester (FAME) analysis. Starting from a FAME database, high quality data sets were generated. Two research questions were investigated: can plant-pathogenic *Pseudomonas* species be discriminated from each other and can the group of plant-pathogenic *Pseudomonas* species be distinguished from the group of non-plant-pathogenic *Pseudomonas* species. In a first stage, a principal component analysis was performed to evaluate the variability within the data. Secondly, the machine learning method Random Forests was evaluated for identification purposes. This intelligent method allows to learn from the variability and patterns in the data and to improve the species identification. The principal component analysis of plant-pathogenic species clearly showed overlapping data clouds. A Random Forests model was developed that achieved a species identification performance of 71.1%. Discriminating the group of plant-pathogenic

grupo de especies no fitopatógenas fue más sencillo, dado el desempeño de los bosques al azar del 85.9%. Por otra parte se demostró que existe una relación estadística entre los perfiles de ácidos grasos y la patogénesis sobre la planta.

Palabras clave adicionales: Diagnóstico, bacterias no patogénicas.

El género *Pseudomonas* está clasificado en el dominio *Bacteria*, phylum *Proteobacteria*, clase *Gammaproteobacteria*, orden *Pseudomonadales* y familia Pseudomonaceae (Brenner *et al*., 2005). La especie tipo del género es *Pseudomonas aeruginosa*, la cual fue originalmente descubierta por Schroeter en 1872. El género, sin embargo fue propuesto por Migula en 1894. En marzo de 2008, 95 especies de *Pseudomonas* fueron válidamente publicadas. Los miembros del género *Pseudomonas* son bastones típicamente rectos o ligeramente curvados con flagelos polares, presentan un mecanismo aeróbico (respiratorio), pero ninguna especie es fermentativa. La mayoría de especies no crecen en condiciones ácidas (pH menor de 4.5) y son habitantes naturales del agua o suelo. Se han establecido diferentes subgrupos dentro de este género basado en las características fenotípicas o patogénicas. Anteriormente, el agrupamiento pudo estar basado en la producción de pigmentos fluorescentes bajo radiación UV (*e.g. P. fluorescens*) posteriormente se formaron grupos de especies patogénicas y no patogénicas. Dos ejemplos son *P. aeruginosa*, un patógeno oportunista de humanos, mientras que *P. syringae* es un patógeno de plantas (Palleroni, 2005; 2008). Los pioneros en la clasificación taxonómica del género *Pseudomonas* fueron Palleroni y colaboradores (Universidad de California, Berkeley, USA) quien en 1963 describió cinco grupos de *Pseudomonas* basados en la hibridación rRNA-DNA. Desde entonces, la taxonomía del género *Pseudomonas* ha sufrido una serie de re-arreglos y hoy en día solamente el grupo I de rRNA corresponde al género *Pseudomonas*. Esto implica que un gran número de especies previamente asignadas al género *Pseudomonas sensu lato* (a menudo referida como las "*Pseudomonas*") fueron transferidos a rangos genéricos o supra genéricos, principalmente pertenecientes a las clases: *Alfa*, *Beta* y *Gammaproteobacteria*. Ejemplos: *Acidovorax*, *Aminobacter*, *Brevundimonas*, *Burkholderia*, *Comamonas*, *Halomonas*, *Methylobacterium*, *Ralstonia*, *Sphingomonas*, *Xanthomonas*, etc., citado por Kersters *et al.*, (1996) y Palleroni (2005). En 1996, Moore *et al.*, diferenciaron 2 grupos intergenéricos basados en la secuencia del gen 16S rRNA del grupo de *Pseudomonas sensu stricto* (= el presente género *Pseudomonas*): el grupo denominado *P. aeruginosa* y el otro grupo *P. fluorescens* cada uno con especies provenientes de diferentes linajes. La mayoría de linajes fueron también agrupados con el análisis FAME por Vancanneyt *et al.*, (1996). En el año 2000, Anzai *et al.*, re-evaluaron 128 especies válidas y no válidas de *Pseudomonas* basados en datos de la secuencia 16S rRNA y varias especies fueron reasignadas a otros géneros. La complejidad de la taxonomía del presente género *Pseudomonas* está también demostrada por los análisis del gen *rpo*B realizado por Tayeb

plant-pathogenic species from the group of non-plant-pathogenic species was more straightforward, given by the Random Forests identification performance of 85.9%. Moreover, it was shown that a statistical relation exists between the fatty acid profiles and plant pathogenesis.

Additional key words: Diagnosis, non-pathogenic bacteria.

The genus *Pseudomonas* is classified in the domain of *Bacteria*, phylum of *Proteobacteria* class of *Gammaproteobacteria*, order of *Pseudomonadales* and family of Pseudomonaceae (Brenner *et al*., 2005). The type species of the genus is *Pseudomonas aeruginosa*, which was originally discovered by Schroeter in 1872. The genus was, however, proposed by Migula in 1894. On 03/2008, 95 *Pseudomonas* species were validly published. *Pseudomonas* members are typically straight or slightly curved motile rods with merely polar flagella, respiratory but never fermentative. Most species fail to grow under acid conditions (pH lower than 4.5) and natural habitats are water or soil. Different subgroupings can be made based on the phenotype or pathogenetic features. In the former case, grouping can be based on the production of pigments that fluorescence under UV radiation (*e.g. P. fluorescens*). In the latter case, a straight forward grouping of pathogenic and non-pathogenic species can be made. Two examples are *P. aeruginosa* which is an opportunistic pathogen of humans, while *P. syringae* is a plant pathogen (Palleroni, 2005; 2008). Pioneers in the taxonomic classificlassification of the genus *Pseudomonas* are Palleroni and coworkers (University of California, Berkeley, USA), who in 1973 described an initial grouping of five discrete *Pseudomonas* clusters based on rRNA-DNA hybridization (Palleroni, 1973; 1984). Since then, the taxonomy of the genus *Pseudomonas* underwent a series of rearrangements and the genus as described today corresponds to rRNA group I. This implies that numerous species, previously assigned to the genus *Pseudomonas sensu lato* (often referred to as the "Pseudmonads") were transferred to the generic or suprageneric ranks, mainly residing in the Alpha-Beta, and Gammaproteobacteria classes. Examples are *Acidivorax*, *Aminobacter*, *Brevundimonas*, *Burkholderia*, *Comamonas*, *Halomonas*, *Methylobacterium*, *Ralstonia*, *Sphingomonas*, *Xanthomonas*, etc. (Kersters *et al*., 1996; Palleroni, 2005). In 1996, Moore *et al*, discriminated two intrageneric clusters in 16S rRNA gene sequences of the *Pseudomonas sensu stricto* group (= the present genus *Pseudomonas*): a *P. aeruginosa* cluster and a *P. fluorescens* cluster, each with different species lineages. Most lineages were also clustered in the FAME analysis as performed by Vancanneyt *et al*., (1996). Anzai *et al*., (2000) re-evaluated 128 valid and invalid *Pseudomonas* species based on 16S rRNA sequence data and reassigned several species to other genera. The complexity of the taxonomy of the present genus *Pseudomonas* is also demonstrated by the *rpo*B gene analysis of Tayeb *et al*., (2005), but validation of the *rpo*B grouping still need DNA-DNA hybridization (DDH) data and extensive phenotypic analysis before emendations at the species level can be proposed.

et al., (2005), pero la validación del agrupamiento del *rpo*B aún necesita datos de la hibridación DNA-DNA (DDH) y de la realización de extensos análisis fenotípicos antes de que se puedan proponer cambios a nivel de especie. Un árbol filogenético construido con el método de máxima verosimilitud del género *Pseudomonas* basado en la secuencia del gen 16S rRNA se puede observar en la Figura 1. Este árbol incluye las especies bacterianas válidamente publicadas en marzo del 2008. Un gran número de patovares

A maximum likelihood tree of the genus *Pseudomonas* based on 16S rRNA gene sequences is visualized in. This tree is based on the validly published list of bacterial species as of Figure 1March 2008. Regarding plant pathogenesis, a multitude of pathovars are described within the species P. syringae and related species. A DDH study showed the existence of nine discrete genomospecies (Gardan *et al*., 1999). In this study, we follow these genomospecies classifications as if they were formal species.

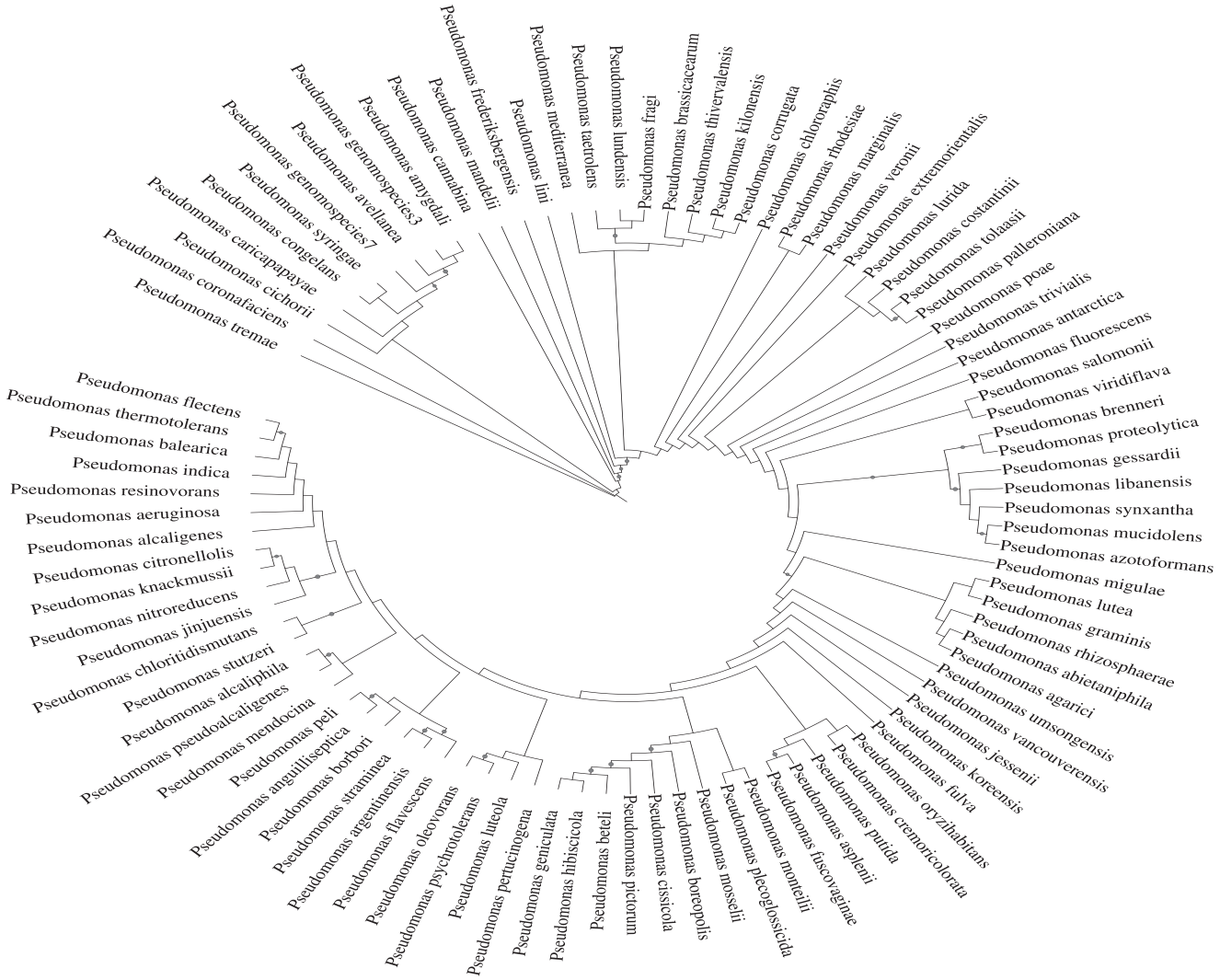

Figura 1. Árbol filogenético de máxima similitud de genéro *Pseudomonas* basado en la secuencia del gen 16S rRNA . Las 95 especies corresponden a la taxonomía valida publicada a marzo del 2008. Cada especie se representa por una secuencia de alta calidad del 16S rRNA extraída de la base de datos SILVA (Pruesse *et al.*, 2007). El árbol filogenetico se elaboró mediante el algoritmo de máxima similitud en el software RAxML y es visualizado con la heramienta de la web iTol (Letunic and Bork, 2007; Stamatakis, 2006). Para especies aisladas, la longitud de las ramas es ignorada debido a que el uso de la longitud de la rama es innecesario. Ramas punteadas corresponden a valores de re-muestreo del 75%.

Figure 1.16S rRNA gene sequence-based maximum likelihood tree of the genus *Pseudomonas*. The 95 species included correspond to the validly published taxonomy of March 2008. Each species is represented by one high-quality 16S rRNA sequence as extracted from the SILVA database (Pruesse *et al*., 2007). The phylogenetic tree is built by the maximum likelihood algorithm as implemented in the RAxML software (based on 1000 bootstraps) and is visualized with the iTol webtool (Letunic and Bork, 2007; Stamatakis, 2006). Tree branch lengths are ignored because of outlier species, making the use of branch lengths pointless. Dotted branches correspond to bootstrap values larger than 75%.

relacionados a patogenicidad en plantas, se han descrito dentro de la especie *P. syringae* y especies relacionadas. Un estudio de DDH mostró la existencia de nueve discretas genomoespecies (Gardan *et al.*, 1999). En este estudio, nosotros seguimos esta clasificación de genomoespecies como si ellos fueran especies válidamente descritas. El género *Pseudomonas* pertenece a las bacterias Gram-negativas. Esto implica que las investigaciones iniciales en la composición de ácidos grasos en este género estuvieron enfocadas en la capa de lipopolisacáridos correspondiente (LPS). Esta capa es responsable de una importante fracción discriminatoria de hidroxi-ácidos grasos. Varios investigadores mostraron inicialmente que la mayor fracción de ácidos grasos de la capa de LPS de *Pseudomonas aeruginosa* estuvo constituida por hidroxi ácidos (Fensom y Grey, 1969; Hancock *et al*., 1970). En los años 1970, la principal investigación en el contenido de ácidos grasos de los miembros del género *Pseudomonas sensu lato* fue llevada a cabo en los laboratorios de Moss (Centro de Control de enfermedades, Atlanta, USA) (Dees and Moss, 1975; Dees *et al*., 1979; Moss *et al.*, 1972; Moss, 1981). Ellos encontraron que los patrones de Esteres matílicos de ácidos grasos (FAME) fueron usados rápidamente para diferenciar especies de *Pseudomonas* y grupos de especies, debido a que los análisis repetidos de FAME dieron patrones similares. Junto con estas investigaciones sobre el contenido de ácidos grasos bacteriales, se llevaron a cabo investigaciones para el mejoramiento del método analítico del contenido de GC. Por supuesto, otros microbiólogos llevaron a cabo investigaciones sobre el contenido celular de ácidos grasos de las *Pseudomonas*. Ikemoto *et al.*, (1978) concluyeron que la presencia de ácidos hidroxi, ciclopropano ácidos y ca ácidos de cadenas ramificadas fue característico para los grupos y especies de este género. Interesantemente, estos autores al igual que los primeros investigadores usaron la longitud de la cadena equivalente (ECL) para los picos de detección de FAME. Los valores de ECL se determinaron del logaritmo del tiempo de retención de la cadena recta saturada FAME contra el número de carbonos. Uno de los principales estudios sobre los ácidos 3-hidroxi fue reportado por Oyaizu (1983). A inicios de los 90', las primeras evaluaciones de los sistemas de evaluación comercial para la identificación bacterial basada en FAME, Sherlock MIS (MIDI Inc. Newark, DE, USA), se llevaron a cabo también con especies del género *Pseudomonas* (Hosterhout *et al*., 1991; Stead, 1992; Stead *et al*., 1992). Aquí, el grupo de Stead se enfocó en las especies de *Pseudomonas* fitopatógenas. Este artículo posteriormente puede estar relacionado con una de las primeras revisiones importantes que se hicieron con FAME donde se analizaron 38, de las 86 especies de *Pseudomonas* válidamente descritas. Seis grupos de cepas fueron discriminadas basadas principalmente en tres tipos de ácidos grasos hidroxi (también core hidroxi ácidos grasos: 2-hidroxi, 3-hidroxi e iso-ramificado 3-hidroxi), a pesar de que ellos tuvieron menos del 10% del área total de picos. Diferencias cuantitativas en ácidos grasos no hidroxi permitieron diferenciar entre taxa dentro de estos grupos y se encontraron pocas diferencias cualitativas entre los perfiles de taxa incluidas en el mismo subgrupo. Incluso únicos perfiles se encontraron para taxa

The genus *Pseudomonas* belongs to the Gram-negative bacteria. This implies that initial research on the fatty acid in this genus was focused on the corresponding lipopolysaccharide (LPS) layer. This layer is responsible for an important discriminatory fraction of hydroxy fatty acids. Several researchers showed initially that the major fatty acid fraction of the LPS layer of *Pseudomonas aeruginosa* was constituded of hydroxy acids (Fensom and Gray, 1969; Hancock *et al*., 1970). In the 1970s, main research on the fatty acid content of the members of the genus *Pseudomonas sensu lato* was performed at the laboratory of Moss (Center for Disease Control, Atlanta, USA) (Dees and Moss, 1975; Dees *et al*., 1979; Moss *et al*., 1972; Moss, 1981). They found that FAME patterns were useful for rapidly distinguishing between *Pseudomonas* species and species groups, and that repeated FAME analysis resulted in similar patterns. Next to research on the bacterial fatty acid content itself, research was performed onimprovements of the analytic GC method. Of course, other microbiologists performed research on the cellular fatty acid content of the Pseudomonads. Ikemoto *et al*., (1978) concluded that the presence of hydroxy acids, cyclopropane acids and branched-chain acids was characteristic for the groups and species in the genus. Interestingly, probably as one of the first these authors used the equivalent chain length (ECL) for FAME peak detection. The ECL value was determined from the logarithm of the retention time of saturated straight-chain FAMEs plotted against their carbon number. A major study of the 3-hydroxy fatty acids was reported by Oyaizu (1983). In the early 1990s, the first evaluations of the commercial system for FAME-based bacterial identification, Sherlock MIS (MIDI Inc. Newark, DE, USA), were also performed with *Pseudomonas* species (Hosterhout *et al*., 1991; Stead, 1992; Stead *et al*., 1992). Herein, the group of Stead focused on the plant-pathogenic *Pseudomonas* species. The latter paper can be regarded as one of the first important review papers as 38 of the, at that time, 86 validly described *Pseudomonas* species were analyzed. Six groups of strains were discriminated, mainly based on three types of hydroxy fatty acids (also core hydroxy fatty acids: 2-hydroxy, 3-hydroxy and iso-branched 3-hydroxy), even though they count for less than 10% of the total peak area. Quantitative differences in non-hydroxy fatty acids allowed differentiating between taxa within those groups and few qualitative differences were found between the profiles of taxa included in the same subgroups. Even unique profiles were found for infraspecific taxa (subspecies, biovar, pathovar) (Stead *et al*., 1992). Also found good correlation between the fatty acid grouping and grouping based on the results of DNA-DNA and DNA-rRNA hybridization. Four years later, Vancanneyt *et al*., (1996) performed a taxonomic evaluation of the Pseudomonads. In this study, 30 *Pseudomonas* species were included. Again the presence of hydroxy fatty acids was shown to be a good taxonomic marker for delineating species and a good correlation was found between the major groups resulting from whole-cell FAME analysis and the groupings based on DNA-rRNA hybridization. However, Vancanneyt *et al*., (1996) also concluded that, from the mean species fatty acid

infraespecíficas (subespecies, biovares, patovares) (Stead *et al*., 1992). Importantemente, Stead *et al.*, (1992) también encontraron buenas correlaciones entre los agrupamientosrealizados con datos de ácidos grasos y los agrupamientos basados en los resultados de hibridación DNA-DNA y DNA-rRNA. Cuatro años después Vancanneyt *et al.*, (1996) realizaron una evaluación taxonómica de las *Pseudomonas*, en este estudio se incluyeron 30 especies de *Pseudomonas*. Nuevamente la presencia de ácidos grasos hidroxi mostró ser un buen marcador para la delimitación de especies y una buena correlación se encontró entre los principales grupos obtenidos del análisis FAME de toda la célula y los grupos basados en hibridación DNA-rRNA. Sin embargo, Vancanneyt *et al.*, (1996) también concluyeron que, el contenido de ácidos grasos de las principales especies, no permite discriminar especies dentro de los diferentes grupos, de estos dos últimos estudios, también se concluyó que los ácidos grasos predominantes en el género *Pseudomonas* son $C_{16:0}$, $C_{18:1}$ y derivados (Stead *et al*., 1992; Vancanneyt *et al*., 1996). Sin embargo se considera que los ácidos grasos hidroxi, $C_{10:0}$ 3-hidroxi, $C_{12:0}$ 3-hidroxi y $C_{12:0}$ 2-hidroxi son predominantes (Palleroni, 2005). Para información mas detallada relacionada al contenido FAME de las diferentes especies de *Pseudomonas*, nosotros nos basamos en las descripciones de las especies y en el Manual Bergey de Bacteriología sietemática (Palleroni, 2005). Con la disponibilidad de un gran número de datos FAME para el género *Pseudomonas* y de un gran número de especies fitopatógenos de este género, nosotros hemos direccionado nuestra investigación al uso de métodos de aprendizaje inteligente para una identificación global o de las especies a nivel de género con datos FAME (Slabbink *et al*., 2008; Slabbink *et al*., 2009) a través de la relación entre la composición bacterial de FAME y la patogénesis en plantas. Sin embargo, en este trabajo, nosotros analizamos la habilidad del análisis FAME para diferenciar entre especies fitopatógenas de *Pseudomonas* y para distinguir estas especies del grupo de especies de *Pseudomonas* no patogénicas. Para entender e interpretar la presente investigación y los resultados correctamente, nosotros primero introducimos nuestros datos, el análisis de componentes principales, el uso de técnicas de aprendizaje inteligente y el subsecuente análisis estadístico. Finalmente, presentamos los resultados obtenidos en esta investigación de FAME en especies de *Pseudomonas* fitopatógenas.

MATERIALES Y MÉTODOS.

Análisis FAME y grupo de datos. Se consideraron las bases de datos FAME del Laboratorio de Microbiología (Ghent University, Belgium) y el BCCM ™/Colección de Bacterias LMG (Bélgica). El análisis FAME de todas las células bacterianas se ha llevado a cabo en este laboratorio desde 1989 y ha resultado en una base de datos que actualmente contiene más de 71,000 perfiles FAME. La investigación basada en FAME está dirigida a bacterias con impacto ambiental, clínico e industrial, pero también se lleva a cabo con propósitos taxonómicos. Básicamente, los perfiles FAME obtenidos por cromatografía de gases son generados

content, species discrimination was not possible within the different groups. From the last two studies, it could also be concluded that predominant fatty acids in the genus*Pseudomonas* are $C_{16:0}$, $C_{18:1}$ and derivatives (Stead *et al*., 1992; Vancanneyt *et al*., 1996). Regarding hydroxy fatty acids, $C_{10:0}$ 3-hydroxy $C_{12:0}$3-hydroxy and $C_{12:0}$2-hydroxy are predominant (Palleroni, 2005). For more detailed information regarding the FAME content of the different *Pseudomonas* species, we refer to the corresponding species descriptions and to Bergey's Manual for Systematic Bacteriology (Palleroni, 2005). With the availability of a large FAME data set for the genus *Pseudomonas* and the presence of a large number of plant pathogens in the genus, we have shifted our research on intelligent learning methods for a global or genus-wide bacterial species identification by FAME data (Slabbink *et al*., 2008; Slabbink *et al*., 2009) towards the relation between bacterial FAME constitution and plant pathogenesis.Therefore, in this work, we analyze the ability of FAME analysis to distinguish between plant-pathogenic *Pseudomonas* species and to distinguish these species from the group of non-plant-pathogenic *Pseudomonas* species. To understand and interpret the presented research and according results pr operly, we first introduce our data, principal component analysis, the used machine learning techniques and the subsequent statistical analysis. Finally, we present the results obtained in this FAME research on plant-pathogenic *Pseudomonas* species.

MATERIALS AND METHODS.

FAME analysis and data sets. The in-house FAME database of the Laboratory of Microbiology (Ghent University, Belgium) and the BCCM™/LMG Bacteria Collection (Belgium) is considered. Whole-cell bacterial FAME analysis has been performed in this laboratory since 1989 and has resulted in a database that currently contains more than 71,000 FAME profiles. FAME research is directed towards bacteria with environmental, clinical and industrial impact, but is also performed for taxonomic purposes. Basically, gas chromatographic FAME profiles are generated following bacterial growth, as described in the TSBA protocol of the Sherlock Microbial Identification System of MIDI Inc. (Newark, DE, USA). This protocol defines a standard for growth, cultur ing and analysis of bacterial strains, and reproducibility and interpretability of the profiles is only possible by working under the described conditions. In this work, we focus on the species of the genus *Pseudomonas* which were grown according to the TSBA protocol. Specifically, this protocol recommends 24h of growth on trypticase soy broth agar at a temperature of 28°C. Following growth, on average 40mg bacterial cells are harvested in the overlapping region of quadrants three and four of streaked plates. Next, FAMEs are extracted by a four-step procedure of saponification, methylation, extraction and sample cleanup, and are finally analyzed by gas chromatography. Following gas chromatographic analysis, the use of a calibration mix and a naming table, the resulting FAME profiles are standardized by calculating relative peak areas for each named peak with respect to the total named peak area. In this work, the TSBA50 peak naming table is used. Whole-cell bacterial FAME profiles resulting from gas chromatographic analysis following these standard growth

durante el crecimiento bacterial, como se describe en el protocolo TSBA del Sistema de Identificación Microbial Sherlock de MIDI Inc. (Sherlock Microbial Identification System de MIDI Inc.) (Newarl, DE, USA). Este protocolo define un estándar para el crecimiento, cultivo y análisis de cepas bacterianas y la reproducibilidad e interpretabilidad de los perfiles es solamente posible trabajando bajo las condiciones descritas. En este trabajo nosotros nos enfocamos en las especies del género *Pseudomonas* las cuales crecieron de acuerdo al protocolo TSBA. Específicamente, este protocolo recomienda el crecimiento de las bacterias durante 24 horas en Tripticasa-caldo de soya-agar a una temperatura de 28°C. Después del crecimiento, se colecta un promedio de 40 mg de células bacterianas de la región

conditions are further indicated as standard FAME profiles.
The *Pseudomonas* FAME data set of Slabbinck *et al*., 2009 is considered, which relies on the bacterial taxonomy as of March 2008. This data set covers 1673 standard FAME profiles of 95 validly published *Pseudomonas* species and 94 named FAME peaks.
In this work, this data set is evaluated in terms of plant pathogenesis. Herein, a species is considered as plant-pathogenic when one of its strains relates to plant or mushroom pathogenesis. Related to our *Pseudomonas* FAME data set, several plant-pathogenic *Pseudomonas* species are considered. These species are listed Table 1 species are considered. These species are listed Table 1, together with the number of corresponding standard FAME

Cuadro 1. Información general de las especies consideradas fitopatógenas. Para cada especie se dan, el número estándar de perfil FAME, el hospedero (s) y las referencias correspondientes.
Table 1. Overview of the considered plant-pathogenic species. For each species, the number of standard FAME profiles, the host(s) and corresponding reference(s) are given.

Especie	No. FAME de perfil	Hospedero(s)	Referencia(s)
P. Agarici	8	*Agaricus bisporus*	Höfte y De Vos, 2007.
P. amygdali	111	*Prunus amygdalus*	Höfte y De Vos, 2007.
P. asplenii	13	*Asplenium nidus*	Gardan, *et al*., 2002.
P. avellanae	4	*Corylus avellana*	Janse *et al*., 1996.
P. beteli	6	*Piper betle*	Van Den Mooter y Swings, 1996
P. cannabina	7	*Cannabis sativa*	Gardan *et al*., 1999.
P. caricapapayae	5	*Carica papaya*	Höfte y De Vos, 2007.
P. cichorii	32	Rango amplio de hospederos	Smith *et al*., 1988; Höfte y De Vos, 2007.
P. cissicola	8	*Cissus japonica*	Hu *et al*, 1997.
P. coronafaciens	44	*Avena sativa*	Gardan, *et al*., 1999.
P. corrugata	22	*Lycopersicon, Chrysanthemum, Geranium, Medicago, , pepper*	Catara *et al*., 2002; Gardan *et al*., 2002;Höfte y De Vos, 2007.
P. costantinii	6	*Agaricus bisporus*	Höfte y De Vos, 2007; Munsch *et al*., 2002.
P. flavescens	7	*Juglans regia*	Hildebrand *et al*., 1994.
P. flectens	6	*Phaseolus vulgaris*	Gardan *et al*., 2002.
P. fuscovaginae	45	*Oryza sativa, Allium, Secalotriticum,Triticum*	Miyajima *et al*., 1983.
P. hibiscicola	6	*Hibiscus rosa-sinensis*	Vancanneyt *et al*., 1996.
P. marginalis	63	*Pastinaca sativa, Allium, Cichorium, Medicago, Phaseolus, Phragmipedium*	Gardan *et al*., 2002;Höfte y De Vos, 2007.
P. mediterranea	10	*Lycopersicon*	Catara *et al*., 2002; Höfte y De Vos, 2007.
P. salomonii	10	*Allium sativum*	Gardan *et al*., 2002.
P. syringae	88	Amplio rango de hospederos	Gardan *et al*., 2002.
P. syringae genomespecie 3	53	Amplio rango de hospederos	Gardan, *et al*., 1999.
P. syringae genomespecie 7	8	*Helianthus annuus, Tagetes erecta*	Gardan, *et al*., 1999.
P. tolaasii	53	*Agaricus bisporus, Agaricus, bitorquis, Allium sativum, Pleurotus ostreatus, Pleurotus eryngii*	Höfte y De Vos, 2007;Munsch *et al*., 2002.
P. tremae	8	*Trema orientalis*	Gardan, *et al*., 1999 ; Gardan *et al*., 2002.
P. viridiflava	17	Amplio rango de hospederos	Höfte y De Vos, 2007.

sobrepuesta del tercer y cuarto cuadrante de la placa estriada. Posteriormente, los FAME se extraen mediante un procedimiento de cuatro pasos: saponificación, metilación, extracción y limpieza de la muestra, finalmente son analizadas por cromatografía de gases. Seguido al análisis por cromatografía de gases, con la ayuda de una calibración y una lista de nombres, los perfiles FAME resultantes son estandarizados mediante el cálculo de las áreas relativas de los picos para cada pico identificado con respecto al área total de los picos. En este trabajo, se usó el cuadro de picos identificados en TSBA. El perfil FAME resultante de toda la célula bacteriana analizado por cromatografía de gases en condiciones de crecimiento estándar es considerado posteriormente como perfiles estándares FAME. En este trabajo se consideraron los datos de las *Pseudomonas* obtenidos con el análisis FAME por Slabbinck *et al.*, 2009 debido a que se encuentra en concordancia con la taxonomía bacterial de Marzo de 2008. Este grupo de datos cubre 1673 perfiles estándares de FAME de 95 especies de *Pseudomonas* válidamente publicadas en los que se ha identificado 94 picos FAME. En la presente investigación, este grupo de datos es evaluado en términos de patogénesis en plantas. Aquí, una especie es considerada como fitopatógena cuando una cepa está relacionada a plantas o patogénicas en hongos o setas. En relación a nuestro grupo de datos FAME de *Pseudomonas* se consideran varias especies de *Pseudomonas* fitopatógenas. Estas especies están listadas en el Cuadro 1, junto con el número de perfiles estándar de FAME correspondientes, su hospedante (s) relacionado y la referencia (s). En relación a los patovares de *Pseudomonas syringae*, la clasificación de genomoespecies se realizó de acuerdo a lo sugerido por Gardan *et al*., (1999).

Aquí, la genomoespecie 3 corresponde a los patovares de *P. syringae*: *antirrhini*, *apii*, *berberidis*, *delphinii*, *maculicola*, *passiflorae*, *persicae*, *primulae*, *ribicola*, *tomato* y *viburni*. La genomoespecie 7 corresponde a los patovares de *P. syringae*: *helianthi* y *tagetis*. Remarcando también que algunas especies de *Pseudomonas* son generalmente mal nominadas y están transferidas a otros géneros, tales como *P. beteli* (*Stenotrophomonas maltophilia*) y *P. hibiscola* (*Stenotrophomonas maltophilia*) (Anzai *et al*., 2000). La especie *P. cissicola* debería ser transferida al género *Xanthomonas*, pero se requiriere de un análisis más extensivo de las características genotípicas y fenotípicas de las especies (Anzai *et al*., 2000; Hu *et al*., 2007; Palleroni, 2005).

Una anomalía taxonómica similar se observó en *P. flectens*, la cual basada en el análisis de la secuencia del gen 16S rRNA se agrupó en la familia Enterobacteriaceae (Anzai *et al*., 2000; Palleroni, 2005). *P. beteli*, *P. cissicola* y *P. hibiscicola* fueron posteriormente anotadas en el grupo *P. beteli*. Cuando se enfocan en especies de *Pseudomonas* fitopatógenas, se evalúan dos casos interesantes de pruebas relacionadas con la identificación bacteriana basada en datos FAME: la discriminación de las especies fitopatógenas están listadas en el Cuadro 1 y la discriminación de especies fitopatógenas y no fitopatógenas basadas en datos FAME de las *Pseudomonas*. En el primer caso, un nuevo grupo de datos se extrajo del grupo de datos completos de FAME cubriendo 25 especies profiles, related host(s) and reference(s). Regarding *Pseudomonas syringae* pathovars, the genomospecies classification is followed as suggested by Gardan *et al*., 1999. Herein, genomospecies 3 corresponds to the *P. syringae* pathovars *antirrhini, apii, berberidis, delphinii, maculicola, passiflorae, persicae, primulae, ribicola, tomato*and *viburni.* Genomospecies 7 corresponds to the *P. syringae* pathovars *helianthi* and *tagetis*. Remark also that some *Pseudomonas* species are generically misnamed and are transferred to another genus, such as *P. beteli,* (*Stenotrophomonas maltophilia*) and *P. hibiscicola* (Stenotrophomonas maltophilia) (Anzai *et al*., 2000). *P. cissicola* should be transferred to the genus *Xanthomonas* but a more extensive analysis of the genomic and phenotypic characteristics of the species is required (Anzai *et al*., 2000; Hu *et al*., 2007; Palleroni, 2005). A similar taxonomic anomaly for *P. flectens* that, based on 16S rRNA gene sequence analysis, clusters in the family of the *Enterobacteriaceae* (Anzai *et al*., 2000; Palleroni, 2005). *P. beteli, P. cissicola* and *P. hibiscicola* are further denoted as *P. beteli* group. When focusing on plant-pathogenic *Pseudomonas* species, two interesting test cases are evaluated regarding bacterial identification based on FAME data: the discrimination of the plant-pathogenic *Pseudomona*s species as listed in Table 1 and the discrimination of plant-pathogenic *Pseudomonas* species from the non-plantpathogenic Pseudomonas species present in the *Pseudomonas* FAME data set. In the first case, a new data set is extracted from the complete *Pseudomona* FAME data set, covering 25 plant-pathogenic species. In the second case, the FAME profiles of the complete *Pseudomonas* data set are re-labeled in accordance with the property of infecting plants or mushrooms. Thus, in this case, only two labels or 'classes' are considered: plant pathogen or non-plant pathogen. Remark that the non-plant pathogen subset contains70 *Pseudomonas* species.

Principal component analysis. When dealing with datasets with an excessive dimensionality (*i.e*. in this work the number of FAME peaks), one approach to reduce the dimensionality is to combine the different features and, as such, project the high-dimensional data in a lower dimensional space. Principal component analysis (PCA) is such a popular dimensionality reduction method by which linear combinations are composed from the different FAME peaks (also called features) (Duda *et al.*, 2001). Initially, the linear combination that represents the largest amount of variability in the data is chosen and called the first principal component (PC). Next, subsequent linear combinations are composed that are orthogonal to the previous PCs, repeatedly based on the combination representing the highest variance. A simple visualization of the PCA is achieved by plotting the variance and accumulated variance of the PCs as ranked by amount of represented variance in a so-called skree plot (see also Figures 2 and 4). From this plot, it is easy to determine how many PCs are needed to cover a certain percentage of variability in the data. These Pcs can subsequently be used for learning with a smaller number of features, resulting in a model of lower dimensions, or thus a less complex model. This is not only advantageous for computational reasons but may also be very

fitopatógenas. En el segundo caso, los perfiles de FAME del grupo completo de *Pseudomonas* son re-calificadas de acuerdo con las propiedades de infectar plantas u hongos. Así, en este caso, solamente dos etiquetas o clases son considerados: fitopatógenos o no fitopatógenos. Obsérvese que el grupo de las no-fitopatógenas contiene 70 especies de *Pseudomonas*.

Análisis de componentes principales. Cuando se trata con grupo de datos de una dimensionalidad excesiva (*i.e.* en este trabajo el número de picos FAME), el enfoque para reducir esta dimensionalidad es combinar las diferentes características y de esta manera, proyectar la gran dimensionalidad de los datos en un espacio dimensional reducido. El análisis de componentes principales (ACP) es un método de reducción de dimensionalidad, en el cual las combinaciones lineares están compuestas de los diferentes picos FAME (también llamado características) (Duda *et al.*, 2001). Inicialmente, la combinación linear que representa la cantidad de variabilidad en los datos es seleccionada y llamado el principal componente principal (PC). Después, las subsecuentes combinaciones lineares son ortogonales al PC anterior, y así repetidamente, estas combinaciones van a representar la varianza más alta. Una visualización simple de los PCA se lleva a cabo dibujando la varianza y la varianza acumulada de los PC, la cantidad de varianza está representada en la gráfica de sedimentación (ver también Figuras 2 y 4). De esta gráfica, es fácil determinar cuántos PC son necesarios para cubrir un cierto porcentaje de la variabilidad en los datos. Estos PC pueden subsecuentemente ser usados para aprender con un número más pequeño de características, resultando en un modelo de bajas dimensiones, o al menos en un modelo menos complejo. Esto no es solamente una ventaja por razones computacionales, sino también puede ser efectivo para incrementar el comportamiento de un modelo de identificación. Graficando los datos en dos o tres espacios dimensionales de componentes principales permite una fácil interpretación de las relaciones (similitud o distancias) entre las diferentes clases, i.e. especies o grupos de especies.

Aprendizaje automatizado.

Terminología. Clasificación e identificación son términos usados con diferentes significados en los campos de la microbiología y en el aprendizaje automatizado. Mientras la identificación se interpreta como la asignación de las clases existentes de un organismo desconocido o datos puntuales, la clasificación debería ser interpretada de manera diferente. En un contexto taxonómico, la clasificación bacterial se refiere a la agrupación de organismos bacterianos basados en similitudes genotípicas y fenotípicas (Madigan *et al*., 2009). Sin embargo, en el aprendizaje automatizado, el objetivo de la clasificación es describir estadísticamente la relación entre los datos señalados de varias clases, considerando las clases establecidas, por la generalización de las tendencias observadas en los datos (Mitchell, 1997). En el contexto de este trabajo, nos referimos a clasificación como el proceso de construir modelos computacionales para diferenciar entre los perfiles FAME de los diferentes grupos de especies bacterias o grupos de especies (Slabbinck *et al*., 2009).

effective for increasing the performance of an identification model. Plotting the data in a two or three dimensional principal component space allows for an easy interpretation of the relations (similarities or distances) between the different classes, *i.e*. species or species groups.

Machine learning.

Terminology. Classification and identification are terms used with different meanings in the fields of microbiology and machine learning. Whereas identification is similarly interpreted as assigning existing class labels to unknown bacterial organisms or data points, classification should be interpreted differently. In a taxonomic context, bacterial classification refers to the grouping of bacterial organisms based on genotypic and phenotypic similarities (Madigan *et al*., 2009). How ever, in machine learning, the goal of classification is to statistically describe the relationship between data points of various classe, given the class labels, by generalizing the observed trends in the data (Mitchell, 1997). In the context of this work, we refer to classification as the process of building computational models to distinguish between the FAME profiles of the different bacterial species or species groups (Slabbinck *et al*., 2009).

Random Forests. Random Forests (RFs) is an ensemble method based on bagging. Ensemble methods generate multiple classifiers (or classification models) and aggregate the results. Specifically, a RF is an ensemble of classification trees. At each node of a classification tree, the best split is chosen among a subset of features randomly chosen at that node. In contrast, standard classification trees split each node based on all features available. RFs performs very well compared to many other classifiers and is robust against overfitting (Breiman, 2001). To achieve optimal classification of the data, two parameters require an optimization step: the number of randomly chosen features at each node (Nf) and the number of trees to be grown in the forest (Nt). Optimization of Nf is done by varying the number of features from 1 to the total number of features in steps of 5 and optimization of Nt is done by varying the number of trees from 1000 to 4000 in steps of 250. Optimization of the parameters is done by a grid-search and the combination of parameters leading the smallest validation error is finally chosen. The first step in the RF algorithm is the generation of Nt bootstrap samples from the original data set. These samples contain about two-thirds of the FAME profiles of the original data set. The remaining profiles are used as test set. For each bootstrap sample, an unpruned tree is grown by the method previously described. Next, for each sample, the corresponding test profiles are predicted by the corresponding tree. At the end of the run, and over all samples, we assign the class label of the class which got most of the votes every time a given profile was present in a test set. The proportion of misclassifications averaged over all test profiles is taken as the overall error rate. New data can be predicted by aggregating the predictions of the Nt trees and selecting the label with the highest number of votes (Breimen, 2001). The RFs used in this study are generated by the RandomForest software package.

Validation. For parameter optimization and model

Árboles aleatorios. Random Forest (RFs) es un método de agrupación basado en datos contenidos idealmente en una bolsa. Los métodos de agrupación generan múltiples clasificadores (o modelos de clasificación) y los resultados se acumulan en forma de agregados. Específicamente, un RF es una agrupación de árboles para clasificación. En cada nodo de un árbol de clasificación, se elige la mejor división entre un subconjunto de características elegidas al azar en ese nodo. En contraste, los árboles de clasificación estándar dividen a cada nodo en base a todas las características disponibles. Los RF funcionan muy bien comparados con muchos otros clasificadores y es robusto frente a re-muestreos (Breiman, 2001). Para lograr una clasificación óptima de los datos, se requiere un paso de optimización en dos parámetros: el número de características elegidas al azar en cada nodo (Nf) y el número de árboles que se produzcan en el análisis (Nt). La optimización de Nf se realiza variando el número de características de 1 al número total de características en pasos de 5 y la optimización de Nt se realiza variando el número de árboles de 1000 a 4000 en pasos de 250. La optimización de los parámetros se realiza con un filtro de búsqueda y finalmente se elige la combinación de los parámetros que llevan a los más pequeños errores de validación. El primer paso en el algoritmo RF es la generación de muestras Nt bootstrap (muestreo con reemplazo) del conjunto de datos originales. Estas muestras contienen cerca de dos tercios de los perfiles FAME del grupo de datos originales. Los perfiles restantes son usados como un grupo de prueba. Para cada muestra bootstrap, se produce un árbol sin ninguna modificación por el método previamente descrito. Después, para cada muestra, los perfiles de prueba correspondientes son predecibles por el árbol correspondiente. Al final de la corrida de todas las muestras, se asigna un valor a la clase marcada que obtuvo la mayoría de las combinaciones en cada perfil y que estuvo presente en la prueba. La proporción de los errores de clasificación promedio de todos los perfiles probados se toma como la tasa de error global. Nuevos datos pueden ser predecidos por la agregación de los árboles Nt y por selección del que tuvo el mayor número de veces que coincidieron (Breiman, 2001). Los Rfs usados en este estudio son generados con el programa "Random Forest".

Validación. Para la optimización de los parámetros y la evaluación del modelo, se realiza una evaluación cruzada con la unión de los resultados de las pruebas (Parker *et al*., 2007; Varma y Simon, 2006; Witten *et al*., 2005). En la validación cruzada, el grupo de datos de entrada se dividen en *k* partes de igual tamaño, generalmente de una forma aleatoria. Para la parte k^{th}, el modelo funciona con la otra parte *k*-1, mientras el comportamiento es evaluado con la parte respectiva. Este proceso de entrenamiento y validación se realiza por cada parte y la validación cruzada de la estimación del error es igual al promedio de los errores obtenidos en las diferentes veces (Bishop, 2006; Hastie *et al*., 2009). En general, cinco y 10 validaciones cruzadas son recomendadas para la selección del modelo (Breiman, 1992; Kohavi, 1995). En muchos mundos reales los grupos de datos, o el número de datos por clase varían. Un buen enfoque para tratar con estas desproporciones es realizar una validación cruzada evaluation, cross-validation is performed with pooling of the test results (Parker *et al*., 2007; Varma and Simon, 2006; Witten *et al*., 2005). In cross-validation, the input data set is split into *k* equally sized parts, generally in a random manner. For the *k* part, the model is trained with the k^{-1} other parts, while the performance is evaluated with the respective part. This training and validation process is done for each part and the cross-validation estimation of the errorequals the average of the errors obtained by the different folds (Bishop, 2006; Hastie *et al*., 2009). In general, five-fold and ten-fold cross-validation is recommended for model selection (Breiman, 1992; Kohavi, 1995). In many real-world data sets, the number of data instances varies per class. A good approach for dealing with these disproportions is to perfor m a stratified cross-validation. Herein, the different folds are so-called stratified so that, based on the different class labels, they contain approximately the same proportions of data points as the original data set (Kohavi, 1995). Because RF do not overfit (Breiman, 2001), parameter optimization is done on the test set. We have chosen to do so because some species correspond to a small number of FAME profiles and learning would be based on an even smaller number of FAME profiles. Nonetheless, it is better experimental practice to perform parameter optimization in a separate cross-validation process and use the test set only for identification purposes. Ultimately, pooling is done of the separate cross-validation folds (or test subsets), implying a model performance estimate based on the complete data set. As only species are considered associated with a minimum of four FAME profiles per species, a four-fold cross-validation is performed.

Evaluation. For both techniques, the probability estimates are statistically analyzed. For each profile, the species label associated with the highest probability estimate is considered. When dealing with two classes, a popular method for evaluation of the performance of a classifier is to construct a confusion matrix. A confusion matrix summarizes the predictions by reporting the number of true positives (TP or positive instances that are predicted as positive), false positives (FP or negative instances that are predicted as positive), true negatives (TN or negative instances that are predicted as negative) and false negatives (FN or positive instances that are predicted as negative). From this matrix different performance measures can be calculated such as sensitivity (TP/(TP+FN)), precision (TP/(TP+FP)) and F-score (2*S*P)/(S+P), with S and P the sensitivity and precision respectively). Note that when no TP and FP results are obtained, precision resolves in a value equal to infinity. For the F-score, this case, together with a denominator of zero, will also resolve in a value of infinity. When, however, confronted with more than two bacterial species, or thus a multi-class problem, a multi-class confusion matrix can be composed. Subsequently, evaluation can be done by decomposing the multi-class confusion matrix to *N* two-class confusion matrix, with *N* the number of classes or species considered. In each two-class confusion matrix, the corresponding species is evaluated against all other species implemented in the classifier. In other words, all other species are considered as one dummy species. Finally, a global

estratificada. Aquí, los diferentes subgrupos son llamados estratificados de modo que, basados en las diferentes clases, tienen aproximadamente las mismas proporciones de puntos de datos como el grupo de datos originales (Kohavi, 1995).
Porque RF no origina errores aleatorios (Breiman, 2001), la optimización de parámetros se realiza en el grupo de prueba. Hemos escogido hacer esto porque algunas especies corresponden a un pequeño número de perfiles FAME y el aprendizaje se basa en un número aún más pequeño de perfiles FAME. Sin embargo, una mejor práctica experimental es realizar la optimización de parámetros en un proceso de validación cruzada y usar los grupos de prueba solamente para propósitos de identificación. Finalmente, el agrupamiento es realizado por la validación cruzada de los "folds" (o de las pruebas de los subgrupos), implicando un modelo de desempeño estimado basado en el conjunto completo de datos. Como las especies son consideradas solamente asociadas con un mínimo de cuatro perfiles FAME por especie, se realizan validaciones cruzadas por cuadriplicado.
Evaluación. Para ambas técnicas, la probabilidad estimada es analizada estadísticamente. Para cada perfil, las especies etiquetadas son consideradas con la probabilidad más alta. Cuando se trata con dos clases, un método popular para la evaluación del comportamiento de un clasificador es la construcción de una matriz confusa. Una matriz confusa resume las predicciones al reportar el número de positivos verdaderos (TP o casos positivos que son previstos como positivos), falsos positivos (FP o casos negativos que son predichos como positivos, verdaderos negativos (TN o casos negativos que son predichos como negativos) y falsos negativos (FN o casos positivos que son previstos como negativos).

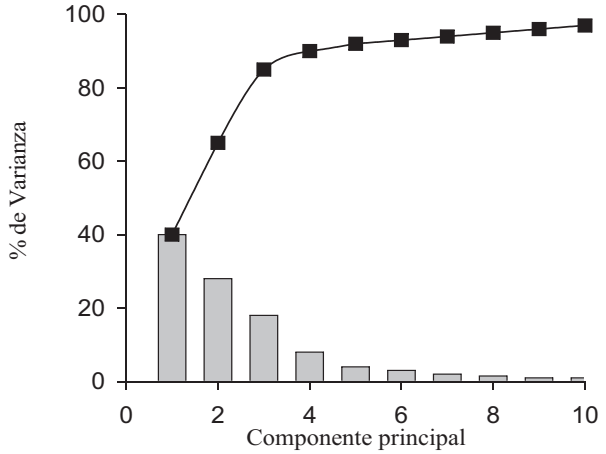

Figura 2. Diagrama de sedimentación derivado del análisis de componentes principales con datos de fitopatógenos. Se muestra el porcentaje de varianza asociado con los 10 primeros componentes principales y la variación acumulada.
Figure 2. Skree plot resulting from principal component analysis of the plant pathogen data set. The percentage of variance associated with the first ten principal components is visualized, together with the cumulative variance.

measure can be calculated by averaging the results of the different confusion matrices.

RESULTS AND DISCUSSION
PCA. We have investigated how FAME data could be used to discriminate between plant-pathogenic *Pseudomonas* species, and between this group and non-plant-pathogenic *Pseudomonas* species. Skree plots are given following principal component analysis on both corrsponding data sets. These plots are given in Figures 2 and 4. From both plots, it becomes clear that three to four principal components can represent 90% to 95% of the variability in the data. This not only indicates that a lot of features (peaks) are correlated with each other, but also that the dimensionality of the data sets (number of peaks) can be significantly decreased, while keeping a similar discrimination possible for the different species or species groups. This correlation may be attributed to the fatty acid biosynthesis pathway through which different fatty acids are typically converted into other fatty acid molecules (Madigan *et al*., 2009). Furthermore, the discrimination between the different species or species groups is visualized by a biplot of the first two principal components, see Figures 3 and 5. From the first biplot, which corresponds to the plant-pathogenic data set, it is clear that plant-pathogenic species are hard to distinguish from each other based on FAME data. Also, some distinct groupings can immediately be seen such as *P. agarici*, *P. corrugata, P. flavescens*, *P. tolaasii*. The species present in the cluster on the left correspond to the *P. syringae* group (Anzai *et al.*, 2000). Remark that in this biplot, the species belonging to the *P. beteli* group and the species *P. flectens* are not included to allow for a better scaling of the other species. This can be explained by the distance between these species and the *Pseudomonas sensu stricto* species, leading to a tight clustering of the latter species in the respective principal component space. When compared to non-plant-pathogenic species, the second biplot shows that the plant-pathogenic FAME profiles mainly cluster in one data cloud, and the non-plant-pathogenic data cluster in two distinct FAME clouds, with one cloud clearly overlapping with the plant-pathogenic FAME clouds.
Machine learninig. The first row of Table 2 summarizes the results of the machine learning experiment by which a possible discrimination of plant-pathogenic *Pseudomonas* species is investigated. These results are similar to those resulting from the machine learning analysis of the complete data set, in the sense that similar metric values are found (Slabbnick *et al*., 2009). Even though a smaller number of species is investigated, it is clearly not straightforward to distinguish the different plant-pathogenic *Pseudomonas* species from one another. These findings are supported by the principal components analysis in the previous section and Figure 1. From the corresponding biplot of the first two principal components, it is obvious that the FAME patterns of the different species overlap, making it not easy for machine learning techniques to find good flexible mathematical functions representing the species boundaries. We can conclude that FAME analysis is not a very good technique to

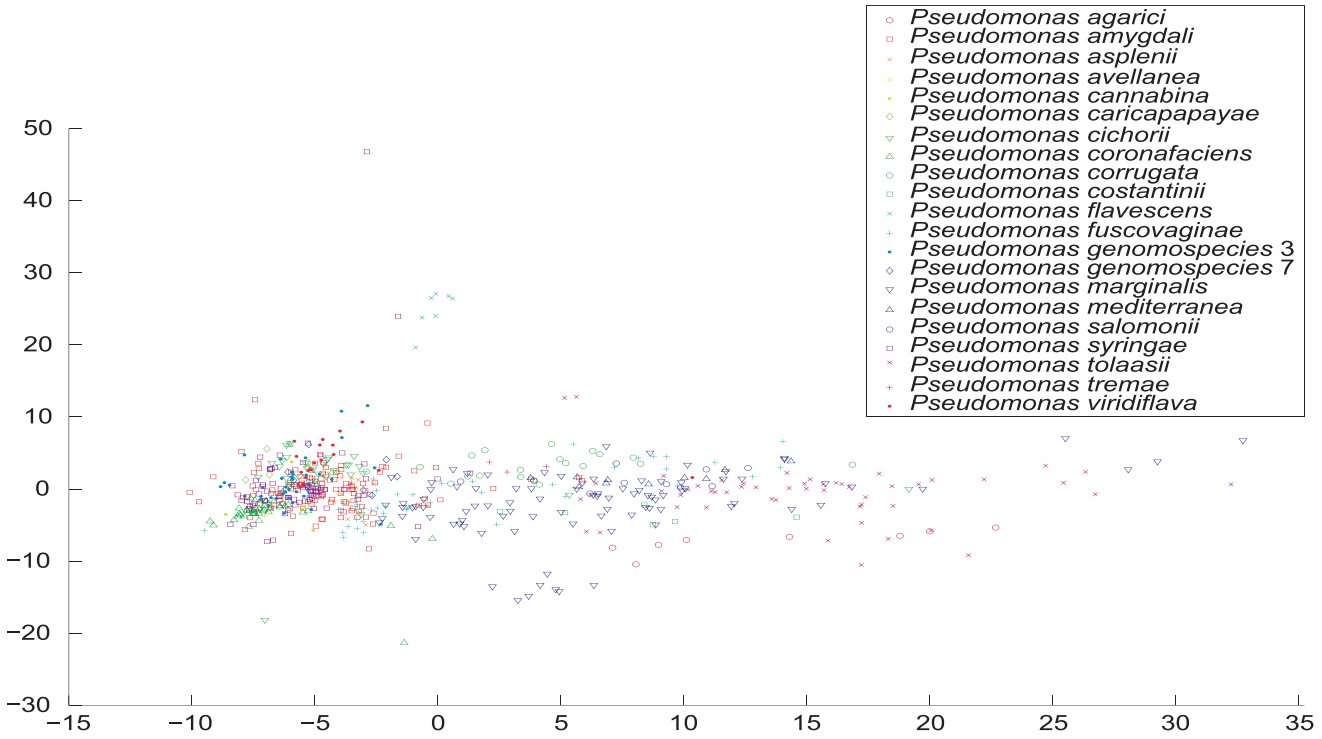

Figura 3. Comportamiento de los dos primeros componentes principales tras un análisis del conjunto de datos de fitopatógenos. Las especies del grupo de *Pseudomonas beteli* y *Pseudomonas flectens* no fueron incluídas para lograr una mejor escala del resto de especies.
Figure 3. Biplot of the first two principal components following analysis of the phytopathogen data set. The species of the *Pseudomonas beteli* group and *Pseudomonas flectens* are not included for achieving a good scaling of the other species.

De esta matriz pueden ser calculadas diferentes mediciones de desempeño tales como sensibilidad (TP/(TP+FN)), precisión (TP/(TP+FP)) y F-score ((2*S*P)/(S+P), con S y P como sensibilidad y precisión respectivamente. Nótese que cuando no se obtienen resultados TP y FP, la precisión lleva a un valor igual al infinito. Para el F-score, este caso, junto con el denominador de cero, puede también llevar a un valor de infinito. Sin embargo, cuando, se confrontan con más de dos especies bacterianas, por lo tanto a un problema de múltiples clases, se puede tener una matriz de confusión compuesta. Subsecuentemente, la evaluación puede ser hecha mediante la fragmentación de la matriz de confusión multiclase a una matriz de confusión *N* dos clases, donde *N* es el número de clases o especies consideradas. En cada matriz de confusión de dos clases, las especies correspondientes son evaluadas contra todas las demás especies implementadas en el clasificador. En otras palabras, todas las demás especies son consideradas como una especie ficticia. Finalmente, una medida global puede ser calculada promediando los resultados de las diferentes matrices de confusión.

RESULTADOS Y DISCUSIÓN

ACP. Los datos que tenemos evaluados hasta ahora de FAME pueden ser usados para discriminar entre especies fitopatogénicas de *Pseudomonas*, y entre este grupo y especies no fitopatógenas de *Pseudomonas*. El análisis de

Pseudomonas species could be discriminated from the group distinguish between all plant-pathogenic *Pseudomonas* species. However, some species could clearly be distinguished. Species with an F-score larger than 0.8 are *P. cissicola, P. corrugata*, *P. flavescens*, *P. flectens*, *P. fuscovaginae*, *P. marginalis*, *P. tolaasii*, *P. tremae* and *P. viridiflava*. For some of these species, distinct groups could also be found in the PCA biplots. Whether this conclusion holds for plant-pathogenic species in other genera remains a topic for future investigation.

In a second machine learning experiment, we have investigated how well the group of plant-pathogenic *P. fuscovaginae*, *P. marginalis*, *P. tolaasii*, *P. tremae* and *P. viridiflava*. For some of these species, distinct groups could also be found in the PCA biplots. Whether this conclusion holds for plant-pathogenic species in other genera remains a topic for future investigation. In a second machine learning experiment, we have investigated how well the group of plant-pathogenic *Pseudomonas* species could be discriminated from the group of non-plant-pathogenic *Pseudomonas* species. The same learning setting as described in the previous experiment is used. However, because a larger number of data points per class is available, we have chosen to perform a ten-fold cross-validation. Again, because RFs do not overfit, parameter optimization is also done by the cross-validation fold. The test results of this

Cuadro 2. Resultados de los experimentos de árboles al azar referentes a identificación de especies de *Pseudomonas*, identificación de especies fitopatógenas y no fitopatógenas dentro del género *Pseudomonas*, prueba de F y desviación estándar.
Table 2. Results of the random forests experiments about *Pseudomonas* species identification, identification of plant-pathogenic species and non-plant-pathogenic species (NPPS) within the genus *Pseudomonas*, precision and F-score and, standard deviations.

Técnica[a]	Sensibilidad[b]	Precisión[c]	F[d]
PPS	0.631 (0.297)[e]	0.811 (0.152)	0.711 (0.181)
PPS vs NPPS	0.961	0.907	0.859

[a]Usando la base de datos de especies patogénicas. [b]Dentro del grupo de *Pseudomonas*. [e, f]Porcentaje entre todas las clases (cada especie contra el resto de especies). [c]Desviación estandar.
[a]Using te data base of pathogenic species. [b]Inside of the group *Pseudomonas*. cPercentaje among all classes (each species against all other species). [d]Standar desviation.

sedimentación obtenido corresponde al análisis de componentes principales en ambos grupos de datos. Estos datos están marcados en las Figuras 2 y 4. De ambas figuras, queda claro que tres a cuatro componentes principales pueden representar el 90 a 95% de la variabilidad en los datos. Esto nos indica solamente que muchas de las características (picos) son correlacionados con cada uno, pero también la dimensionalidad del grupo de datos (número de picos) pueden ser significativamente disminuidos, mientras mantenga una discriminación similar posible para las diferentes especies o grupos de especies. Esta correlación puede ser atribuida a la ruta biosintética de ácidos grasos a través de los cuales diferentes ácidos grasos son típicamente convertidos en otras moléculas de ácidos grasos (Madigan *et al*., 2009). Además, la discriminación entre las diferentes especies o grupos de especies es visualizada por gráficos de los primeros dos componentes principales, ver Figura 3 y 5. Del primer gráfico, el cual corresponde al grupo de datos de fitopatógenos, es claro que las especies fitopatógenas son fuertes para distinguir de otros basados en los datos FAME. También, algunos grupos distintivos pueden inmediatamente ser vistos tales como *P. agarici*, *P. corrugata*, *P. flavescens*, *P. tolaasii*. Estas especies presentes en el agrupamiento de la izquierda corresponde al grupo de *P. syringae* (Anzai *et al.*, 2000). Se remarca que en esta gráfica, las especies pertenecientes al grupo *P. beteli* y la especie de *P. flectens* no están incluidos permitiendo su cambio a otras especies. Esto puede ser explicado por la distancia entre estas especies y las especies de *Pseudomonas sensu stricto*, permitiendo posteriormente un estricto agrupamiento de las especies en el espacio respectivo de componentes principales. Cuando comparamos a una especie no fitopatógenas, el segundo biplot muestra que el perfil de FAME de las fitopatógenas se agrupa principalmente en una nube de dispersión y el grupo de datos de las especies no fitopatógenas en dos distintas nubes de dispersión FAME, con una nube claramente sobrepuesta con las nubes FAME de las fitopatógenas.

Aprendizaje automatizado. La primera columna del Cuadro 2 resume los resultados del experimento de aprendizaje two-class classification experiment are reported in the second row of Table 2. It can immediately be concluded that a high discrimination is possible between plant-pathogenic and non-plant-pathogenic *Pseudomonas* species using FAME data. The high metric values are somewhat surprising given the overlapping data clouds as visualized by the principal component analysis (see Figure 5). It is clear that some relation exists between plant-pathogenicity and the fatty acid content.The probability estimates as given by the RF experiment are statistically analyzed by a Wilcoxon rank-sum test. This test assumes randomly sampled observations that are independent and continuous and that the shapes of the underlying distributions are identical (Higging, 2004; Sheskin, 2004). In this experiment, the first assumptions are satisfied, while the last assumption is considered to be met. A *p* value of approximately zero is obtained, implying statistically different probability estimates at the significance level of 0.05. A possible explanation for these strong identification results could be found in the paper of Stead (1992) who describes a clustering of a multitude of the plant-pathogenic *Pseudomonas* species by hydroxy fatty acids. Though, in this study, the genus *Pseudomonas sensu lato* is considered and major discriminations are discussed. Herein, one major group corresponds to the genus *Pseudomonas sensu stricto*. This group consisted of about 35 plant-pathogenic taxa that could mainly be discriminated based on the hydroxy fatty acids $C_{10:0}$ 3-hydroxy and $C_{12:0}$ 3-hydroxy. Subgrouping could also be achieved by these hydroxy fatty acids, together with the hydroxy fatty acid $C_{12:0}$ 2-hydroxy.

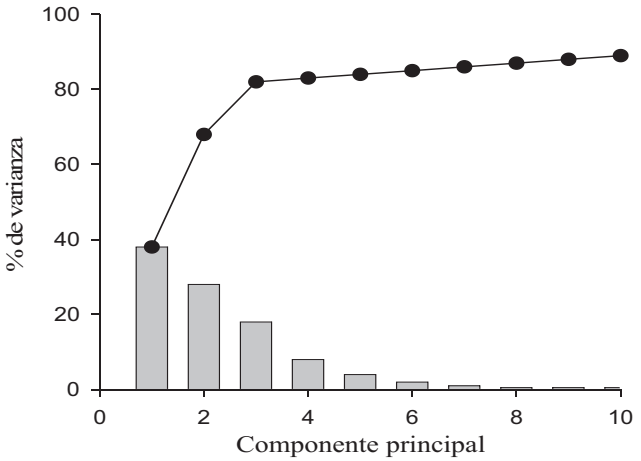

Figura 4. Diagrama de sedimentación del Análisis de componente principal de fitopatógenos *vs* no fitopatógenos, porcentaje de varianza acumulado y porcentaje de varianza relacionado con el primero de los 10 componentes principales.
Figure 4. Skree plot resulting from principal component analysis of the plant pathogen vs. non-plant pathogen data set. The percentage of variance associated with the first ten principal components is visualized, together with the cumulative variance.

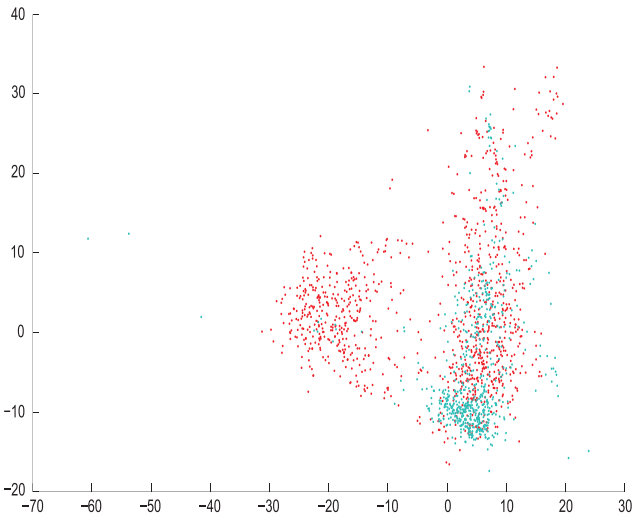

Figura 5. Primeros dos componentes principales derivados del análisis de datos de fitopatógenos *vs* no fitopatógenos. Los puntos verdes señalan especies con una o más razas fitopatógenas, puntos rojos son referidos a especies no asociadas a razas fitopatógenas.

Figure 5. Biplot of the first two principal components following analysis of the plant pathogen *vs* non-plant-pathogen data set. Green dots denote species with one or more plant-pathogenic strains, while red dots denote species not associated with plant-pathogenic strains.

automatizado por los cuales se investiga la posible discriminación de las especies de *Pseudomonas* fitopatógenas. Estos resultados son similares a los resultados del análisis de aprendizaje automatizado de los grupos de datos completos, en el sentido de que son encontrados similares los valores métricos (Slabbnick *et al*., 2009). Aunque un número más pequeño de especies es investigado, queda claro que no es sencillo para diferenciar las diferentes especies fitopatógenas de otras *Pseudomonas*. Estos resultados son soportados por los análisis de componentes principales en la sección previa y Figura 1. De los correspondientes gráficos de los primeros dos componentes principales, es obvio que los patrones FAME se sobreponen en las diferentes especies, haciendo que la técnica de aprendizaje automatizado no sea fácil para encontrar una buena función matemática flexible que represente la amplitud de las especies.Podemos concluir que el análisis FAME no es una buena técnica para diferenciar entre todas las especies fitopatógenas de *Pseudomonas*. Sin embargo, algunas especies podrían ser claramente diferenciadas.

Especies con un F-score más grande que 0.8 son *P. cissicola*, *P. corrugata*, *P. flavescens*, *P. flectens*, *P. fuscovaginae*, *P. marginalis*, *P. tolaasii*, *P. tremae* y *P. viridiflava*. Para algunas de estas especies, grupos distintos podrían también ser encontrados en la gráfica del PCA. Si esta conclusión permanece para especies fitopatógenas en otros géneros este es un tópico para investigaciones futuras. En un segundo experimentos sobre aprendizaje automatizado, se ha investigado como un grupo de especies fitopatógenas de *Pseudomonas* podrían ser discriminados de los grupos de especies no fitopatógenas. Se estableció el mismo sistema de aprendizaje automatizado usado anteriormente. Sin embargo, For this group, Stead (1992) concluded that qualitative and quantitative differences in many of these fatty acids were found. No comparison was, however, made with non-plant-pathogenic *Pseudomonas* species. Nonetheless, these discrimination can also be assumed to be very valuable for discrimination between plant-pathogenic and non-plant-pathogenic species. Future study of this relation should, however, reveal all determining FAME constituents and the corresponding qualitative and quantitative differences.

CONCLUSIONS

We have evaluated the possibilities of FAME analysis for the identification of plant-pathogenic *Pseudomonas* species. Two cases have been considered: identification at species level and a discrimination of the group of plant-pathogenic species from the non-plant-pathogenic species. An initial simple analysis of the data was done by principal component analysis. From these experiments, it has become clear that only a few principal components exploit a large amount of the variability in the data. This is mainly due to the correlation between different fatty acids. Nonetheless, when looking at the biplot of the first two principal components, the fatty acid profiles of different plant-pathogenic species are clearly overlapping. Regarding the plant pathogen versus non-plant pathogen data set, a distinct data cloud can be seen. However, an overlap is still seen in the data. This data analysis results in impor tant knowledge when considering machine learning for identification purposes. Due to overlapping FAME clouds, flexible species boundaries need to be learned. In the case of distinguishing between plant-pathogenic *Pseudomonas* species, this learning process appears to be quite hard as only a moderate identification performance is achieved. Nonetheless, some species could clearly be identified. When discriminating plant-pathogenic *Pseudomonas* species from non-plant-pathogenic *Pseudomonas* species, RFs are able to achieve a good identifica tion of either group. Besides this, we have statistically also shown that the FAME profiles of both groups are significantly different. In other words, a clear statistical relation exists between certain fatty acids and plant pathogenesis. Future work should reveal which constituents are major players in this relation. Other interesting topics for future research may comprise the relations to and between plant pathogens in other bacterial genera.

LITERATURA CITADA

Anzai, Y., Kim, H., Park, J.-Y., Wakabayashi, H., and Oyaizu, H. 2000. Phylogenetic affiliation of the Pseudo- monads based on 16S rRNA sequence. International Journal of Systematic and Evolutionary Microbiology 50: 1563–1589.

Bishop, C.M. 2006. Pattern Recognition and Machine Learning, 1st Edition, Springer, New York. 738 p.

Breiman, L., Spector, P. 1992. Submodel selection and evaluation in regression: the X-random case. International Statistical Review 60: 291-319.

Breiman,L. 2001. Random Forests. Machine Learning 45: 5-32.

debido al gran número de datos disponibles en los diagramas de dispersión disponibles, se decidió llevar a cabo la validación de "ten-fold cross-validation". Nuevamente, debido a un no sobre ajuste de RF, la optimización de parámetros es también hecho por la "cross-validation fold". Los resultados de la prueba de esta segunda clase de experimento para clasificación son reportados en la segunda columna del Cuadro 2. Se puede concluir inmediatamente que es posible una alta discriminación entre especies fitopatógenas y no fitopatógenas usando datos FAME. El alto valor métrico es algo que sorpresivamente se observa con las nubes de dispersión sobrepuestas visualizado en el análisis de componentes principales (ver Figura 5).

Está claro que existen algunas relaciones entre el contenido de ácidos grasos de especies fitopatógenas. La probabilidad estimada obtenida en el experimento RF son estadísticamenteanalizadas con la prueba de Wilcoxon Rank-sum. Esta prueba asume observaciones de muestras al azar que son independientes y continuas y que la forma de la distribuciones subsecuentes son idénticas (Higging, 2004; Sheskin, 2004).En este experimento, los primeros supuestos están confirmados, mientras el último supuesto se considera como presente. Un valor-*p* de aproximadamente cero es obtenido, implicando probabilidades estimadas estadísticamente diferentes en el nivel de significancia de 0.05. Una posible explicación para esta fuerte identificación podría ser encontrado en el artículo de Stead (1992), quien describe un agrupamiento de un gran número de especies fitopatógenas de *Pseudomonas* por hidroxi-ácidos grasos. Aunque, en este estudio, se ha considerado principalmente datos del género *Pseudomonas sensu lato* para discusión. Aquí, un grupo principal corresponde al género *Pseudomonas sensu stricto*. Este grupo consistió de cerca de 35 taxa de fitopatógenas que podrían ser principalmente discriminadas con base en los ácidos grasos hidroxy $C_{10:0}$ 3-hidroxy y $C_{12:0}$ 3-hidroxy. Subgrupos podrían ser también obtenidos por estos ácidos grasos hidroxy, juntos con el ácidos grasos $C_{12:0}$ 2-hidroxy. Para este grupo, Stead (1992) concluyó que se encontraron muchas diferencias cuantitativas y cualitativas en estos ácidos grasos. No se realizaron comparaciones con especies no fitopatógenas de *Pseudomonas*. Sin embargo, estas discriminaciones pueden ser también asumidas y es muy valorable para la discriminación entre especies patogénicas y no patogénicas. Futuros estudios de esta relación debería, sin embargo, revelar todos los constituyentes que determina FAME y las diferencias cualitativas y cuantitativas correspondientes.

CONCLUSIONES

Hemos evaluados las posibilidades del análisis FAME para la identificación de especies fitpatogénicas. Se han considerado dos casos; identificación a nivel de especie y una discriminación del grupo de especies fitpopatógenas de las no fitopatógenas. Un simple análisis inicial de los datos fue hecho por el análisis de componentes principales. De estos experimentos, surge claramante que solamente unos pocos componentes principales explican la gran variabilidad en los datos. Esto es principalmente debido a la correlación entre

Brenner, D.J., Krieg, N.R., and Staley, J.T. 2005. Bergey's Manual of Systematic Bacteriology Volume 2: The Proteobacteria Part A Introductory Assays, Springer, New York, NY, USA. 304 p.

Catara, V., Sutra, L., Morineau, A., Achouak, W., Christen, R., and Gardan, L. 2002. Phenotypic and genomic evidence for the revision of *Pseudomonas corrugata* and proposal of *Pseudomonas mediterranea* sp. nov. International Journal of Systematic and Evolutionary Microbiology 52:1749-1758.

Dees, S.B., Moss, C.W. 1975. Cellular fatty acids of *Alcaligenes* and *Pseudomonas* species isolated from clinical specimens. Journal of Clinical Microbiology: 1:414-419.

Dees,S.B., Moss, C.W., Weaver, R.D., and Hollis D. 1979. Cellular fatty acid composition of *Pseudomonas paucimobilis* and groups IIk-2, Ve-1, and Ve-2. Journal of Clinical Microbiology 10: 206-209.

Duda, R.O., Hart, P.E., and Stork, D.G. 2001. Pattern Classification, John Wiley & Sons, New York, USA. 654 p.

Fensom, A.H., and Gray, G.W. 1969. The chemical composition of the lipopolysaccharide of *Pseudomonas aeruginosa*. Biochemical Journal 114: 185-196.

Gardan, L., Shafik, H., Belouin, S., Broch, R., Grimont, F., and Grimont, P.A.D. 1999. DNA relatedness among the pathovars of *Pseudomonas syringae* and description of *Pseudomonas tremae* sp. nov. and *Pseudomonas cannabina* sp. nov. (*ex* Sutic and Downson 1959). International Journal of Systematic Bacteriology 49: 469-478.

Gardan, L., Bella, P., Meyer, J.M., Christen, R., Rott, P., Achouak, W., and Samson, R., 2002. *Pseudomonas salomonii* sp. nov., pathogenic on garlic, and *Pseudomonas palleroniana* sp. nov., isolated on rice. International Journal of Systematic and Evolutionary Microbiology 52: 2056-2074.

Hancock, I.C., Humphreys, G.O., and Meadow, P.M. 1979. Characterisation of the hydroxy acids of *Pseu- domonas aeruginosa* 8602. Biochimica et Biophysica Acta (BBA) 202: 389-391.

Hastie, T., Tibshirani, R., and Friedman J. 2009. The Elements of Statistical Learning: Data Mining, Infer- ence and Prediction, 2nd Edition, Springer-Verlag, New York, USA. 763 p.

Higgings, J.J. 2004. An Introduction to Modern Nonparametric Statistics, Brooks/Cole-Thomson Learning, CA, USA. 366 p.

Hildebrand, D.C., Palleroni, N.J., Hendson, M., Toth, J., and Johnson, J.L. 1994. *Pseudomonas flavescens* sp. nov., isolated from walnut blight cankers. International Journal of Systematic Bacteriology 44: 410-415.

Höfte, M., and De Vos, P. 2007. Plant-pathogenic *Pseudomonas* species. p. 507-536. In: G.S. S. (Eds.) Plant-associated bacteria, Springer, Dordrecht, The Netherlands. 712 p.

diferentes ácidos grasos. Sin embargo, cuando se observa el biplot del primero de los dos componentes principales, los perfiles de ácidos grasos de diferentes especies fitopatógenas están claramente sobrepuestos. En relación al grupo de datos de los patógenos de plantas versus no patogénicos, una distinta nube de dispersión de los datos puede ser vista. Sin embargo, una sobreposición es aún vista en los datos. Los resultados del análisis de datos es un conocimiento importante cuando se considera el aprendizaje automatizado para propósitos de identificación. Debido a los diagramas de dispersión FAME sobrepuestos, se necesita conocer más de la amplitud de especies flexibles. En el caso de de distinguir entre especies de *Pseudomonas* fitopatógenas, este proceso de aprendizaje parece ser muy difícil cuando solamente se lleva a cabo una identificación moderada. Sin embargo, algunas especies podrían ser claramente identificadas. Cuando discriminamos las especies fitopatógenas de *Pseudomonas*, los RFs son capaces de llevar a cabo una buena identificación de los grupos. Además de esto, tenemos estadísticamente también mostrado que los perfiles de FAME de ambos grupos son significantemente diferentes. En otras palabras, una clara relación estadística existe entre ciertos ácidos grasos y lapatogénesis en plantas. Trabajos futuros deberán revelar cual constituyentes juegan un rol principal en esta relación. Otro tópico interesante para futuras investigaciones puede comprender las relaciones entre fitopatógenos en otros géneros bacteriales.

Hu, F.-P., Yount, J.M., Stead, D.E., and Goto, M. 1997. Transfer of *Pseudomonas cissicola* (Takimoto 1939) Burkholder 1948 to the genus *Xanthomonas*. International Journal of Systematic Bacteriology 47:228-230.

Ikemoto, S., Kuraishi, H., Komagata, K., Azuma, R., Suto T., and Murooka, H. 1978. Cellular fatty acid com- position in *Pseudomonas* species. Journal of General and Applied Microbiology 24 :199-213.

Janse, J.D., Rossi, P., Angelucci, L., Scortichini, M., Derks, J.H.J., Akermans, A.D.L. and De Vrijer,R. 1996. P.G. Psallidas, Reclassification of *Pseudomonas syringae* pv. *avellanae* as *Pseudomonas avellanae* spec. nov. the bacterium causing canker of hazelnut (*Corylus avellanae* L.). Systematic and Applied Microbiology 19:589-595.

Kersters, K., Ludwig, W., Vancanneyt, M., De Vos, P., Gillis, M., and Schleifer, K.H. 1996. Recent changes in the classification of pseudomonads: an overview. Systematic and Applied Microbiology 19: 465-477.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montréal, Canada, 1995. p. 1137-1145.

Letunic, I., and Bork, P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127-128.

Madigan, M.T., Martinko, J.M., Dunlap, P.V., and Clarck, D.P. 2009. Brock Biology of Microorganisms, 12th Edition, Pearson Education Inc., San Francisco, USA. 1061 p.

Mitchell, T.M., 1997. Machine Learning, McGraw-Hill, Boston, USA. 414 p.

Miyajima, K., Tanii, A., and Akita, T. 1983. *Pseudomonas fuscovaginae* sp. nov., nom. rev. International Journal of Systematic Bacteriology 33:656-657.

Moore,E.R.B., Mau, M., Arnscheidt, A., Böttger, E.C., Hutson, R.A., Collins, M.D., Van de Peer, Y., De Wachter, R., and Timmis, K.N. 1996. The determination and comparison of the 16S rRNA gene sequences of species of the genus *Pseudomonas* (*sensu stricto*) and estimation of the natural intrageneric relation- ships. Systematic and Applied Microbiology 19:478-492.

Moss, C.W., Samuels, S.B., and Weaver, R.B., Cellular Fatty Acid Composition of Selected *Pseudomonas* Species. Applied Microbiology 24:596-598.

Moss, C.W., and Samuels, S.B. 1974. Short-chain acids of *Pseudomonas* species encountered in clinical specimens. Applied Microbiology 27: 570-574.

Moss, C.W., 1975. Gas-liquid chromatography as an analytical tool in microbiology. Journal of Chromatography 203:337-347.

Moss, C.W., and Dees, S.B. 1975. Identification of microorganisms by gas chromatographic-mass spectrometric analysis of cellular fatty acids. Journal of Chromatography 112:595-604.

Moss, C.W., and Dees, S.B.1976. Cellular fatty acids and metabolic products of *Pseudomonas* species ob- tained from clinical specimens. Journal of Clinical Microbiology 4: 492-502.Moss, C.W. 1981. Gas-liquid chromatography as an analytical tool in microbiology. Journal of Chromatography 203:337-347.

Munsch, P., Alatossava, T., Marttinen, N., Meyer, J.M., Christen, R., and Gardan, L. 2002. *Pseudomonas costantinii* sp. nov., another causal agent of brown blotch disease, isolated from cultivated mushroom sporophores in Finland. International Journal of Systematic and Evolutionary Microbiology 52:1973-1983.

Osterhout,G.J., Shull, V.H., and Dick, J.D. 1991. Identification of clinical isolates of Gram-negative nonfermentative bacteria by an automated cellular fatty acid identification system. Journal of Clinical Microbiology 29:1822-1830.

Oyaizu, H., and Komagata, K., 1983. Grouping of *Pseudomonas* species on the basis of cellular fatty acid composition and the quinone system with special reference to the existence of 3-hydroxy fatty acids. Journal of General and Applied Microbiology 29:17-40.

Palleroni, N.J., Kunisawa, R., Contopoulou, R., Doudoroff, R., 1973. Nucleic acid homologies in the ge- nus Pseudomonas. Systematic and Applied Microbiology 23: 333-339.

Palleroni, N.J., 1984. Genus I. *Pseudomonas* Migula 1984. p. 141-199. *In*: N.R. Krieg, J.G. Holt (Eds.), Bergey's Manual of Systematic Bacteriology Volume 1, Williams and Wilkins, Baltimore, USA. 721 p.

Palleroni, N.J., 2005. Genus I. *Pseudomonas Migula* 1894, $237^{AL.}$. p. 322-379. *In*: D.J. Brenner, N.R. Krieg, J.T. Staley (Eds.), Bergey's Manual of Systematic Bacteriology Volume 2: The Proteobacteria Part B The

Gammaproteobacteria, Springer, New York, USA. 1108 p.

Palleroni, N.J. 2008. The Road to the Taxonomy of *Pseudomonas.* p. 1-18. *In*: P. Cornelis (Ed.) *Pseudomonas*: Genomics and Molecular Biology, Caister Academic Press, Norfolk, USA. 244 p .

Parker, J.P., Günter, S., and Bedo, J. 2007. Stratification bias in low signal microarray studies. BMC bioinformatics 8:1-16.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W.G., Peplies, J., and Glöckner, F.O. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data com- patible with ARB. Nucleic Acids Research: 35:7188-7196.

Sheskin, D.J. 2004. Handbook of Parametric and Nonparametric Statistical Procedures, 3rd Edition, Chapman and Hall/CRC, Florida, USA. 1193 p.

Slabbinck, B., De Baets, B., Dawyndt, P., and De Vos P. 2009, Towards large-scale FAME-based bacterial species identification using machine learning techniques. Systematic and Applied Microbiology 32:163-176.

Smith, I.M., Dunez, J., Philips, D.H., Lelliot, E.A., and Archer, S.A. 1988. European Handbook of Plant Diseases, Blackwell Scientific Publications, Oxford, UK. 583 p.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analysis with thou- sands of taxa and mixed models. Bioinformatics: 22:2688-2690.

Stead, D.E. 1992. Grouping of plant-pathogenic and some other *Pseudomonas* spp. by using cellular fatty acid profiles. International Journal of Systematic Bacteriology 24:281-295.

Stead, D.E., Sellwood, J.E., Wilson, J., and Viney, I. 1992. Evaluation of a commercial microbial identification system based on fatty acid profiles for rapid, accurate identification of plant-pathogenic bacteria. Journal of Bacteriology 72:315-321.

Tayeb, L.A., Ageron, E., Grimont, F., and Grimont P.A.D. 2005. Molecular phylogeny of the genus *Pseudomonas* based on *rpoB* sequences and application for the identification of isolates. Research in Microbiology 156:763-773.

Van Den Mooter, M., and Swings, J. 1990. Numerical analysis of 295 phenotypic features of 266 *Xanthomonas* strains and related strains and an improved taxonomy of the genus. International Journal of Systematic Bacteriology 40: 348-369.

Vancanneyt, M., Witt, S., Abraham, W., Kersters, K., and Frederickson, H.L. 1996. Fatty acid content in whole-cell hydrolysates and phospholipid fractions of Pseudomonads: a taxonomic evaluation. Systematic and Applied Microbiology 19:528-540.

Varma, S., and Simon, R. 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7: 91.

Witten, I.H., and Frank, E. 2005. Data Mining Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, San Francisco, USA. 525 p.