



**From learning taxonomies to phylogenetic learning:
a computational approach to FAME-based bacterial
species identification**

ir. Bram Slabbinck

Supervisors

Prof. dr. Bernard De Baets

Research Unit Knowledge-based Systems

Department of Applied Mathematics, Biometrics and Process Control

Faculty of Bioscience Engineering

Ghent University

Prof. dr. Paul De Vos

Laboratory of Microbiology

Department of Biochemistry and Microbiology

Faculty of Sciences

Ghent University

Prof. dr. Peter Dawyndt

Department of Applied Mathematics and Computer Science

Faculty of Sciences

Ghent University

Dean

Prof. dr. ir. Guido Van Huylenbroeck

Rector

Prof. dr. Paul Van Cauwenberge

From learning taxonomies to phylogenetic learning:
a computational approach to FAME-based bacterial
species identification

ir. Bram Slabbinck

Thesis submitted in fulfillment of the requirements for the degree of
Doctor (PhD) in Applied Biological Sciences

Dutch translation of the title:

Van het leren van taxonomieën tot fylogenetisch leren: een computationele benadering voor een vetzuur-gebaseerde identificatie van bacteriële soorten.

Please refer to this work as follows:

Slabbinck, B. (2009). From learning taxonomies to phylogenetic learning: a computational approach to FAME-based bacterial species identification. PhD thesis, Ghent University, Ghent, Belgium.

© Photograph on cover by Tom Couckuyt.
Gjendesheim, Jotunheimen, Norway.

ISBN-number: 978-90-5989-339-9

This research was funded by the Federal Science Policy (BELSPO), projects C3/00/12 and IAP VI-PAI VI/06.

The author and the supervisors give the authorisation to consult and to copy parts of this work for personal use only. Every other use is subject to the copyright laws. Permission to reproduce any material contained in this work should be obtained from the author.

Acknowledgement

Dankwoord

There and back again. A research tale...

Om met de woorden van J.R.R. Tolkien in huis te vallen. Het is een verhaal geworden van tien jaar Universiteit Gent. Een echt genre kan ik er niet op plakken maar het omvat in elk geval een stukje spanning en avontuur, een beetje thriller, ietwat knutsel- en schilderwerk en zelfs enige fantasy. Gedurende dit schrijven zijn veel herinneringen terug even naar boven gekomen. Niet alleen van vier jaar doctoraat maar ook van tijdens m'n studies bio-ingenieur en toegepaste informatica. Laat ik in dit eindwerk, letterlijk en figuurlijk, dan ook maar handig gebruik maken om een kort dankwoord te richten tot de vele mensen die me op weg hebben geholpen en die een steun hebben betekend in deze jaren.

2004, afgestudeerd als bio-ingenieur cel- en genbiotechnologie met een thesis die m'n zin in bioinformatica serieus had aangewakkerd. Met nog een extra jaartje toegepaste informatica op zak ging ik vol goede moed de wereld in op zoek naar een job in de bioinformatica. Dat was echter niet zo vanzelfsprekend als initieel gedacht. Ik bleek al snel een bredere kennis nodig te hebben om in deze branche terecht te kunnen, lees: een doctoraat. Het was niet zo simpel vier jaar geleden om een doctoraatspositie te vinden en al zeker niet als je je enkel en alleen toelegt op bioinformatica en moleculaire biologie. Al snel vond ik in m'n 'heimat' ofte boerekot een vacature voor data mining in de microbiologie. Officieel was de titel als volgt: *Dynamische geautomatiseerde identificatiesystemen voor Prokaryoten door integratie van databeheer en geavanceerde computationele strategieën*. Klonk goed, al moet ik eerlijk toegeven dat de vermelding dat één van de promotoren Prof. Bernard De Baets was toch voor enige twijfels zorgde. Laat ik met de glimlach maar gewoon stellen dat de cursus probabiliteit in 2e kandidatuur bio-ingenieur niet echt een populair vak was. Geen nood Bernard, alle twijfels waren snel van tafel geveegd! Uiteindelijk werden het drie promotoren. Bernard, Paul en Peter, bedankt om me de kans te geven om aan onderzoek te doen en om de nodige geldstromen te vinden om me vier jaar te laten doctoreren. Zoals elke doctoraatstudent het zal beamen, het was een periode met vallen en opstaan. De wereld van machine learning was mij onbekend en leren doe je pas door fouten te maken en ook effes op je bakkie te gaan. Bernard, bedankt voor het vele geduld, het vele verbeterwerk (ik mag hopen dat m'n Engels schrijven toch enigszins verbeterd is), het constante enthousiasme die je aan de dag legt en om me toch wat m'n eigen weg te laten gaan. In m'n start- en eindjaar is het voor jou niet altijd even gemakkelijk geweest om je met ons, doctoraatstudenten, bezig te houden maar toch bleef je zwoegen op alle details met een ongelofelijke passie voor wetenschap. Ik apprecieer ten zeerste alle werk en tijd die je in mijn onderzoek gestoken heb. Peter, ik bewonder je drang naar nieuwe kennis en je onuitputtelijke drive. Bedankt voor de hulp en steun die je me gegeven hebt deze vier jaar. Paul, wij hebben misschien in mindere mate met elkaar samengewerkt maar ik wil je bedanken voor je hulp, om de grote hoeveelheid data te laten genereren en ik hoop dat dit 'neural network research' je overtuigd heeft van de mogelijkheden en kwaliteiten van deze wetenschapstak voor verdere

integratie in de microbiologie.

Drie promotoren dus. Ondanks dat het niet altijd even soepel verliep, vind ik het een groot voordeel om onderzoek te mogen doen en om deel te kunnen uitmaken van twee totaal verschillende onderzoeksgroepen. In de ene groep bio-ingenieurs, wiskundigen en informatici, in de andere groep microbiologen en biotechnologen. En het was eigenlijk best grappig hoe beide groepen elkaar aankijken. Machine learning en statistische modellen zijn voor de microbiologen fancy, veel te ingewikkeld en een ver-van-m'n-bed-show. Voor de ingenieurs zijn de biologische data daarentegen verschrikkelijk aantrekkelijk om even een modelletje op te testen of er een algoritme op los te laten. Je ziet dus: een groot voordeel (lacht). Nu, het grootste deel van m'n tijd heb ik vertoefd op het boerekot of op 'de coupure' in de onderzoeksgroep Knowledge-based Systems. Eén persoon wil ik hier speciaal bedanken. Willem, bedankt voor alle nuttige suggesties en verbeteringen die je hebt aangebracht in m'n onderzoek. We hebben veel discussies gehad maar dankzij deze samenwerking heb ik m'n onderzoek kunnen bijsturen en verbeteren. Ik moest meermaals uitleggen hoe de data nu weer in elkaar zat en hoe ik alles in elkaar gestoken had, maar jij moest ook verschillende keren diverse machine learning aspecten uitleggen. Ik denk dat we elkaar goed aanvoelden en ik hoop dat we nog vaak een pint gaan pakken, een looptje gaan doen, en dat de lang verwachte ardennentrekking toch nog van de grond gaat komen! De donderdag en/of vrijdag was ik meestal te vinden op de Ledeganck in het labo Microbiologie. Ook hier moet ik enkele personen speciaal bedanken. Een machine learner is niets zonder data! An en Liesbeth, bedankt voor al het werk dat het genereren van de vetzuren met zich meebracht. Jullie zeiden misschien 'jis do were wi met z'n vetzuurtjes' maar wees ervan overtuigd dat zonder deze data niet de nodige resultaten tot stand zouden zijn gekomen! Ik wil ook alle andere onderzoekers en laboranten bedanken die in dit project werden ingeschakeld om me van de nodige data te voorzien: Cindy, Jeroen H., Emly, Jeroen, Caroline en Stefanie. With data generation comes data sharing. Working on the data of other researchers is challenging and results in a positive research drive! Therefore, I also want to thank Johannes Sikorski for sharing some of his data and for his interest in this project. I greatly appreciate your contribution to my research!

Vier jaar werken in twee onderzoeksgroepen betekent ook dat je met veel mensen in contact komt. Bij deze wil ik dan ook m'n KERMIT-bureauleden, Michael, Willem, Karolien, Hilde en Ester, bedanken voor de leuke sfeer, discussies en toffe momenten. Bij een doctoraat komen ook congressen aan bod en op deze gelegenheden leer je je collega's ook beter kennen. Verschillende momenten zijn me bijgebleven en ik moet nog steeds hartelijk lachen met de avondklim in de bossen van Spa waar Willem al vloekend z'n trolley naar boven sleurt. Ook op de ledeganck wil ik een aantal mensen speciaal bedanken. Bart en Tom, als '(bio)informatici' van het eerste uur hebben we toch heel wat afgelachen. Het was voor mij iedere week uitkijken naar onze uren samen in het bureautje op't vierde. Tom, met jou was het altijd een stek alhier en aldaar, samen de vrijdag uitwerken met wat techno op de achtergrond, Cercle vs. Buffalo ... Ik hoop dat we samen nog eens een pint gaan pakken! Wim G., bedankt voor de vele efforts om FAME-bank van de grond te krijgen. Zonder jou was m'n laatste hoofdstuk er helemaal niet gekomen! Doctoreren voor mij was ook meer dan onderzoek alleen. Het Biomatch minivoetbalteam was een tweewekelijks op de adem trappen en een leuke 'derden tiem'. Bert, Brecht, Jan VB, Jo, Peter, Sem, Timpe en anderen, bedankt voor de leuke sportieve momenten. Ik moet toegeven dat ik waarschijnlijk niet altijd de makkelijkste team-mate was. Ik ben nu eenmaal een winnerstype en stil toekijken op de bank is me nu eenmaal niet gegeven. In de eerste jaren was er ook de

Thank God It's Friday-drink met het labo Microbiologie in de Monopole, De Tobbe of Speakers Corner, wat een leuke afsluiter was van de week maar ook goede start was van het weekend. An, Katrien, Lies en andere genietters, merci voor de leuke momenten en, ja je kan het raden, de vele roddels op café. Ik hoop dat we aan die momenten nog een paar keer een vervolg kunnen geven. An en Katrien, Sevilla is om niet te vergeten. Tapas, spaghetti flamenco, zomerse temperaturen in november, veel lachen en, om niet te vergeten, een toren die nog steeds roept om beklommen te worden. Ik moet gewoon nog terug zeker? Om finaal het plaatje compleet te maken, wil ik iedereen van de vakgroep Biomath en het Labo Microbiologie bedanken voor de leuke tijd die ik heb mogen doorbrengen in beide groepen. Een speciale dank voor Viv en Annemie voor de administratieve rompslomp en de hulp om daar een weg in te vinden.

A journey is best measured in friends rather than miles

In 28 jaar bouw je vriendschappen op en het is belangrijk om daar op te kunnen terugvallen. Brecht en Eva, Hannelore, Jeroen en Anneleen, Stijn en Jasmijn, Lieze en Nico, David en Jasmien, Dries en Shirley, Klaas en Hanne, en Tim en Vanessa bedankt voor alle steun en vriendschap, een pint op café, een etentje, een quiz of simpelweg wat aflachen. Elke twee weken is het ook uitkijken naar de thuismatch van Cercle, niet enkel voor het voetbal maar ook omwille van de vele mensen die je hebt leren kennen en die dezelfde passie delen. Tuur, Ruiz, en de andere mannen en vrouwen van 'ons bende', Wouter DV, familie Dhaenekint en veel andere groenzwarte zielen. 'En we zien de manne van de Cercle Brugge...', merci! Eén van de voordelen van werken aan de universiteit is de mogelijkheid om een ietwat langere en/of verdere reis te kunnen ondernemen. Noorwegen, Canada, New York en Zweden zijn om niet te vergeten. In Noorwegen was het magnifiek en een lucky slag! Een prachtland doorkruisen met een machtige groep: Tom C., Mieke en Katrien C., bedankt voor de mooie en onvergetelijke momenten en ik hoop dat we nog een paar keer al lachend die toffe herinneringen kunnen ophalen.

Dit dankwoord kan en wil ik niet afsluiten zonder m'n ouders hierin te betrekken. Bedankt voor alle steun, kansen en vrijheid die je me altijd gegeven hebt, me nog steeds geeft en gewoon om er altijd te zijn en klaar te staan. Zonder jullie kon ik niet bereiken waar ik nu sta! Woorden zijn hier te beperkt om me goed te kunnen uitdrukken. M'n broers, m'n grootouders en m'n familie, bedankt voor alle steun en om er simpelweg te zijn. En last but not least, Katrien, samen leven, samen werken, samen Sevilla, New York en Zweden veroveren, samen iets gaan eten en drinken, en een dikke knuffel. We doen dat goed samen! Het laatste jaar was echter ook een tijd van samen een doctoraat schrijven, wat niet altijd even gemakkelijk was. Laten we die tijd nu maar snel inhalen!

Bram Slabbinck

04/12/2009

Contents

Acknowledgement - Dankwoord	v
List of Figures	xiii
List of Tables	xxiii
List of Abbreviations	xxvii
Preface	1
PART I General Introduction	
1 Machine Learning	7
1.1 Introduction	7
1.1.1 General Definitions and Concepts	7
1.1.2 Classification Settings	11
1.1.3 Balanced versus Imbalanced Data Sets	12
1.2 Machine Learning Techniques	13
1.2.1 Artificial Neural Networks	14
1.2.2 Support Vector Machines	26
1.2.3 Random Forests	33
1.3 Model Evaluation	42
1.3.1 Confusion Matrix	43
1.3.2 ROC Curve	45
1.3.3 Wilcoxon Rank-Sum Statistic	46
2 Bacteriology	49
2.1 Introduction	49
2.2 A Taxonomic View on the World of the Bacteria	50
2.2.1 Introduction	50
2.2.2 Bacterial Identification	56
2.2.3 A General Focus on the Genus	57
2.2.4 Where Machine Learning Meets Bacteriology	63
2.3 Bacterial FAME Analysis	65

2.3.1	Towards FAME Profiling	66
2.3.2	FAME Analysis of Species in the Genera <i>Bacillus</i> , <i>Paenibacillus</i> and <i>Pseudomonas</i>	77

PART II Data Mining and Machine Learning

3	Data Analysis	83
3.1	Introduction	83
3.2	Data Selection	83
3.3	Data Analysis and Visualization	89
3.3.1	Average FAME Profile	89
3.3.2	Clustering	93
3.3.3	TaxonGap	97
3.3.4	Principal Component Analysis	101
3.4	Conclusion	107
4	FAME-based Bacterial Species Classification	109
4.1	Introduction	109
4.2	<i>Bacillus</i> species classification: an ANN Approach	110
4.2.1	Methodologies	110
4.2.2	Results and Discussion	112
4.2.3	Publication	117
4.3	Three Genera - Three Techniques	117
4.3.1	Methodologies	117
4.3.2	Experimental Design	118
4.3.3	Results	122
4.3.4	Discussion	124
4.3.5	Comparison with Sherlock MIS	129
4.3.6	Independent Test Sets	131
4.3.7	The Plant-pathogenic <i>Pseudomonas</i> Species	134
4.3.8	Publication	136
4.4	Conclusions	136
5	Phylogenetic Learning	139
5.1	Introduction	139
5.2	From Learning Taxonomies to Phylogenetic Learning	140
5.2.1	Methodologies	140
5.2.2	Results and Discussion	147
5.2.3	Publication	161
5.3	Putting Bacterial Species Identification into Context	161
5.3.1	Statistically Significant Nodes	161
5.3.2	Visualization and Evaluation	162

5.4	Conclusions	170
-----	-----------------------	-----

PART III Online Database

6	FAME-bank.net	175
6.1	Introduction	175
6.2	Construction and Content	177
6.2.1	Database Schema and Implementation	177
6.2.2	Data Sources and Quality Control	178
6.2.3	Web Interface	179
6.3	Utility and Discussion	179
6.3.1	User Interface	179
6.3.2	Use, Benefits and Future Development	180
6.4	Conclusions	183
6.5	Availability	184

PART IV General Summary

7	General Conclusions	187
7.1	Data Sets	187
7.2	Data Analysis and Machine Learning	188
7.2.1	Data Analysis	188
7.2.2	FAME-based Bacterial Species Classification	189
7.2.3	Phylogenetic Learning	191
7.3	FAME-bank.net	193
7.4	Main Contributions to the Community	194
8	Future Perspectives	197
8.1	General	197
8.2	Data Sets	198
8.3	Data Analysis	198
8.4	Phylogenetic Learning	199
8.5	FAME-bank.net	200

PART V Appendices

A	Data Sets	203
A.1	Strain Tables	203
A.1.1	Data set 2006	203
A.1.2	Data sets 2008	207
A.2	Average FAME Profiles: Major Constituents	217
A.3	PCA Biplots	227

B The Sherlock MIS	231
B.1 TSBA50 Peak Naming Table	231
B.2 TSBA50 Identification Library	234
C 16S rRNA Gene Sequences	237
Bibliography	243
Summary	259
Samenvatting	263
Curriculum Vitae	269

List of Figures

1	Dissertation road map. The different chapters are marked by their chapter number and title.	4
1.1	Generalization. Simple illustration of the errors on the training (red) and validation (green) set for varying parameter settings. In this curve, the error on both data sets decreases until a certain parameter value, denoted by the dashed line. Beyond this line, the error on the training data further decreases while the error on the validation set increases. In other words, the model learns the data values and starts to overfit. The generalization power of the model decreases and is given by the increased error on the validation set. The final model is constructed by the parameter setting corresponding to the minimum validation error.	9
1.2	k-fold cross-validation. The initial data set is split into k equal parts, with $k = 4$ in this figure. Each part is once used for validation (green) while the other $k - 1$ parts are used for training (red). The final model performance equals the average performance over the k folds.	10
1.3	Structure of biological neuron and signal transmission. Signals are retrieved by the multiple dendrites and transported to the soma. If the sum of all signals exceeds a certain threshold, the signal is further transported to the axon for transmission to the other neurons (McMurry and Castellion, 2002).	15
1.4	Mathematical representation of the perceptron. The input values of training point \mathbf{x}_i correspond to x_{id} , with $d = 0, \dots, D$ and D the number of features. Input value x_0 is called the bias neuron and its value is set to 1. To each input is associated a weight w_d . The weighted sum of the input values is calculated and the resulting sum is converted into a specific output value y as defined by the bipolar step function. A threshold on the weighted input signal is set equal to zero (adapted from Mitchell (1997)).	15
1.5	Linear decision boundary. Linear decision boundary corresponding to $f(\mathbf{x}_i) = 0$ or $\sum_{d=1}^D w_d x_{id} = -w_0$ in a two-dimensional input space. The weight vector \mathbf{w} defines the orientation of the boundary, while the bias weight w_0 defines the position of the boundary in terms of its perpendicular distance from the origin. The boundary separates the green data points (positive class) from the red data points (negative class).	16

- 1.6 **Architecture of a fully connected multi-layer artificial neural network.** The first layer of the network consists of the input data values x_{id} with $d = 1, \dots, D$ and D the number of features (the bias not considered). The input data is forwarded over weighted neuron connections to the hidden neurons h_m with $m = 1, \dots, M$, M the optimal number of hidden neurons and w_{dm} the weights on the corresponding connections. A weighted input-output mapping is performed at the hidden neurons and the resulting values are forwarded to the output neurons p_k over connections with weights w_{mk} , with $k = 1, \dots, K$. K output neurons are considered with K the number of classes in the data set in case of classification and $K = 1$ in the case of regression. Again, a weighted input-output mapping is performed, resulting in the output values o_k 19
- 1.7 **Maximum margin classifier.** The linear decision boundary corresponding to $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ is visualized in a two-dimensional input space. Parallel to linear decision boundary, the two optimal boundaries are set as dashed lines and the distance $2M$ between both is called the margin. Data points lying on the dashed lines are called the support vectors and are encircled in black. The weight vector \mathbf{w} defines the orientation of the boundary, while the bias weight b defines the position of the boundary in terms of its perpendicular distance from the origin. The boundary separates the green data points (positive class) from the red data points (negative class). On the left, a linearly separable problem is visualized and the margin is denoted as ‘hard margin’. On the right, a non-linear separable problem is visualized. One point of each class is beyond its boundary and corresponds to slack variable ξ . By allowing but penalizing misclassifications in this case, the margin is denoted as ‘soft margin’. 27
- 1.8 **Concept of feature mapping.** The function ϕ maps the non-linear separable input data into a feature space \mathcal{F} that allows linear separability. 30
- 1.9 **Decision tree in a setting with five classes A, B, C, D and E and two features x_1 and x_2 .** Four thresholds θ_r are used as splitting criteria with $r = 1, \dots, 4$. The decision tree is visualized both in input space (left figure) and in feature space (right figure) (Bishop, 2006) 36
- 1.10 **ROC curve.** Example of a ROC curve for a two-class classification experiment. On the X axis, the false positive rate (FPR) is given, while the Y axis represents the true positive rate (TPR). The area under the curve is an estimation of the classification performance and equals here 0.965. 46
- 2.1 **Trend in novel valid species descriptions.** Number of valid bacterial species descriptions since 1980. For 2009, the number of validly published species is given as counted on 03/11/2009. Data captured from Euzéby (1997). 54

- 2.2 **The major bacterial phyla.** 16S rRNA gene-based phylogenetic parsimony tree showing the major bacterial phyla. The triangles indicate groups of related organisms, while the angle at the root of the group roughly indicates the number of sequences available and the edges represent the shortest and longest branch within the group. All available almost complete homologous sequences from *Archaea* and *Eukarya* were used as outgroup references to root the tree, indicated by the arrow (Ludwig and Klenk, 2001). In this study, we focused on two genera of the red phyla of the Firmicutes and on one genus of the green phyla of the Proteobacteria. 55
- 2.3 **Monthly nomenclatural changes in the bacterial taxonomy of the genus *Bacillus* as published by the IJSEM between January 2006 and November 2009.** The number of novel described species is given, together with the number of species renamed within the genus *Bacillus* as the number of valid *Bacillus* species renamed to a new species outside the genus. Information extracted from the List of Prokaryotic Names with Standing in Nomenclature (Euzéby, 1997). 57
- 2.4 **16S rRNA gene sequence-based maximum likelihood tree of the genus *Bacillus*.** The 147 species included correspond to the validly published taxonomy of May 2008. One high-quality 16S rRNA sequence per species is selected 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 (Stamatakis, 2006; Letunic and Bork, 2007). The tree branch lengths are ignored because of outlier species that make the use of branch length pointless. The green group corresponds to the *Bacillus subtilis* group, while the blue group corresponds to the *Bacillus cereus* group. Dotted branches correspond to bootstrap values larger than 75%. 59
- 2.5 **Monthly nomenclatural changes in the bacterial taxonomy of the genus *Paenibacillus* as published by the IJSEM between January 2006 and November 2009.** The number of novel described species is given, together with the number of valid *Paenibacillus* species renamed to a new species outside the genus. Information extracted from the List of Prokaryotic Names with Standing in Nomenclature (Euzéby, 1997). 60
- 2.6 **16S rRNA gene sequence-based maximum likelihood tree of the genus *Paenibacillus*.** The 101 species included correspond to the validly published taxonomy of May 2008. One high-quality 16S rRNA sequence per species is selected 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 (Stamatakis, 2006; Letunic and Bork, 2007). Dotted branches correspond to bootstrap values larger than 75%. 61

2.7	Monthly nomenclatural changes in the bacterial taxonomy of the genus <i>Pseudomonas</i> as published by the IJSEM between January 2006 and November 2009. The number of novel described species is given, together with the number of species renamed within the genus <i>Pseudomonas</i> as the number of valid <i>Pseudomonas</i> species renamed to a new species outside the genus. Information extracted from the List of Prokaryotic Names with Standing in Nomenclature (Euzéby, 1997).	62
2.8	16S rRNA gene sequence-based maximum likelihood tree of the genus <i>Pseudomonas</i>. The 117 species included correspond to the validly published taxonomy of May 2008. One high-quality 16S rRNA sequence per species is selected 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 (Stamatakis, 2006; Letunic and Bork, 2007). Tree branch lengths are ignored because of outlier species, that make the use of branch lengths pointless, and dotted branches correspond to bootstrap values larger than 75%.	64
2.9	Nomenclature of fatty acids (Dawyndt, 2004).	68
2.10	The Sherlock Microbial Identification System workflow (Sasser, 1990).	71
2.11	Sherlock MIS example report. Report of the analysis of the whole-cell FAME composition of <i>Bacillus subtilis</i> LMG7135 ^T . The report consists of three main parts: the chromatogram with the different FAME peaks, a report with detailed peak information and an identification report. In this report, only one entry was given for the identification of <i>Bacillus subtilis</i> LMG7135 ^T	74
2.12	Trend in whole-cell FAME analysis. Number of whole-cell FAME profiles generated at the Laboratory of Microbiology and BCCM TM /LMG Bacteria Collection (Ghent University, Belgium) from the start in 1989 to 21/11/2009. The total number of FAME profiles equals 71,267.	78
3.1	Number FAME profiles per <i>Bacillus</i> and <i>Paenibacillus</i> species.	90
3.2	Number FAME profiles per <i>Pseudomonas</i> species.	91
3.3	Average of peak percentages in the genus <i>Bacillus</i>	94
3.4	Peak distribution in average species profiles the genus <i>Bacillus</i>	94
3.5	Average of peak percentages in the genus <i>Paenibacillus</i>	95
3.6	Peak distribution in average species profiles the genus <i>Paenibacillus</i>	95
3.7	Average of peak percentages in the genus <i>Pseudomonas</i>	96
3.8	Peak distribution in average species profiles of the genus <i>Pseudomonas</i>	96

- 3.9 **Heatmap of data with clustering of the genus *Bacillus* data set.** The left figure illustrates only peak clustering, while the right plot illustrates peak and profile clustering. Clustering is based on the Canberra metric. Rows correspond to the different FAME profiles, which are alphabetically ordered and coloured by species name (presented at the left of each heatmap). Columns correspond to the different FAME peaks. Peak percentages are coloured from green to red with an increasing value, as represented by the colour key in the top-left corner. 98
- 3.10 **Heatmap of data with clustering of the genus *Paenibacillus* data set.** The left figure illustrates only peak clustering, while the right plot illustrates peak and profile clustering. Clustering is based on the Canberra metric. Rows correspond to the different FAME profiles, which are alphabetically ordered and coloured by species name (represented in the colour bar at the left of each heatmap). Columns correspond to the different FAME peaks. Peak percentages are coloured from green to red with an increasing value, as represented by the colour key in the top-left corner. 98
- 3.11 **Heatmap of data with clustering of the genus *Pseudomonas* data set.** The left figure illustrates only peak clustering, while the right plot illustrates peak and profile clustering. Clustering is based on the Canberra metric. Rows correspond to the different FAME profiles, which are alphabetically ordered and coloured by species name (represented in the colour bar at the left of each heatmap). Columns correspond to the different FAME peaks. Peak percentages are coloured from green to red with an increasing value, as represented by the colour key in the top-left corner. 98
- 3.12 **Intra- and inter-species discrimination by FAME in the genus *Bacillus*.** Visualization generated by TaxonGap 2.4.1. The axis denotes the percentage of heterogeneity within a species and the percentage of separability from other species. Dark grey bars denote the minimum separability from the other species, while the light grey bars denote the maximum species heterogeneity. The arrow indicates a line corresponding to the minimum separability over all species. . . . 102
- 3.13 **Intra- and inter-species discrimination by FAME in the genus *Paenibacillus*.** Visualization generated by TaxonGap 2.4.1. The axis denotes the percentage of heterogeneity within a species and of the percentage separability from other species. Dark grey bars denote the minimum separability from the other species, while the light grey bars denote the maximum species heterogeneity. The arrow indicates a line corresponding to the minimum separability over all species. 103
- 3.14 **Intra- and inter-species discrimination by FAME in the genus *Pseudomonas*.** Visualization generated by TaxonGap 2.4.1. The axis denotes the percentage of heterogeneity within a species and of the percentage separability from other species. Dark grey bars denote the minimum separability from the other species, while the light grey bars denote the maximum species heterogeneity. The arrow indicates a line corresponding to the minimum separability over all species. . . . 104

- 3.15 **Principal component analysis of the three genera *Bacillus*, *Paenibacillus* and *Pseudomonas*.** Skree plots are given for the data set of each genus, with only the ten first principal components. 105
- 3.16 **Principal component analysis of the genus data set.** A biplot is visualized of the first two principal components. 106
- 3.17 **Principal components analysis of the plant-pathogenic *Pseudomonas* data set.** A biplot is visualized of the first two principal components. The species belonging to the *P. beteli* group and *P. flectens* are not included in the plot. . . . 107
- 3.18 **Principal components analysis of the 2008 *Pseudomonas* data set with species labeled as being plant-pathogenic or not.** A biplot is visualized of the first two principal components. Red points correspond to non-plant-pathogenic species, while green points correspond to plant-pathogenic species. 108
- 4.1 **An example of optimization of the number of hidden neurons.** Optimization was done by stratified cross-validation. Mean squared error (MSE) values are plotted for different numbers of hidden neurons (step size of 5). The final number of hidden neurons is pointed by the arrow. 111
- 4.2 **Mean overall area under the ROC curve (AUC) and standard deviation of each experiment type for each activation combination.** 113
- 4.3 **Mean true positive (TP) percentages for each experiment type and for each activation combination.** Percentages are given for the identification of the correct species name as highest output score (First), as present in the five highest output scores (Five Best) and as highest output score when considering the *B. cereus* and *B. subtilis* species groups (Group). An experiment type consists of a balance type (bal/imbal) and a validation type (val/cv). 116
- 4.4 **Schematic presentation of the experimental design. 1A/B. Stratified identification strategy.** 1A. Genus identification is performed by the genus identification model. This model relies on the complete FAME data set in which the profiles are annotated by genus name (dark grey box). 1B. For each genus, a species identification model is built based on the FAME profiles corresponding to that specific genus. The respective FAME profiles are annotated by species name (light grey boxes). In both cases, each profile is labelled with the genus or species name associated with the highest output value. However, species identification is only performed for the genus associated with the highest output value following genus identification. 2. **Straight species identification strategy.** The complete data set of FAME profiles is annotated by genus and species name (dark grey box). Identification is performed by a single identification model. Each profile is labelled with the genus and species name associated with the highest output value. 119

- 4.5 **Average ROC.** Average ROC plots of the ANN, RF, SVM with RBF and linear (lin) kernel and BSVM with RBF and linear kernel classification experiments for each genus, the genera classification experiment and the straight identification experiment. Each time, the experiment with the highest AUC was selected. ROC values were calculated for each class separately and vertically averaged over the classes. 121
- 4.6 **Comparison of the identification performance of random forests and Sherlock MIS in a stratified setting.** Blue bars correspond to random forests and red bars with Sherlock MIS. In the left diagram, evaluation is performed by the test set with averaging of the identification results over each species and globally over all species. In the right diagram, evaluation is performed by randomly sampling ten subsets consisting of one profile for each species from the test set and by calculating an average performance over each subset. In this setting, the performance of each subset is calculated by the percentage of correct identifications. The bottom and upper bar respectively denote the 25% and 75% percentile of the identification results. 131
- 5.1 **FAME tree.** Phylogenetic tree resulting from the divisive clustering of the FAME data of 15 *Bacillus* species based on classification by Random Forests. Clustering is based on AUC and average linkage of the probability estimates calculated from identification by Random Forests. At the different nodes the corresponding AUC value is reported. 149
- 5.2 ***Bacillus* 16S rRNA gene neighbour-joining tree as constructed by PHYLIP 3.68 and based on sequences selected from the SILVA database.** Only the species present in the original 2008 data set are visualized. The tree was visualized using the iTol webtool (Letunic and Bork, 2007). The *Bacillus cereus* and *Bacillus subtilis* groups are coloured in blue and green, respectively. 151
- 5.3 ***Bacillus* 16S rRNA gene UPGMA tree as constructed by PHYLIP 3.68 and based on sequences selected from the SILVA database.** Only the species present in the original 2008 data set are visualized. The tree was visualized using the iTol webtool (Letunic and Bork, 2007). The *Bacillus cereus* and *Bacillus subtilis* groups are coloured in blue and green, respectively. 152
- 5.4 **Sensitivity and F-score values by phylogenetic learning based on a 16S rRNA gene NJ tree.** For each *Bacillus* species, the corresponding sensitivity and F-score value of phylogenetic learning based on a 16S rRNA gene NJ tree is displayed. Sensitivity is denoted by the light blue bars, F-score by the green bars. The tree is visualized using the iTol webtool (Letunic and Bork, 2007). The *Bacillus cereus* and *Bacillus subtilis* groups are coloured in blue and green, respectively. 155

- 5.5 **Sensitivity and F-score values by phylogenetic learning based on a 16S rRNA gene UPGMA tree.** For each *Bacillus* species, the corresponding sensitivity and F-score value of phylogenetic learning based on a 16S rRNA gene UPGMA tree is displayed. Sensitivity is denoted by the light blue bars, F-score by the green bars. The tree is visualized using the iTol webtool (Letunic and Bork, 2007). The *Bacillus cereus* and *Bacillus subtilis* groups are coloured in blue and green, respectively. 156
- 5.6 **Sensitivity and F-score values for flat multi-class classification.** For each *Bacillus* species, the corresponding sensitivity and F-score value for flat multi-class classification is displayed along the 16S rRNA gene NJ tree. Sensitivity is denoted by the light blue bars, F-score by the green bars. The tree is visualized using the iTol webtool (Letunic and Bork, 2007). The *Bacillus cereus* and *Bacillus subtilis* groups are coloured in blue and green, respectively. 157
- 5.7 **Performance comparison at class level.** For each class, sensitivity and F-score values resulting from phylogenetic learning based on a 16S rRNA gene NJ or UPGMA tree were compared to those obtained by flat multi-class classification. Four plots are given. The X-axis corresponds to thresholds set on the corresponding metric values. Threshold steps of 0.01 were chosen. For each threshold, flat multi-class classification was evaluated at class level and those classes with metric values smaller than or equal to the threshold were selected. Classification performance by phylogenetic learning was analyzed at class level for each set of classes. The Y-axis on the left projects the number of phylogenetic learning classes that had a higher metric value than those obtained by flat multi-class classification. The red line expresses this number, relative to the size of the corresponding set (Y-axis on the right). 158
- 5.8 **Average misclassification depth of phylogenetic learning based on a NJ tree.** The average depth of the misclassified test points of each class is visualized for phylogenetic learning based on an NJ tree. Depth equals the number of nodes along the classification path until misclassification occurs (the corresponding node also included) and corresponds to the green bars. The maximum or correct depth is shown by the red bars. Maximum depth equals the number of nodes along the true phylogenetic path (final leaf included). 159
- 5.9 **Average misclassification depth of phylogenetic learning based on a UPGMA tree.** The average depth of the misclassified test points of each class are visualized for phylogenetic learning based on a UPGMA tree. Depth equals the number of nodes along the classification path until misclassification occurs (the corresponding node included) and corresponds to the green bars. The maximum or correct depth is shown by the red bars. Maximum depth equals the number of nodes along the true phylogenetic path (the final leaf included). 160

- 5.10 **Statistical evaluation of phylogenetic learning for the genus *Bacillus* with a 16S rRNA gene NJ phylogenetic tree.** Phylogenetic learning was performed with the 2008 *Bacillus* data set. The number of FAME profiles of each species is reported following the species name. For each node, the number of FAME profiles of the subtree (top value) is given, together with the p -values (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval [0.01*,0.05*] and green colour to p -values below 0.01*. 164
- 5.11 **Statistical evaluation of phylogenetic learning for the genus *Bacillus* with a 16S rRNA gene UPGMA phylogenetic tree.** Phylogenetic learning was performed with the 2008 *Bacillus* data set. The number of FAME profiles of each species is reported following the species name. For each node, the number of FAME profiles of the subtree (top value) is given, together with the p -value (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval [0.01*,0.05*] and green colour to p -values below 0.01*. 165
- 5.12 **Statistical evaluation of phylogenetic learning for the genus *Paenibacillus* with a 16S rRNA gene NJ phylogenetic tree.** Phylogenetic learning was performed with the 2008 *Paenibacillus* data set. The number of FAME profiles of each species is reported following the species name and for each node the number of FAME profiles of the subtree (top value) is given, together with the p -value (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval [0.01*,0.05*] and green colour to p -values below 0.01*. 166
- 5.13 **Statistical evaluation of phylogenetic learning for the genus *Paenibacillus* with a 16S rRNA gene UPGMA phylogenetic tree.** Phylogenetic learning was performed with the 2008 *Paenibacillus* data set. The number of FAME profiles of each species is reported following the species name. For each node, the number of FAME profiles of the subtree (top value) is given, together with the p -value (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval [0.01*,0.05*] and green colour to p -values below 0.01*. 167

- 5.14 Statistical evaluation of phylogenetic learning for the genus *Pseudomonas* with a 16S rRNA gene NJ phylogenetic tree.** Phylogenetic learning was performed with the 2008 *Pseudomonas* data set. The number of FAME profiles of each species is reported following the species name. For each node, the number of FAME profiles of the subtree (top value) is given, together with the p -value (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval [0.01*,0.05*] and green colour to p -values below 0.01*. 168
- 5.15 Statistical evaluation of phylogenetic learning for the genus *Pseudomonas* with a 16S rRNA gene UPGMA phylogenetic tree.** Phylogenetic learning was performed with the 2008 *Pseudomonas* data set. The number of FAME profiles of each species is reported following the species name. For each node, the number of FAME profiles of the subtree (top value) is given, together with the p -value (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval [0.01*,0.05*] and green colour to p -values below 0.01*. 169
- 6.1 Entity-Relation Diagram of the FAME-bank database structure.** Rectangular and oval entities correspond to table names and attributes, respectively. Attributes are connected to tables by dotted lines. Diamond squares represent relations between the different entities. Full lines represent at least one relationship between two entries of both entities, while dashed lines correspond to a possible relationship. The primary key of each table is underlined. 179
- 6.2 FAME profile FB00000647 of *Bacillus subtilis* LMG 7135^T.** Screenshots are given for the different information tabs. 181
- A.1 Biplot of the first two principal components of PCA analysis of the *Bacillus* data set.** Species denoted by the character '×' belong the *Bacillus cereus* species group, while species denoted by a circle belong the *Bacillus subtilis* group. Besides those marks, single species that form a separable cluster are also marked, though by a different character. 228
- A.2 Biplot of the first two principal components of PCA analysis of the *Paenibacillus* data set.** Single species that cluster in a distinct group are marked by a different character for each species individually. 229
- A.3 Biplot of the first two principal components of PCA analysis of the *Pseudomonas* data set.** Species denoted by the character '×' belong the *Pseudomonas aeruginosa* species group, while species denoted by a circle belong the *Pseudomonas syringae* group and species denoted by a plus character belong to the *Pseudomonas beteli* outgroup. Besides those marks, single species that form a separable cluster are also marked, though by a different character. 230

List of Tables

1.1	Overview of commonly used ANN activation functions. The mathematical formula and representation are given for the step and bipolar (or symmetric) step function, the logistic sigmoid function and hyperbolic tangent sigmoid function.	20
1.2	Confusion matrix. Two-by-two matrix summarizing the predictions of a two-class classification.	43
1.3	Performance measures. Overview of several performance measures as calculated from a two-class confusion matrix.	44
1.4	Multi-class confusion matrix. K -by- K matrix summarizing the predictions of a multi-class classification, with K the number of classes. The cells on the main diagonal represent the number of TP while the non-diagonal cells correspond to the errors made.	44
3.1	Statistics of the generated data sets. For each data set year, type of data set, and number of corresponding species, strains, FAME profiles and FAME peaks are reported.	86
3.2	Overview of the considered plant and mushroom pathogenic <i>Pseudomonas</i> species. For each species, the number of FAME profiles, the host(s) and corresponding reference(s) are given.	88
4.1	Overview of the identification results of each experiment type with the activation combination leading to the highest mean AUC. Number of training, validation and test profiles, and the mean and standard deviation of the area under the ROC curve (AUC) are reported.	112
4.2	p-values of the Wilcoxon rank-sum test based on the mean overall AUC of each experiment type for each activation combination. Each activation combination corresponds to a triangle: b/b (top-left), b/s (top-right), s/s (bottom-left) and s/b (bottom-right). Significantly better identification performance is indicated by an asterisk ($p < 0.05$).	114
4.3	Effect of the activation type. Number of activation combinations for the four experiment types leading to significantly different AUC values based on the Wilcoxon rank-sum test ($p < 0.05$).	115

- 4.4 **Overview of the results for genus and species identification.** Classification performance is indicated by the Area Under the ROC Curve (AUC). Identification performance is indicated by sensitivity (Se), precision (Pr) and F-score (F). Results are reported for two identification (ID) strategies: stratified and straight identification. The stratified identification strategy performs identification at genus level and at species level for the three genera *Bacillus*, *Paenibacillus* and *Pseudomonas*. Three machine learning techniques are used: artificial neural networks (ANNs) with a sigmoid (s) and/or bipolar sigmoid (b) activation function on the hidden and output neurons ($f_{\text{hidden}}/f_{\text{output}}$), random forests (RFs) and support vector machines (SVMs) with RBF and linear (lin) kernel. Performance values and standard deviations are reported. Highest performance values are indicated in bold face. 120
- 4.5 **Multi-class confusion matrix resulting from genus identification by the best RF experiment.** The number of correct predictions are presented on the main diagonal, the other cell values show the number of incorrect predictions. Row labels correspond to the true genus names, column labels correspond to the predicted genus names. 122
- 4.6 **Identification of the J. Sikorski independent FAME data set.** The identification results of the best random forests experiment and of Sherlock MIS are given for 131 *Bacillus simplex* FAME profiles. The percentage of *B. simplex* identification is marked in bold. 132
- 4.7 **Results of *Pseudomonas* species identification by the 2008 *Pseudomonas* plant-pathogenic species data set.** Two machine learning techniques were evaluated: random forests (RFs) and support vector machines (SVMs). For support vector machines, the linear kernel (lin) and the RBF kernel were considered. Nested cross-validation was performed, with no inner cross-validation in the case of RFs. The AUC, sensitivity (Se), precision (Pr) and F-score (F) were calculated as an average over all classes in a one-versus-others settings. Standard deviations are also reported. 134
- 4.8 **Results of *Pseudomonas* identification of plant-pathogenic species in the 2008 *Pseudomonas* data set.** Two machine learning techniques were evaluated: random forests (RFs) and support vector machines (SVMs). For support vector machines, the linear (lin) and RBF kernel was considered. 10-fold nested cross-validation was performed without an inner cross-validation in the case of RFs. The AUC, sensitivity (Se), precision (Pr) and F-score (F) were calculated from the two-class identification results. 135

5.1	Results from the hierarchical single-label multi-class classification, phylogenetic learning and flat multi-class classification experiments. In this table, the three classification strategies are abbreviated as ‘HSMC’, ‘PhyLearn’ and ‘Multi-class’, respectively. The results of these three strategies are reported in the upper, middle and bottom part of the table, respectively. The results of hierarchical single-label multi-class classification were based on the FAME tree resulting from the divisive clustering experiment. Only the 15 species data set was considered and 3-fold and 11-fold stratified cross-validation (CV) was performed. In the case of phylogenetic learning, two 16S rRNA gene trees were used as template: neighbour-joining (NJ) and unweighted pair group method with arithmetic mean (UPGMA). For PhyLearn, both the 15 and the 74 species data set were considered and all PhyLearn experiments were performed using 3-fold stratified CV. Also the flat multi-class experiments were validated by this CV strategy. In the three strategies, classification performance was evaluated based on the pooled test set. Metrics reported are sensitivity, precision and F-score. Based on a multi-class confusion matrix, statistics were calculated in a one-versus-other setting with averaging of the corresponding statistic over the different classes. Standard deviations are reported between brackets. NaN denotes the number of classes that have resulted in a value ∞ (only in case of precision and F-score).	150
A.1	Strain table corresponding to the 2006 <i>Bacillus</i> data set. Strain numbers and corresponding number of included FAME profiles are reported. Also, exceptional growth and culturing conditions are reported (column ‘EC’).	203
A.2	Strain table corresponding to the 2008 <i>Bacillus</i>, <i>Paenibacillus</i> and <i>Pseudomonas</i> data sets. Strain numbers and corresponding number of included FAME profiles are reported. Also, exceptional growth and culturing conditions are reported (column ‘EC’). Regarding the <i>Pseudomonas</i> strains, the plant-pathogenic strain reallocation by Gardan et al. (1999) is followed and integrated in the table. For this genus, strains with one or more plant-pathogenic strains are denoted by superscript ‘p’.	207
A.3	Average FAME peaks in the genus <i>Bacillus</i>. Averages and standard deviations of the peaks with a prevalence in more than ten species.	218
A.4	Average FAME peaks in the genus <i>Paenibacillus</i>. Averages and standard deviations of the peaks with a prevalence in more than ten species.	220
A.5	Average FAME peaks in the genus <i>Pseudomonas</i>. Averages and standard deviations of the peaks with a prevalence in more than twenty species.	222
B.1	The FAME peaks named by the Sherlock MIS TSBA50 peak naming table.	231

- B.2 Entries of the Sherlock MIS TSBA50 identification library corresponding to the genus *Bacillus*, *Paenibacillus* and *Pseudomonas*.** Entries validly described as a species according to the taxonomy of 03/2008 are marked, otherwise the correct species name is given (only if species was validly described). Subspecies or infrasubspecific ranks are not considered. 234
- C.1 Strain list with according 16S rRNA gene accession numbers.** List of the species included in the 2008 data set with the type strain number and the accession number according to the selected 16S rRNA gene sequence. 237

List of Abbreviations

ACP	Acyl carrier protein
ANI	Average nucleotide identity
ANN	Artificial neural network
ARDRA	Amplified ribosomal DNA restriction analysis
AUC	Area under the ROC curve
b	Bipolar sigmoid activation function
<i>B.</i>	<i>Bacillus</i>
BAL	Balanced
CV	Cross-validation
DDAG	Decision directed acyclic graph
DDH	DNA-DNA hybridization
DNA	Deoxyribonucleic acid
EC	Exceptional condition(s)
ECL	Equivalent chain length
EPS	Enhanced postscript
F	F-score
FAME	Fatty acid methyl ester
FN	False negative
FP	False positive
FPR	False positive rate
FTIR	Fourier transform infrared spectroscopy
GC	Gas chromatography
ICNB	International Code of Nomenclature of Bacteria
ICSP	International Committee on Systematics of Prokaryotes
IJSEM	International Journal of Systematic and Evolutionary Microbiology
IMBAL	Imbalanced
INSDC	International Nucleotide Sequence Database Collaboration
LMG	Laboratory of Microbiology Ghent
LPS	Lipopolysaccharide
LSPN	List of prokaryotic names with standing in nomenclature
MIDI	Microbial ID Inc.
MIS	Microbial Identification System
ML	Maximum likelihood
MLSA	Multi-locus sequence analysis
MLST	Multi-locus sequence typing
MP	Maximum parsimony
MSE	Mean squared error

ODBC	Open database connectivity
OTU	Operational taxonomic unit
<i>P.</i>	<i>Pseudomonas</i>
<i>Pa.</i>	<i>Paenibacillus</i>
PC	Principal component
PCA	Principal components analysis
PR	Precision
RBF	Radial basis function
RPROP	Resilient propagation
RF	Random forest
ROC	Receiver operating characteristic
rRNA	Ribosomal ribonucleic acid
RT	Retention time
s	Sigmoid activation function
SE	Sensitivity
SI	Similarity index
STDEV	Standard deviation
SVM	Support vector machine
TN	True negative
TP	True positive
TPR	True positive rate
TSBA	Trypticase soy broth agar
VAL	Simple validation

Introduction

In 1989, Tim Berners-Lee proposed a data and information management system that was the basis for the creation of the World Wide Web we experience today in every day life. In twenty years time, data sharing has grown explosively. Networking has become integrated in every corner of our society and databases are among the main pillars it is relying on. When focusing on biology, online databases of every kind pop up rapidly in many different fields. Importantly, the main part of current research relies on stored and shared data. This is also true in the field of bacteriology, where online gene and protein sequence databases are an important basis for bacterial identification, comparison and analysis. Moreover, this genotypic information has also extensively changed bacterial taxonomy, that was originally based on phenotypic features such as morphology, presence of flagella, pathogenicity, etc. Nevertheless, phenotypic analysis methods are still routinely used because they are cheap, fast and possibly allow for an automated high-throughput analysis. It is clear that these advantages allow phenotypic methods to be used for first-line bacterial identification. However, one of the major drawbacks is that these methods mostly lack a good identification scheme up-to-date with the standing bacterial taxonomy. Nonetheless, in view of a polyphasic taxonomy, the description of a bacterial species asks for phenotypic data to confirm the findings based on genotypic information. It is clear that the phenotype was, is and will remain an important player in bacterial taxonomy.

In this dissertation, we focus on bacterial whole-cell fatty acid methyl ester (FAME) analysis. This is a phenotypic and chemotaxonomic typing method. Relying on gas chromatography, FAME analysis allows for an easy, cheap, automated and high-throughput typing of bacteria, which in combination with a particular identification library results in a rapid identification of bacterial isolates. Twenty years of bacterial FAME research at the Laboratory of Microbiology and the BCCMTM/LMG Bacteria Collection (Ghent University, Belgium) has led to a FAME database that currently consists of more than 71,000 whole-cell FAME profiles. It is clear that such a large database is a perfect environment for data mining and knowledge discovery. With this study, we lifted FAME analysis for bacterial identification to a higher level by an intelligent computational analysis using machine learning techniques.

Bacteriology and, more specifically, bacterial taxonomy is still a rich field to explore for bioinformaticians and computer scientists. The number of machine learning applications in this field is still very small and this is especially true with regard to FAME analysis. Where FAME analysis has already been performed and optimized for almost 50 years, computational FAME

data analysis is still limited. The main reason lies in the fact that, in contrast to genotypic data such as gene sequences, phenotypic data are not shared in an easy way to handle. Many of the data are published in papers, though are not easy to access for electronic data analysis. Moreover, data resulting from phenotypic analysis methods are also stored in private databases, making it difficult to perform any extensive computational study. By using the LMG FAME database as a starting point, we investigated how bacterial FAME-based species identification could be improved by the use of machine learning techniques and by focusing on three particular branches of the bacterial tree of life. More specifically, three bacterial genera were of main interest: *Bacillus*, *Paenibacillus* and *Pseudomonas*. The choice of these three genera was based on the extensive experience at the Laboratory of Microbiology on these genera and on the number of profiles available in the FAME database.

In this work, three main research steps have been investigated. As a first step, we analyzed the FAME data with standard data analysis methods. Hereby, we pursued to gain insight in the composition and the patterns of the data. With the goal of improving bacterial species identification by FAME data, it was investigated how the FAME patterns relate to each other at the species level. Especially in the framework of a machine learning study, data analysis is typically performed to determine how the data will confine the performance of machine learning techniques. As a second step, FAME-based bacterial species identification by three machine learning techniques was investigated. Different data sets, techniques and parameter settings were considered with the goal of improving species identification. Where bacterial species identification over multiple genera was already investigated in a very small setting of only a few species per genus, we focused on bacterial taxonomy and performed machine learning research in a genus-wide spectrum. In other words, we investigated bacterial species identification by considering a model for each genus separately. Two identification strategies were investigated and a comparison of the identification performance was made with the commercial identification system MIDI. As a third step, we tried to put the research of the previous topic in view of the hierarchical framework of bacterial taxonomy. FAME data was combined with 16S rRNA data and the machine learning approach of binary tree classifiers was analyzed. Where the identification performance was compared to the machine learning setting of the second topic, this approach was also considered as a way to put the results of FAME identification in a taxonomic context. Herefore, a statistical analysis of the results was performed. The work and resulting models described in this dissertation can easily be generalized and extended towards a larger bacterial spectrum. Moreover, FAME-based machine learning research even showed promising for other identification purposes, such as the identification of plant-pathogenic strains. Therefore, this work may contribute to the field of bacteriology by improving and rapidly updating the routinely used FAME-based bacterial species identification and by allowing simple implementation in laboratory information management systems.

A final but distinct research topic relates to the first paragraphs, where we mentioned that phenotypic data is mainly stored in private databases. As this restricts research possibilities, this last topic handles the creation of public FAME database, or FAME-bank, for sharing bacterial FAME profiles. A particular database structure was constructed, together with a user-friendly web application. This web application is currently in its alpha phase but can easily be extended

with multiple features. The main purpose of this FAME-bank.net project is to allow for more extensive FAME research that is only possible by inter-laboratory collaboration. The taxonomic scope of research will be extended and data set sizes will be able to grow. In this way, we close the circle and contribute also to the scientific community by extending the data sharing network. Moreover, with a larger FAME database, microbiologists and bioinformaticians will be allowed to develop personal and custom identification libraries, to analyze FAME data within a broader bacterial scope and to extend and generalize the presented machine learning research. In general, we regard this work as a steppingstone towards a further and better convergence of the two distinct scientific fields of bacteriology and machine learning.

A road map to this dissertation

A general road map to this dissertation is visualized in Figure 1. This figure illustrates how the four parts of this dissertation are organized and how this dissertation can be read.

In the first part of this dissertation, a general introduction is given on the fields of machine learning (Chapter 1) and bacteriology (Chapter 2). With this part, the reader becomes familiar with theory, principles, concepts and definitions needed for a better understanding of the research performed. Both chapters can be read separately and are not directly referring to one another. First, the field of machine learning is handled. In this chapter, we discuss the general principles involved in machine learning research such as learning, generalization, overfitting, etc. In the main part of this chapter, we focus on the three machine learning techniques used in our research. Each technique is put in a historical perspective and topics required for a good understanding of the technique are discussed. Next, a mathematical representation is given, together with the major (dis)advantages, though these mostly hold for many machine learning techniques. Finally, model evaluation is described in terms of the confusion matrix, performance measures, the ROC curve and statistics. In the second chapter, we deal with bacteriology. We first give a general overview on what bacterial taxonomy encompasses. Topics as taxonomy, species definition and concept, and bacterial identification are briefly explained. Importantly, to prevent any further misunderstanding, the terms classification and identification are explained in view of the two introductory chapters. Also, an overview of machine learning applications in bacteriology is given. In a second section, bacterial FAME analysis is discussed. The history of bacterial fatty acid analysis is reported and we explain what fatty acids are and how these are analyzed by the commercial Sherlock MIS system. Finally, fatty acid research performed at the Laboratory of Microbiology and the resulting in-house database are described. In each of the two sections of this chapter, the three genera *Bacillus*, *Paenibacillus* and *Pseudomonas* are described in the context of the respective sections.

The second part of this dissertation handles the scientific research performed during this doctoral study. This part comprises three chapters. In Chapter 3, we describe an initial data analysis. Information is given concerning the construction of the data sets, together with corresponding statistics. Next, for each species, average FAME profiles are analyzed. Again, these are discussed for each genus separately. Also, clustering experiments of the data are described followed by a TaxonGap analysis of each genus. Finally, a principal components analysis is

described and discussed. In the following chapter (Chapter 4), we discuss the FAME-based species classification experiments performed with the three machine learning techniques artificial neural networks, support vector machines and random forests. First, artificial neural network experiments for species identification in the genus *Bacillus* are described. Second, two strategies are reported and discussed in which the three machine learning techniques are handled for species identification of the three bacterial genera. In this section, the results are also compared with the performance of the commercial identification system Sherlock MIS and the identification of three independent third-party data sets is described. In Chapter 5, we discuss the integration of taxonomic knowledge in the identification models described in the Chapter 4. For a better understanding of the research described in this chapter, we briefly introduce binary tree classifiers and bacterial phylogeny. A new approach, *phylogenetic learning*, is described and discussed. In the second section of this chapter, we also propose a new method for easy visualization of the results obtained by phylogenetic learning.

The third part of this dissertation comprises a single chapter (Chapter 6), which describes the FAME-bank.net project. The construction of a public FAME database is described and the concept, main motives and advantages are discussed.

The final and fourth part is a general summary of the dissertation. General conclusions are given and future perspectives are proposed.

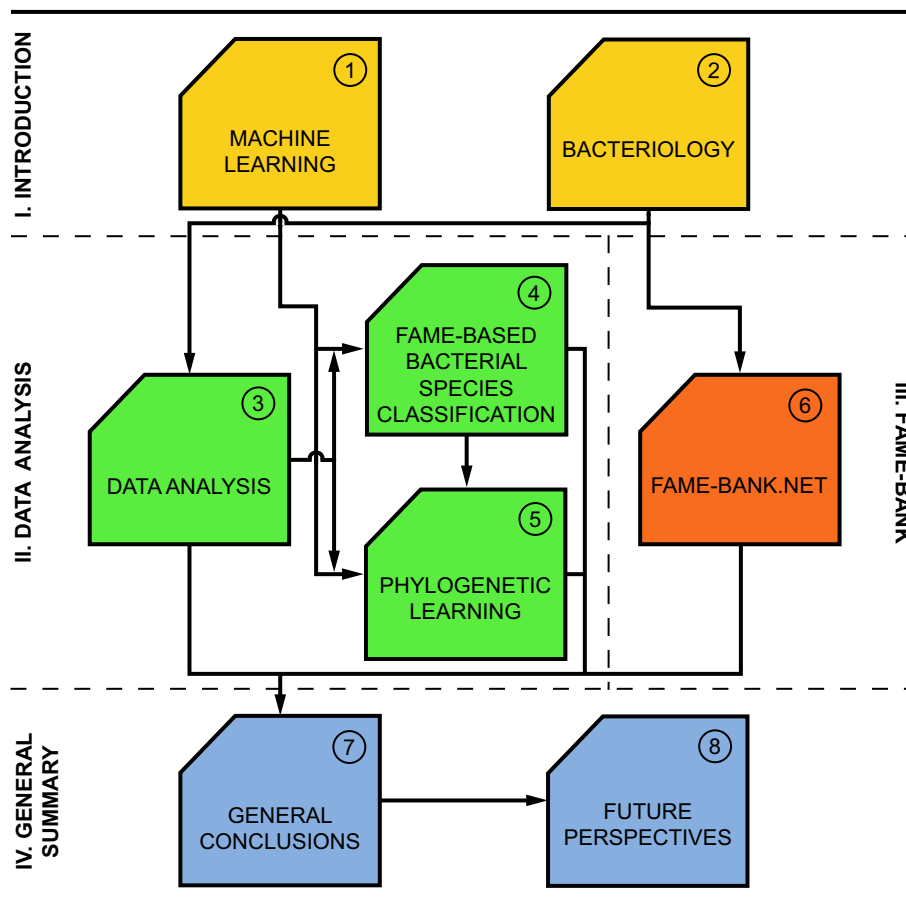


Figure 1: Dissertation road map. The different chapters are marked by their chapter number and title.

PART I

GENERAL INTRODUCTION

CHAPTER 1

Machine Learning

Intelligence is the ability to adapt to change.

STEPHEN HAWKING

... the story of the sheep dog who was herding his sheep, and serendipitously invented both large margin classification and sheep vectors...

ANA MARTÍN LARRAÑAGA

1.1 Introduction

When diving into literature and surfing on the World Wide Web, machine learning is described in many formal ways. One of the best descriptions of what machine learning really encompasses is given by Tom Mitchell in his book ‘Machine Learning’. He states in the opening sentence of the preface to his book that: “*The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience*”. Moreover, Mitchell states that machine learning is related to concepts and results from many fields, including statistics, artificial intelligence, philosophy, information theory, biology, cognitive science, computational complexity and control theory. In the following sections I summarize the concepts, theories and techniques needed for further understanding the machine learning background of the scientific research performed. This summary is mainly based on the excellent books of Mitchell (1997), Duda et al. (2001), Hastie et al. (2001, 2009) and Bishop (2006), to which I refer for a more detailed reading.

1.1.1 General Definitions and Concepts

Learning corresponds to acquiring new knowledge and, in machine learning terms, starts from a particular data set $S = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$, with N the data set size. Suppose that the N data points or instances comprise D features, also called the independent variables or predictors. These features can be binary, continuous or categorical. To each data point corresponds a vector $\mathbf{x}_i \in \mathbb{R}^D$ with $i = 1, \dots, N$, together with an associated output \mathbf{y}_i . Machine learning focuses on learning a model or ‘machine’ whose task it is to learn the mapping $\mathbf{x}_i \mapsto \mathbf{y}_i$. The machine is actually defined by a set of possible mappings $\mathbf{x} \mapsto f(\mathbf{x})$, or also $f(\mathbf{x}, \alpha)$ with α an adjustable parameter or a vector of such parameters. This machine is

assumed to be deterministic as for a given input \mathbf{x} and a particular choice of α , it will always give the same output. Setting α to a certain value results in a so-called ‘trained’ machine. Therefore, the input data set is generally called the training set and the phase of determining the function $f(\mathbf{x})$ is called the learning or training phase. Once the machine, model, learner or classifier is trained, it can be used to test the identity of unknown data points, that are comprised in a so-called test set (Burges, 1998; Bishop, 2006; Hastie et al., 2009).

When the output values y_i is known beforehand, it is possible to learn or train the model in a supervised way. When the output or target variables consist of one or more discrete categories or classes, the problem setting is called classification. In case of continuous output(s), the setting is referred to as regression. When the targets of the input data points are not known beforehand, it is possible to find groups of data points with a certain degree of similarity. This problem setting is called unsupervised learning or clustering (Bishop, 2006).

The aim in classification and regression problems is to find a good balance between the ability of predicting the training data correctly, called memorization, and the ability of achieving a similar performance on a set of unknown data, defined as generalization. The error rate on an independent test set is a good estimator of the generalization power of the model. The concept of generalization is very important as in real-world situations, one often has only a small subset of all possible data points at hand. Consequently, it is highly important to generalize over the pattern information or variability in the present subset (Fausett, 1994; Duda et al., 2001; Bishop, 2006; Hastie et al., 2009). In other words, generalization is about capturing the underlying trends in the data and distinguishing them from noise. Thus, the generalization ability determines the quality of the learned model. A commonly used method for obtaining a good generalization power is to use an additional validation set. Especially in data-rich situations, this is one of the best approaches. Typically, the input data set is randomly divided in a training, validation and test set. The training set is used to fit the models, the validation set is used for model selection based on the generalization ability and the test set is used to estimate the performance of the final model. Typically used split ratios are 50%, 25% and 25%, respectively, or 33% for each set (Hastie et al., 2009). In general, this approach is executed ten-fold to hundred-fold with random sampling of the different data sets. Final model evaluation is subsequently performed by averaging of the results. In this work, this type of generalization estimation is further denoted as simple validation. An example of the estimation of the generalization performance is visualized in Figure 1.1. Before training a model, the error on the training and validation set is typically high. Through learning from the training data, the error lowers. The minimum error on the validation set corresponds to the best generalization error that the model is able to achieve and the final model is constructed with the corresponding parameter setting. For other parameter value settings, the model is unable to fit the training data well due to a too low number of free parameters or the model has tuned its parameters to the training data values and becomes too complex. Both situations lead to a bad generalization and, thus, to a high validation error. This phenomenon is better known as under- and overfitting, respectively (Duda et al., 2001).

Mitchell (1997) defines overfitting in a very clear way: given a hypothesis space \mathcal{H} , hypothesis $h \in \mathcal{H}$ is said to overfit the training data if there exists some alternative hypothesis $h' \in \mathcal{H}$

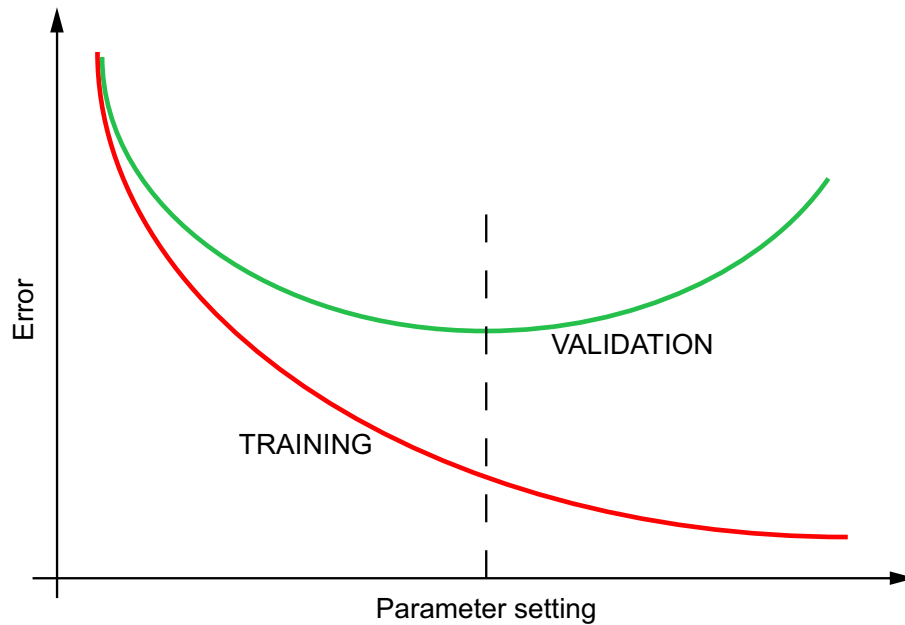


Figure 1.1: Generalization. Simple illustration of the errors on the training (red) and validation (green) set for varying parameter settings. In this curve, the error on both data sets decreases until a certain parameter value, denoted by the dashed line. Beyond this line, the error on the training data further decreases while the error on the validation set increases. In other words, the model learns the data values and starts to overfit. The generalization power of the model decreases and is given by the increased error on the validation set. The final model is constructed by the parameter setting corresponding to the minimum validation error.

such that h has a smaller error than h' over the training examples, but h' has a smaller error than h over the entire distribution of instances. Finding a good balance is particularly difficult when dealing with small data sets. In this case, the problem of overfitting becomes very severe and can be solved by cross-validation (Mitchell, 1997).

The principle of cross-validation is illustrated in Figure 1.2. The input data set is split into k equally sized parts, generally in a random manner. For the k^{th} part, the model is trained with the $k - 1$ other parts, while the performance is evaluated with the respective part. This training-validation process is done for each part and the cross-validation estimation of the error equals the average of the errors obtained by the different folds. When k is set equal to the data set size, leave-one-out cross-validation is performed (Bishop, 2006; Hastie et al., 2009). The choice of a good value for k depends on the bias-variance trade-off (more information below) and the computational load, that increases with the number of folds. In general, five-fold and ten-fold cross-validation is recommended for model selection (Breiman and Spector, 1992; Kohavi, 1995). In the context of cross-validation it is also important to note that in many real-world data sets the number of data instances varies per class, which was also the case for the data sets dealt with in this work. A good approach for dealing with these disproportions is to perform 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). In this work, we first splitted a stratified test set from the original data set and subsequently used stratified cross-validation for parameter optimization.

With the stratified test set the performance error of the final classifier was estimated. Often this latter step is skipped and performance estimation is done by the cross-validation. In this perspective, Varma and Simon (2006) state that, when dealing with small data sets, the calculated cross-validation error could be a biased estimate of the true error of the final classifier trained on all the data and using the optimal classifier parameters. The solution proposed in their paper is nested cross-validation in which two cross-validation loops are performed. An outer cross-validation loop for performance estimation and an inner cross-validation loop for parameter optimization. Ultimately, pooling is done of the different outer folds (or test data), implying a model performance estimate based on the complete data set.

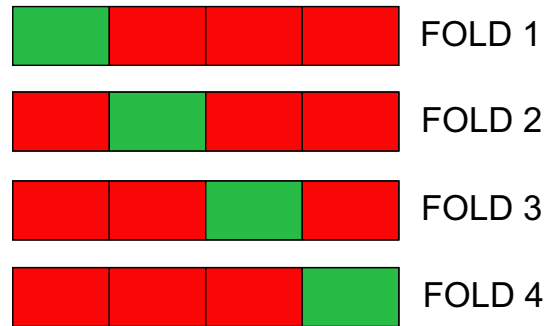


Figure 1.2: k -fold cross-validation. The initial data set is split into k equal parts, with $k = 4$ in this figure. Each part is once used for validation (green) while the other $k - 1$ parts are used for training (red). The final model performance equals the average performance over the k folds.

The goal of a learning model is to minimize the error on an independent test set. The expected squared prediction error of a certain test point \mathbf{x}_0 can be decomposed in:

$$\text{Error}(\mathbf{x}_0) = \text{Irreducible error} + \text{Bias}^2 + \text{Variance} \quad (1.1)$$

The first term corresponds to the variance introduced by the new test point and is regarded as noise. This variance cannot be avoided, no matter how well we estimate, unless it equals zero. The second term is the squared bias, the amount by which the average of the estimate differs from the true mean. This bias is also known as statistical bias. It can be estimated by repeatedly drawing training points from the input data set, by constructing the corresponding models and by averaging the resulting outcomes. The statistical bias is given by the difference between the average prediction and the true target. The third term is the variance or the expected squared deviation of the prediction around its mean. When minimizing the error on the test set, consequently, the goal is to minimize the statistical bias and variance (Dietterich and Bakiri, 1995; Bishop, 2006; Hastie et al., 2009). This decomposition is better known as the bias-variance tradeoff. One important factor in model selection is to choose the complexity of the model by trading bias off with variance, such that the test error is minimized. This tradeoff is also visualized by the curves in Figure 1.1. The X axis denotes model complexity and the validation curve represents the test error. If a model is not complex enough, it will underfit and may result in a large bias, implying a poor generalization. In the other extreme, if the model is too complex, the training data is fitted too well. Predictions will have a large variance and, thus,

also show a poor generalization (Hastie et al., 2009).

When talking about model selection, overfitting and model complexity, an important aspect to touch is regularization. If too many free parameters are used, model complexity will become too high which leads to overfitting and, conversely, when too few parameters are implemented, model complexity becomes low, leading to underfitting. In both cases generalization will be poor. In view of model complexity, a common approach to tackle overfitting is to reformulate the error minimization problem by adding an additional term for penalizing model complexity:

$$\min_{f \in \mathcal{H}} \left[\sum_{i=1}^N L(\mathbf{y}_i, f(\mathbf{x}_i)) + \lambda J(f) \right], \quad (1.2)$$

where $L(\mathbf{y}, f(\mathbf{x}))$ is a particular error function. The mean squared error is a popular example. $J(f)$ is a penalty function that quantifies the complexity of the function f and \mathcal{H} is a space of functions on which $J(f)$ is defined. This concept is defined as regularization. Specifically, the minimization problem becomes a trade-off between the error function and its complexity. The parameter λ defines this trade-off and is an additional parameter that needs to be optimized (Duda et al., 2001; Hastie et al., 2009).

1.1.2 Classification Settings

1.1.2.1 From Two-class to Multi-class

In this work, we were only confronted with supervised classification. This learning type is very general and can be subdivided in many different settings. The most popular and most studied setting is two-class or binary classification. In binary classification problems, only two classes are considered, typically represented by a positive and a negative label. Each data sample corresponds to one of the two labels. However, in many real-world situations, data sets comprise multiple classes. In supervised learning, this problem setting is called multi-class classification, which asks for a different approach. With respect to the learning aspect, the three popular solutions for solving the multi-class classification problem are learning a single multi-class classification model, the one-versus-others (also one-versus-rest or one-versus-all) approach and the one-versus-one approach. In the first setting, one classifier is trained to distinguish between all classes by solving one optimization problem. Of the different techniques discussed below, this setting is typically used in artificial neural networks (ANNs) and random forests (RFs). In the second approach, different binary classification models are learned that distinguish each class from all other classes, while in the third approach, different binary classifiers are trained to learn to distinguish between all pairs of classes. This latter setting is typically used in support vector machines (SVMs). A more detailed discussion about both approaches is given in Subsection 1.2.2.2. Besides the two-class and multi-class setting, different other learning types exist, such as

- Multi-label learning: each data sample is associated with a set of labels
- Structured learning: the objective is to learn a more complex structure such as graphs, trees, sequences, etc.

- Hierarchical classification: detailed information in the following subsection
- etc.

1.1.2.2 Multi-class Hierarchical Classification

A particular multi-class classification approach is that of classifying the multiple classes by means of a hierarchical structure or tree. Herein, a classification model is trained on each node of the tree to distinguish between the subset of data instances corresponding to the underlying classes. At present, machine learning papers describing multi-class classification with classes structured in a tree topology mainly focus on the area of web-, document-, text- and ontology-based classification. Many research problems involve multi-furcating tree nodes, and most papers deal with data instances corresponding to multiple classes structured in this kind of hierarchical setting. Classification problems related to this issue are better known as multi-label classification. In machine learning terms, learning by exploiting hierarchical structure information is called hierarchical classification (Koller and Sahami, 1997; McCallum et al., 1998; Dumais and Chen, 2000; Blockeel et al., 2002; Dekel et al., 2004; Kriegel et al., 2004; Barutcuoglu et al., 2006; Cesa-Bianchi et al., 2006; Rousu et al., 2006; Vens et al., 2008), learning with taxonomies (Hofmann et al., 2003) and structured label learning (Wu et al., 2005). Most of these studies do not involve hierarchical classification for single-label multi-class classification, meaning that each instance is classified at leaf level. Also, the hierarchical topologies are mostly predefined by a certain ontology or pre-existing class structure.

In view of the optimization problem, hierarchical classification is also proposed by handling multi-class classification by means of a tree of binary classifiers, also called binary tree classifiers (Lee and Oh, 2003; Cheong et al., 2004; Vural and Dy, 2004; Fei and Liu, 2006; Lu et al., 2007; Xia et al., 2007). In these studies, a tree architecture is constructed from the considered data set where tree inference is based on different algorithms for calculating distance measures or similarities between the considered classes. Ultimately, the resulting tree is used for multi-class classification by training a binary classifier at each node of the tree. We focused on this approach of hierarchical single-label multi-class classification. Evaluation of this classification approach is similar to that of the approaches mentioned in Subsection 1.1.2.1, which is further denoted as flat multi-class classification. Herein, a test instance is identified along the tree until classification into one of the leaves of the tree.

To complete the list of approaches, a totally different hierarchical classification method is the decision directed acyclic graph (DDAG) proposed by Platt et al. (2000). The basis of this multi-class classification algorithm is the one-versus-one approach where different binary SVM classifiers are structured in a bifurcating tree structure. The authors state that, compared to other multi-class SVM algorithms, the DAGSVM algorithm is superior in both training and evaluation time.

1.1.3 Balanced versus Imbalanced Data Sets

As in many real-world data sets, in this study we were confronted with imbalanced classes. In other words, the multiple classes of the data set consisted of a different number of data points.

Training with these type of data sets may become problematic. This class imbalance problem is addressed in a few overview papers such as Japkowicz and Stephen (2002) and Weiss and Provost (2003). For more detailed information, I refer to the references in these papers. Various strategies are proposed for dealing with class imbalances (Japkowicz and Stephen, 2002):

- Over-sampling: oversampling the minor class until it contains as many elements as the major class. Sampling may be performed at random or by focusing on specific data points or patterns
- Under-sampling: eliminating elements of the major class. Possible strategies are the same as for over-sampling
- Cost-modifying: modification of the relative misclassification cost by compensating for the imbalance ratio of the minor and major class

For these re-sampling techniques, different results are reported. Japkowicz and Stephen (2002) report that under-sampling is the least effective in many cases, while Weiss and Provost (2003) report that neither approach outperforms the other nor does any sampling rate consistently yields the best results. This can be regarded somewhat dubious, as classes with a massive amount of data points may contain a lot of redundant information that is irrelevant for the classification task. But when this is not the case, crucial information is lost, ultimately leading to a degradation of the classification performance. Note that with oversampling, due to copying of data, a possible overfitting may occur and that computation time for model construction will increase (Japkowicz and Stephen, 2002; Weiss and Provost, 2003). Japkowicz and Stephen (2002) report that for classifiers sensitive to the class imbalance problem, simple re-sampling techniques can improve performance (e.g. in case of ANNs) but some techniques may suffer from it (e.g. SVM training with undersampling). It is also clear that imbalanced classes may cause a performance degradation when analyzed by learning methods that assume balanced class distributions (Japkowicz and Stephen, 2002).

A totally different approach of handling skewed class distributions is reported by Weiss and Provost (2003). Instead of modifying the class distribution of the data sets, the authors suggest to adjust the resulting probability estimates. This work only took decision trees into account even though the authors believe their conclusions will also hold for other learners. Importantly, the authors conclude that, if no additional information is provided about the true class distribution and a class distribution must be chosen without any experimentation, the natural distribution and a balanced distribution are reasonable default training distributions. In this work, these approaches were followed. Due to imbalanced data sets with a high number of classes with a small number of data instances, balanced data sets with under-sampling and imbalanced data sets were considered for classification.

1.2 Machine Learning Techniques

In this study, we focused on three popular black-box models: artificial neural networks, random forests and support vector machines. In the following sections, a brief introduction to

the theory behind the three models is reported. Detailed reading can be done in many machine learning books.

1.2.1 Artificial Neural Networks

1.2.1.1 Introduction

The idea of artificial neural networks started approximately 70 years ago by a motivation to understand the brain and to exploit the strengths of biological neural systems. Information processing by biological neural systems is distributed over many neurons in a parallel fashion. This ability has been one of the cornerstones in the development of ANNs. The first neural network was designed in 1943 by McCulloch and Pitts and the first neural learning was implemented by Hebb in 1949. In the 1950s and 1960s, one of the most popular models for artificial neural networks was developed by Rosenblatt and other researchers: the perceptron. Next to the Hebb rule and the perceptron, different other learning algorithms were developed such as the popular Widrow-Hoff learning rule. Besides this rule, Widrow also developed the adaptive linear neuron (adaline) and its multi-layer extensions. In the preceding decades, a multitude of artificial neural models were developed such as the Kohonen self-organising maps, Hopfield nets, the neocognitron and the Boltzmann machine. One very popular learning method is the backpropagation method which is also applied in this work (Fausett, 1994).

An ANN is a mathematical model, regarded as an artificial information-processing system, that incorporates the major parts of the biological neuronal model. To understand the idea behind ANNs, this biological neuron model needs an explanation first (Fausett, 1994; Mitchell, 1997). A biological neuron consists of three main components: dendrites, the soma and the axon. The structure of a biological neuron is also displayed in Figure 1.3. The many dendrites of the neuron may receive signals from other neurons. This signal transmission is chemically regulated over a synaptic gap between neurons. The dendrites are connected with the soma which receives all incoming signals. When the sum of all signals exceeds a certain threshold, the cell fires and the soma transmits a signal towards the axon for signal forwarding to other neurons. Transmission of the signal through the neuron is accomplished by an action potential resulting from different concentrations of ions on either side of the neuron's axon sheath (Fausett, 1994).

As a mathematical model of the biological neuron, ANNs consists of the same structural elements. Information processing occurs at many simple units, called neurons, signals are transmitted between neurons over connection links, each connection link is associated with a weight that multiplies the signal and each neuron applies an activation function to the sum of the weighted input signals to determine an output signal. These elements are, thus, similar to the soma, the dendrites and axon, and the neuron activation for signal transmission, respectively (Fausett, 1994; Mitchell, 1997). As an example, a mathematical representation of an artificial neuron is displayed in Figure 1.4.

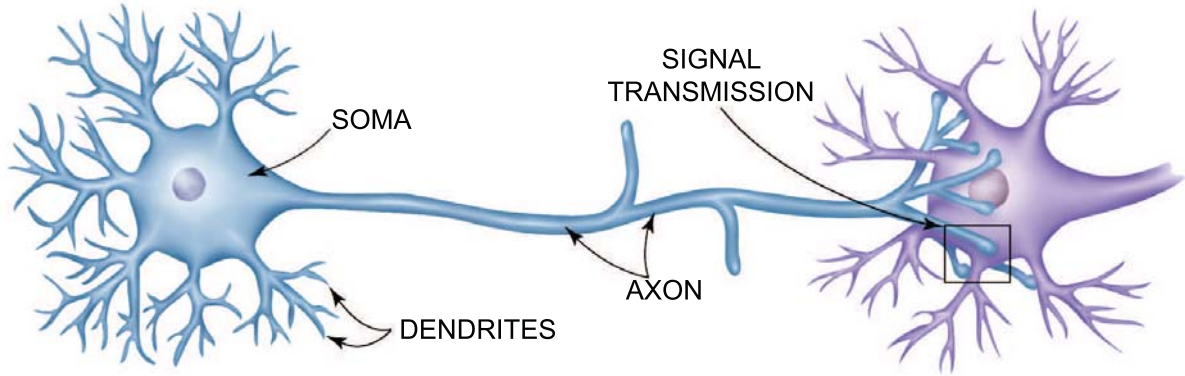


Figure 1.3: Structure of biological neuron and signal transmission. Signals are retrieved by the multiple dendrites and transported to the soma. If the sum of all signals exceeds a certain threshold, the signal is further transported to the axon for transmission to the other neurons (McMurry and Castellion, 2002).

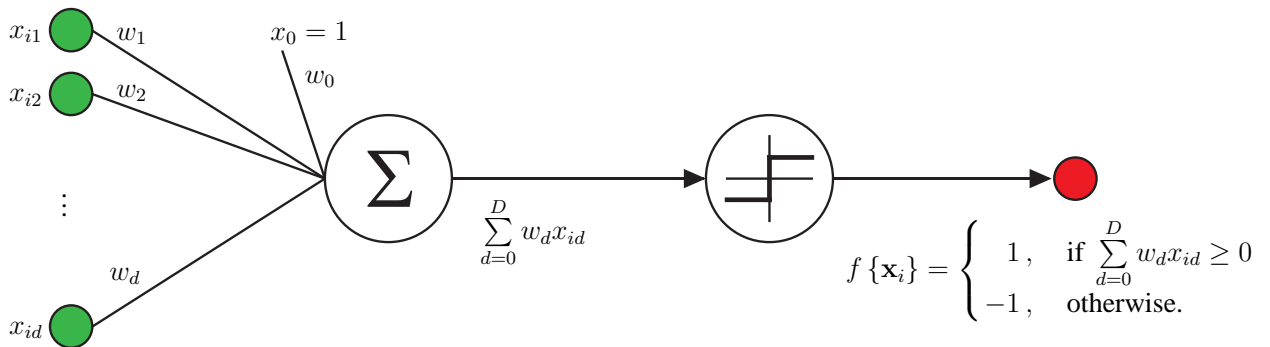


Figure 1.4: Mathematical representation of the perceptron. The input values of training point \mathbf{x}_i correspond to x_{id} , with $d = 0, \dots, D$ and D the number of features. Input value x_0 is called the bias neuron and its value is set to 1. To each input is associated a weight w_d . The weighted sum of the input values is calculated and the resulting sum is converted into a specific output value y as defined by the bipolar step function. A threshold on the weighted input signal is set equal to zero (adapted from Mitchell (1997)).

1.2.1.2 The Perceptron: A Basic Unit

The perceptron is one of the most popular neural network types. In fact, the model is a linear discriminant model for two-class classification (Duda et al., 2001; Bishop, 2006). A perceptron takes an input data point or vector \mathbf{x}_i from the training set and calculates a linear combination of the input values with their corresponding weights

$$\sum_{d=1}^D w_d x_{id} + w_0, \quad (1.3)$$

with the value of the bias neuron x_0 set to 1 and D the number of features. This linear combination is also called the decision boundary, representing a line in \mathbb{R}^2 or a hyperplane in higher dimensions. The weight vector \mathbf{w} is perpendicular to any vector lying in the hyperplane and, as such, defines its orientation. This formulation is also visualized in Figure 1.5. Consequently, for the two-class situation, the data points are separated in a positive class (region above the

hyperplane) and a negative class (region below the hyperplane). The weights determine the contribution of input value x_{id} ($d = 1, \dots, D$) to the output of the perceptron. Next, a non-linear transformation is performed to map the weighted input signal to an output value. The latter method is also better known as activation and the corresponding function is called the activation function. The perceptron uses the (bipolar) step function for non-linear transformation (Fausett, 1994; Bishop, 1995; Mitchell, 1997; Duda et al., 2001; Hastie et al., 2009). Where in Figure 1.4 a bipolar step function is used for activation, the linear combination can similarly be evaluated by

$$f(\mathbf{x}_i) = \begin{cases} 1, & \text{if } \sum_{d=0}^D w_d x_{id} \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (1.4)$$

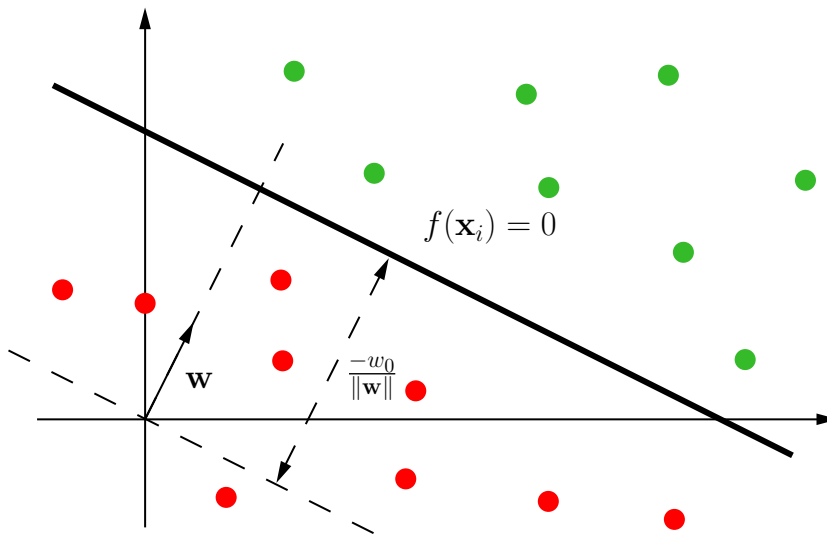


Figure 1.5: Linear decision boundary. Linear decision boundary corresponding to $f(\mathbf{x}_i) = 0$ or $\sum_{d=1}^D w_d x_{id} = -w_0$ in a two-dimensional input space. The weight vector \mathbf{w} defines the orientation of the boundary, while the bias weight w_0 defines the position of the boundary in terms of its perpendicular distance from the origin. The boundary separates the green data points (positive class) from the red data points (negative class).

From (1.3) and (1.4) follows that the bias weight w_0 acts as a threshold in the non-linear transformation of the netto input signal and determines the location of the decision hyperplane (Fausett, 1994; Bishop, 1995; Duda et al., 2001). Alternatively, an additional threshold θ can also be defined on the activation function (Fausett, 1994). Hereby, the width of the hyperplane can be changed and the boundaries of the hyperplane can be represented as

$$\begin{cases} \sum_{d=0}^D w_d x_{id} > \theta, & \text{representing the positive class} \\ \sum_{d=0}^D w_d x_{id} < -\theta, & \text{representing the negative class} \end{cases} \quad (1.5)$$

A single perceptron can be used to represent all primitive boolean functions (AND, NAND, OR and NOR). Learning these functions corresponds to making good choices for the different

weights \mathbf{w} . In terms of the hypothesis space \mathcal{H} , this implies that all candidate hypotheses considered correspond to the set of all possible weight vectors. Different learning algorithms can be used to converge to the correct weights. Initially, one mostly starts with a random initialization of the weights, drawn from a single uniform distribution. Next, in an iterative process the different training points are applied to the perceptron. The perceptron learning rule modifies the weights with a certain value Δ_d if a certain point is misclassified according to the following two rules:

$$w_d \leftarrow w_d + \Delta w_d \quad (d = 1, \dots, D) \quad (1.6)$$

$$\Delta w_d = \alpha(\mathbf{y}_i - f(\mathbf{x}_i))x_d, \quad (1.7)$$

with \mathbf{y}_i and $f(\mathbf{x}_i)$ the target and generated output, and α the learning rate which is a positive constant. The learning rate is used to moderate the modification degree of the different weights and is typically set to a small value. As can be seen, weights remain unchanged when a correct output value is obtained. Ultimately, the training data is repeatedly applied to the perceptron until correct classification of all training points. The iterative process can be represented as moving the separating hyperplane $w_d x_{id}$ towards the boundary corresponding with correct classification (Mitchell, 1997). The perceptron convergence theorem states that if an exact solution exists, thus, when the data is linearly separable, then the perceptron learning algorithm is guaranteed to find an exact solution in a finite number of steps (Minsky and Papert, 1969; Fausett, 1994; Duda et al., 2001).

1.2.1.3 Feed-forward Multi-layer Networks with Backpropagation

In this work, feed-forward multi-layer neural networks with backpropagation are used. To make the algorithm of this technique clear, the first paragraphs of this section briefly explain the principles and concepts behind this technique.

1.2.1.3.1 Gradient Descent

If the training data is not linearly separable, the delta rule can be applied as an alternative learning rule. This rule approximates the correct solution of the problem by converging to the best fit. The basis of this rule is the gradient descent algorithm, which searches the hypothesis space of possible weight vectors to find the weights best fitting the data points. Herein, gradient descent searches for a weight vector that minimizes an error function. By starting with an arbitrary initial weight vector, this vector is repeatedly modified in small steps in the direction of the steepest descent of the error function. Therefore, gradient descent is also known as steepest descent. This direction can easily be calculated by taking the derivative of the error function. This also motivates the choice of a low learning rate value, as too large values lead to larger steps in the gradient descent with the possibility of not converging to, or even diverging from, the global minimum of the error function. Note that too small values will lead to a very slow convergence and very long training time. Gradient descent learning has become a basic algorithm for searching hypothesis spaces and is the main principle behind basic backpropagation (more

information below). A major risk of this method is the convergence to a local minimum of the error function, if present. Different alternative strategies have been constructed to overcome this problem and are discussed in Subsection 1.2.1.3.5 (Bishop, 1995; Mitchell, 1997).

1.2.1.3.2 Activation Functions

To learn functions more complex than boolean functions, different neuron units should be combined. The perceptron is one of the most popular neuron units for learning in these multi-unit neural networks (Mitchell, 1997). Moreover, when non-linear functions need to be learned, multiple layers of neurons have to be considered. The most simple and popular example of a non-linear function is the XOR function. Learning this function can only be achieved by an ANN with two neuron layers (Duda et al., 2001). Consequently, a common setting for learning complex non-linear functions is to train a multi-layer network. An example of this type of ANN is visualized in Figure 1.6. However, multi-layer networks consisting of linear units as discussed in Subsection 1.2.1.2 will still only be capable of learning linear functions. An activation function for a backpropagation ANN should be non-linear, continuous and differentiable (more information below). Therefore, alternative activation functions are introduced in the single neurons and only these functions are further considered in this study. To keep the weights and activations bounded, and to keep training time limited, saturated functions are chosen. Monotone functions can be convenient but are not essential. In most cases, for each neuron layer the same activation function is used. This function may, however, differ between layers (Fausett, 1994; Bishop, 1995; Mitchell, 1997; Duda et al., 2001). Different types of functions exist and the sigmoid function is most widely used. Next to the step function, the logistic sigmoid and hyperbolic tangent sigmoid activation functions with the respective input-output mapping are defined and visualized in Table 1.1. The presented activation functions may be customized by modifying their slope. For instance, in case of the sigmoid function, this can be attained by changing e^x to $e^{\sigma x}$. By choosing for σ a very small value, linear functions can be approximated. Note that the hyperbolic tangent sigmoid function is equivalent to the hyperbolic tangent function $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. It is clear that small inputs and weights will result in a summed input near the origin of the sigmoid function which, approximates a linear transformation, while larger values will approximate a step function mapping (Fausett, 1994; Bishop, 1995; Mitchell, 1997). Different other activations exist such as the arctangent function or the radial basis function, which is used in radial basis function networks (Fausett, 1994).

Because the defined activation functions are differentiable, the derivatives of the error function with respect to the weight parameters can easily be evaluated and the gradient descent becomes a highly suitable learning rule for approximating the objective function (Bishop, 1995; Mitchell, 1997). Activation functions, such as the sigmoid function, also motivate the use of random initialization of the network weights. It can be seen that too small or too large weights will possibly result in too small derivatives of the activation function, leading to too small weight updates. This will ultimately result in extremely slow training. A common procedure is to initialize weights to random values in the intervals $[-0.5, 0.5]$ or $[-1, 1]$.

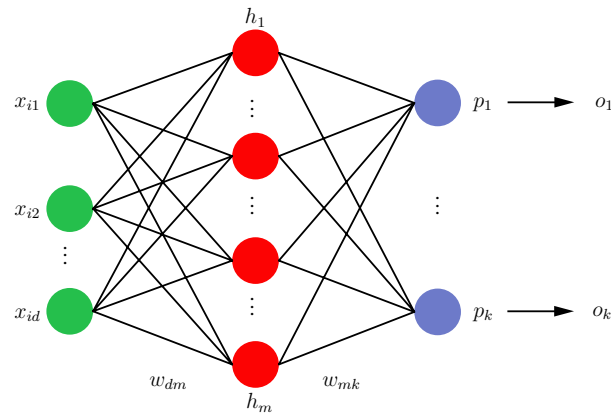


Figure 1.6: Architecture of a fully connected multi-layer artificial neural network. The first layer of the network consists of the input data values x_{id} with $d = 1, \dots, D$ and D the number of features (the bias not considered). The input data is forwarded over weighted neuron connections to the hidden neurons h_m with $m = 1, \dots, M$, M the optimal number of hidden neurons and w_{dm} the weights on the corresponding connections. A weighted input-output mapping is performed at the hidden neurons and the resulting values are forwarded to the output neurons p_k over connections with weights w_{mk} , with $k = 1, \dots, K$. K output neurons are considered with K the number of classes in the data set in case of classification and $K = 1$ in the case of regression. Again, a weighted input-output mapping is performed, resulting in the output values o_k .

1.2.1.3.3 Feed-forward Multi-layer ANN

In this work, the architecture of a feed-forward multi-layer ANN is chosen. Herein, N input values are presented to the ANN by weighted connections, with D the number of features present in the respective data set plus a bias value with an associated weight of 1. A first layer of neurons deals with these input values according to the perceptron concept as described above. A non-linear mapping of the input variables is also called a basis function of the model and the resulting input-output mapping is presented to a second and final layer of K neurons, also called the output neurons. Also in this case, a bias term is considered and the mapped values are transformed according to the perceptron concept. Moreover, in this setting a mapping occurs of a linear combination of the basis functions with the respective weights. At this level, the non-linear separability problem can be transformed into a linear separability problem. In most cases, for each neuron layer the same activation function is used. This function may, however, differ between layers (Fausett, 1994; Mitchell, 1997; Duda et al., 2001; Bishop, 2006). In case of classification problems, K equals the K classes of the respective data set and K equals 1 in case of regression.

The principle of input values flowing through the network towards the output neurons without any feed-back connections is called feed-forward. An ANN with backtracking or feed-back connections between the units is known as a recurrent net. Neurons positioned between the input layer of input values and the layer of output neurons are also called hidden neurons. Importantly, the number of these hidden neurons determines the number of connections in the network and, as such, the hidden neurons govern the power of the ANN and determine the complexity of the decision boundary. According to the respective learning problem, the number

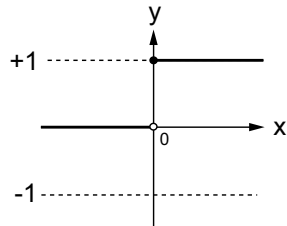
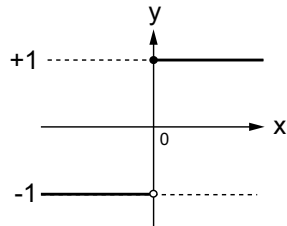
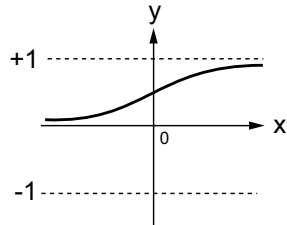
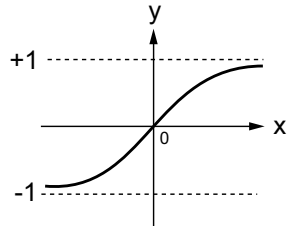
Function	Mathematical representation	Graph
Step	$a(x) = \begin{cases} 1, & \text{if } y \geq 0 \\ 0, & \text{if } y < 0 \end{cases}$	
Bipolar step	$a(x) = \begin{cases} 1, & \text{if } y \geq 0 \\ -1, & \text{if } y < 0 \end{cases}$	
Logistic sigmoid	$a(x) = \frac{1}{1+e^{-x}}$	
Hyperbolic tangent sigmoid	$a(x) = \frac{2}{1+e^{-2x}} - 1$	

Table 1.1: Overview of commonly used ANN activation functions. The mathematical formula and representation are given for the step and bipolar (or symmetric) step function, the logistic sigmoid function and hyperbolic tangent sigmoid function.

of these neurons needs to be optimized for optimal convergence to the global minimum of the error function. Optimization of the number of hidden neurons is commonly done by calculating the network error over a series of numbers of hidden neurons and by choosing the number that minimizes the error. Hereby, the complexity of the network is controlled in order to avoid overfitting. Another possibility to control model complexity is weight decay, which is an example of regularization. This method is not further considered in this study (Fausett, 1994; Bishop, 1995; Duda et al., 2001; Bishop, 2006).

The multi-layer aspect can be regarded two-fold. In terms of nodes, three layers exist: an input layer of input values, and a hidden and output layer of neurons. In terms of the weighted connections, two layers exist to connect the nodes of the input layer to the output layer. In this study, only one hidden layer is implemented, even though multiple hidden layers can be integrated in the ANN. For most learning problems, one hidden layer is sufficient to approximate any continuous function with arbitrary accuracy, provided sigmoidal activation functions and a sufficiently large number of hidden neurons. An important consequence of integrating multiple

hidden layers is a longer training time due to the larger number of connections and weights. Also, an ANN with multiple hidden layers is more prone to converging to local minima. As such, an ANN with this architecture is regarded to be a general parametrized multivariate non-linear functional mapping (Fausett, 1994; Bishop, 1995; Duda et al., 2001).

1.2.1.3.4 Backpropagation

For training in this kind of architecture the backpropagation algorithm is mostly applied. Backpropagation gained a lot of popularity due to its computational efficiency. This algorithm, also known as the generalized delta rule, uses the concept of perceptron learning. The backpropagation algorithm involves three stages: feed-forward of the input values towards the output neurons over the hidden neurons, backpropagation of the error calculated between target and output value from output layer to input layer and, third, update of the ANN weights according to the backpropagated error (Fausett, 1994; Bishop, 2006). Basic backpropagation employs gradient descent to find those weights \mathbf{w} that minimize the squared error between output and target values

$$E(\mathbf{w}) = \sum_{i=1}^N \sum_{k=1}^K (y_{ik} - f^k(\mathbf{x}_i))^2 \quad (1.8)$$

over all training points N in the data set and over all output neurons p_k with $k = 1, \dots, K$, with y_{ik} and $f^k(\mathbf{x}_i)$ the target and output values associated with the k^{th} output neuron (Hastie et al., 2009). As this error function is a differentiable function of the network weights, the derivatives of the error function can be used to find weight values that minimize the error function. Note that different alternative error functions may be used for solving learning problems (Fausett, 1994; Bishop, 1995).

This tri-fold basic backpropagation algorithm is iterated multiple times until a certain stopping condition is satisfied. One iteration is called an epoch, which corresponds to one cycle through the entire set of training data. A variety of stopping conditions can be applied, but stopping is mostly defined by a certain number of epochs or a minimum error between the target and output values (Fausett, 1994). Alternatively, also one of the several parameters of the network can be used. In this work, we mainly focused on the ANN error as stopping criterium, with simple validation or cross-validation. If the error on the validation set starts to increase, the iterative backpropagation process is halted. This method is also known as early stopping, which is also visualized in Figure 1.1, where the X axis corresponds to the number of training epochs. Early stopping halts training of the ANN at the epoch corresponding with the minimum validation error, given by the dashed line (best generalization). Early stopping is used as an alternative to regularization to control network complexity. By this method, a good generalization can be achieved and overfitting becomes prevented (Fausett, 1994; Bishop, 1995; Mitchell, 1997; Bishop, 2006; Hastie et al., 2009).

The principle of backpropagation for learning in a feed-forward multi-layer ANN is described in Algorithm 1. This basic algorithm considers a feed-forward multi-layer ANN consisting of a layer of input values x_{id} , with $d = 1, \dots, D$, of a particular data point \mathbf{x}_i ; a hidden neuron layer and an output neuron layer. These layers are fully interconnected by links with weights

w_{dm} and w_{mk} , $m = 1, \dots, M$ with M the number of hidden neurons, and $k = 1, \dots, K$ with K the number of output neurons. An input bias x_0 and a hidden bias h_0 are considered with weights w_{0m} and w_{0k} , respectively. The hidden and output neurons map their netto weighted input signal by the activation functions a_h and a_o , respectively, to the respective values h_m and o_k , respectively. Optimization of the network weights in the basic backpropagation algorithm is performed by gradient descent. In the presented algorithm, the weights are updated after all training points are presented, which is also called batch updating. More often used is incremental, online or sequential learning in which the weights are updated after the presentation of each training point. Training by this method but with randomly selection of the data points is called stochastic learning. Sequential methods are preferred over batch methods because of a higher computational efficiency and a more easy escape from local minima (Fausett, 1994; Bishop, 1995; Mitchell, 1997; Duda et al., 2001; Bishop, 2006). Backpropagation also allows to calculate the Jacobian and the Hessian matrices. The former matrix consists of the derivatives of the outputs with respect to the inputs, thus showing how the output changes with respect to the input. The Hessian matrix consists of the second derivatives of error function in terms of the weights. This matrix is used in alternative weight optimization algorithms, and allows to determine the least significant weights in the network, error intervals to the predictions and regularization parameters (Bishop, 1995).

1.2.1.3.5 Alternative Training Algorithms

Backpropagation is known to suffer long training times of tens to thousands of epochs due to slow weight optimization. To decrease training time, numerous alternative optimization algorithms are proposed. weight optimization is based on two main principles: the choice of a direction on the error function and the choice of a distance to move. In basic gradient descent, the gradient of the error function determines the direction in which the function increases more rapidly and, as such, the negative of the gradient defines the direction of the most rapid decrease. For this algorithm, the second question related to the weight optimization issue is addressed by the learning rate α (Bishop, 1995; Mitchell, 1997).

The most simple adaptation of the basic backpropagation algorithm regards the choice of updating the network weights incrementally or in batch (see also Subsection 1.2.1.3.4). Where gradient descent is used for training towards the negative direction of the error function, too large or too small learning rate values may possibly result in not converging to the global minimum. A too large learning rate may even result in divergent oscillations. As an alternative, the learning rate can be made variable by gradually decreasing the rate with a higher number of steps or even allow each weight to have its own learning rate, known as the delta-bar-delta rule (Fausett, 1994; Bishop, 1995; Mitchell, 1997). Also, each weight update can be made dependent of the previous weight update by adding a constant μ , called the momentum, with $0 \leq \mu < 1$. For the hidden neurons h_m , the corresponding weight update with momentum becomes (analogous for the output neurons):

$$\Delta w_{dm}(t+1) = \alpha \delta_m x_{id} + \mu \Delta w_{dm}(t) \quad (1.9)$$

Algorithm 1 Basic backpropagation algorithm.

Require: Initialize w_{dm} , w_{0m} , w_{mk} and w_{0k} at small random values

Require: Initialize the learning rate α

```

1: while Stopping criterium = false do
2:   STAGE 1: FEED-FORWARD OF VALUES OF DATA POINTS  $\mathbf{x}_i$ 
3:   for all hidden neurons  $h_m: m = 1 \rightarrow M$  do
4:      $h_m \leftarrow a_h \left( w_{0m} + \sum_{d=1}^D x_{id} w_{dm} \right)$ , with  $d = 1, \dots, D$  and  $D$  the number of features
5:   end for
6:   for all output neurons  $o_k: k = 1 \rightarrow K$  do
7:      $o_k \leftarrow a_o \left( w_{0k} + \sum_{m=1}^M h_m w_{mk} \right)$ 
8:   end for
9:   STAGE 2: BACKPROPAGATION OF THE ERROR
10:  Calculate the error  $\delta$  and the corresponding weight and bias weight correction  $\Delta w$ 
11:  for all output neurons  $o_k: k = 1 \rightarrow K$  do
12:     $\delta_k \leftarrow (y_{ik} - f^k(\mathbf{x}_i)) a'_o \left( w_{0k} + \sum_{m=1}^M h_m w_{mk} \right)$ , with  $y_{ik}$  the corresponding target value
13:     $\Delta w_{mk} \leftarrow \alpha \delta_k h_m$  and  $\Delta w_{0m} \leftarrow \alpha \delta_m$ 
14:  end for
15:  for all hidden neurons  $h_m: m = 1 \rightarrow M$  do
16:     $\delta_m \leftarrow a'_h \left( w_{0m} + \sum_{d=1}^D x_{id} w_{dm} \right) \sum_{k=1}^K \delta_k w_{mk}$ 
17:     $\Delta w_{dm} \leftarrow \alpha \delta_m x_{id}$  and  $\Delta w_{0m} \leftarrow \alpha \delta_m$ 
18:  end for
19:  STAGE 3: WEIGHT UPDATES
20:  for  $d = 1 \rightarrow D$  do
21:    for  $m = 1 \rightarrow M$  do
22:       $w_{dm}(\text{new}) \leftarrow w_{dm}(\text{old}) + \Delta w_{dm}$ 
23:    end for
24:  end for
25:  for  $m = 1 \rightarrow M$  do
26:    for  $k = 1 \rightarrow K$  do
27:       $w_{mk}(\text{new}) \leftarrow w_{mk}(\text{old}) + \Delta w_{mk}$ 
28:    end for
29:  end for
30:  Evaluate stopping criterium
31: end while

```

Momentum keeps the direction of training in the same direction of the previous direction. By training with a momentum, the weight updates are larger when training is performed in the same direction. As a consequence, the likelihood that the algorithm converges to local minima is reduced and the algorithm continues to convergence in regions where the gradient is unchanging. The momentum can, however, result in a weight change that increases the error (Fausett, 1994; Mitchell, 1997). Including a momentum generally results in significant performance improvement of the gradient descent algorithm, but implies the optimization of an additional model parameter (Bishop, 1995). Instead of considering a direction on the error function and choosing how far to move, one could also consider some search direction in weight space and find the minimum of the error function along that direction. This parameter optimization method is known as line search. Another popular optimization technique is training with con-

jugate gradient. In this training algorithm, each gradient is chosen orthogonal to the previous gradient where the search direction is chosen such that the component of the gradient is parallel to the previous search direction. So, the new search direction is said to be conjugate to the previous direction. These methods with their numerous variations, advantages and disadvantages are not further considered in this study (Bishop, 1995). Nonetheless, Mitchell (1997) states that even though alternative error minimization methods may lead to improved efficiency in ANN training and computation, no significant generalization improvement may be attained. The only likely impact is on the final error, as that different methods may converge to different local minima.

1.2.1.4 Properties

Even though ANNs are already beyond the state-of-the-art in machine learning, ANNs have been, and still are, successfully applied in many scientific domains. ANNs have a multitude of advantages. A short overview is given (Mitchell, 1997; Hastie et al., 2009):

1. Input and output values can be real-valued or binary-valued
2. Backpropagation is basically a simple training algorithm that can efficiently be parallelized
3. The target function may be real-valued, discrete-valued and vector-valued
4. Availability of a multitude of ANN learning algorithms and optimization techniques
5. Robustness to noisy data
6. ANNs can handle many features
7. An ANN has both a long-term and short-term memory where the first corresponds to the connection weights and the second to signals sent by the neuron

Nonetheless, several disadvantages are inherent to ANN algorithms. Hastie et al. (2009) state that there is quite an art in training ANNs. You can only agree upon this, as the main disadvantage of ANNs is their over-parametrized nature. Other disadvantages are (Fausett, 1994; Mitchell, 1997; Duda et al., 2001; Hastie et al., 2009):

1. Most error functions are complex, mostly non-convex (implying multiple local minima) and are possibly unstable. Therefore, ANNs possibly converge to one of the local minima and not to the global minimum. Mitchell (1997) denotes that more weights, corresponding to a higher dimensional space, provide more dimensions that may provide an 'escape route' route for local minima. This problem can be alleviated by changing the learning rule, altering the learning rate and/or adding a momentum. However, this leads to the additional disadvantage of the optimization of additional model parameters. A rough method of just training multiple networks using the same data with different random weights could be performed, but is not preferred for efficiency and computational reasons. This method may lead to different local minima and the network with best performance is finally selected

2. Depending on the training algorithm and the stopping condition, the iterative backpropagation process in a feed-forward multi-layer ANN may be executed ten to thousand times leading to a large training time
3. ANNs are quite sensitive to overfitting and can easily achieve a (too) high complexity
4. ANNs cannot handle categorical variables. To solve this problem, dummy features need to be introduced. Herefore, each categorical variable is represented by binary encoding in which the number of bits equals the number of categorical variables minus 1
5. Standardization of the input is required. Numerical values of different magnitude will result in an adjustment of the weights in favor of the largest values and the error will hardly depend on the small input values. It is best to standardize all inputs to have a zero mean and a standard deviation of one
6. ANNs cannot handle missing values
7. ANNs do not provide probability estimates

1.2.1.5 Software

ANN implementation, training and testing was done using the Matlab software versions 7.0 (R14) and 7.3 (R2006b) and the Matlab Neural Network toolbox versions 4.0.3 and 5.0.1. Wrapper code was written in Perl for the automation of data import and conversion, parameter optimization and statistical analysis of the identification results.

Numerous neural network software packages, toolboxes, wrapper codes, ... exist and ANNs are implemented in (almost) every statistical and mathematical software package. A list of all these implementations would constitute a section on itself. Making a choice of one package, however, restricts the use of different types of neural networks, training algorithms, activation functions ... In the Matlab Neural Network Toolbox, however, a large set of ANN types and training functions are available.

For training feed-forward multilayer ANNs, the resilient propagation (RPROP) algorithm was selected for reasons of fast computation time and good accuracy. RPROP is a variant on the gradient descent algorithm. When dealing with sigmoid activation functions, gradient descent may become problematic when the gradient has a very small magnitude and cause small changes in the weights, even if these are far from optimal. The purpose of RPROP is to eliminate this harmful effect of the magnitude of the partial derivative. As such, the weight optimization process does not consider the magnitudes of the weight changes, but focuses only on the signs of the partial derivatives for weight optimization. This leads to a higher efficiency in computation time and storage.

More specific, RPROP performs batch update gradient descent by integrating for each weight w_{dm} an additional update term Δ_{dm} in the weight update step. These update values are adapted during the learning process based on the local gradient of the error function E ,

expressed as the partial derivative $\frac{\partial E}{\partial w_{dm}}$. The learning rule is as follows:

$$\Delta_{dm}^{(t)} = \begin{cases} \mu^+ \times \Delta_{dm}^{(t-1)}, & \text{if } \frac{\partial E}{\partial w_{dm}}^{(t-1)} \times \frac{\partial E}{\partial w_{dm}}^{(t)} > 0 \\ \mu^- \times \Delta_{dm}^{(t-1)}, & \text{if } \frac{\partial E}{\partial w_{dm}}^{(t-1)} \times \frac{\partial E}{\partial w_{dm}}^{(t)} < 0 \\ \Delta_{dm}^{(t-1)}, & \text{otherwise,} \end{cases} \quad (1.10)$$

where $0 < \mu^- < 1 < \mu^+$. This can be interpreted as follows: each time the sign of the partial derivative of the corresponding weight w_{dm} changes, the update value Δ_{dm} is decreased by a factor μ^- . This indicates that the last update was too large and the algorithm has jumped over a local minimum. Conversely, if the derivative retains its sign, the update value is slightly increased to accelerate convergence. Once the individual update value is adapted, the weight update is as follows:

$$\Delta w_{dm}^{(t)} = \begin{cases} -\Delta_{dm}^{(t)}, & \text{if } \frac{\partial E}{\partial w_{dm}}^{(t)} > 0 \\ +\Delta_{dm}^{(t)}, & \text{if } \frac{\partial E}{\partial w_{dm}}^{(t)} < 0 \\ 0, & \text{otherwise.} \end{cases} \quad (1.11)$$

One exception is considered: if the partial derivative changes its sign, the previous weight update is reverted: $\Delta w_{dm}^{(t)} = -\Delta w_{dm}^{(t-1)}$. To avoid a double punishment of the update value, no adaptation should be done of the update value in the succeeding step. This can be done by setting $\frac{\partial E}{\partial w_{dm}}^{(t-1)} = 0$ in the learning rule above (Riedmiller and Braun, 1993; Demuth et al., 2006). Riedmiller and Braun (1993) state that setting the two parameters μ^- and μ^+ to the respective values of 0.5 and 1.2 provides very good results independent of the problem setting. These values are also implemented by the Matlab Neural Network Toolbox (Demuth et al., 2006). In this study, we chose to set these parameters at their default values.

1.2.2 Support Vector Machines

While kernel theory was introduced in 1964 by Aizerman et al. (1964), only since the early 1990s the group of Vapnik (Computer Learning Research Center, University of London, UK) has brought it back unto attention by combining this theory with large margin classifiers, leading to the SVM (Boser et al., 1992; Cortes and Vapnik, 1995). Since the introduction of the SVM, the use of kernel functions has become one of the hot topics in machine learning research. Together with the growing popularity of kernels, the number of applications has witnessed a boost in the last ten years. Especially in the field of engineering and bioinformatics, the number of publications describing the application of SVMs on particular topics, problems and data sets has grown extensively.

As with ANNs, a lot of theoretical information is available about support vector machines and kernel-based learning. In the next subsections, only a short and basic report is given about SVMs and I refer for a more detailed reading to the books of Vapnik (1995, 1998), Cristianini and Shawe-Taylor (2000) and Schölkopf and Smola (2002). More references to publications, books and software can also be found on the website <http://www.kernel-machines.org>.

1.2.2.1 Hyperplanes and Support Vectors

In the previous section, a brief introduction was given about linear discriminant analysis in the perspective of the perceptron and its extension towards multi-layer artificial neural networks. As described, training an ANN comes along with searching the hypothesis space for those weights best fitting the data points and, thus, with the approximation of the decision boundary between two (or more) classes. However, Vapnik (1995) states that, for the separable case, the optimal separating hyperplane separates the two classes and maximizes the distance to the closest point from either class. This concept is also better known as a maximum margin classifier where the sum of the maximum distance to either class is called the margin. By this last constraint on the optimization problem, not only a unique solution is provided for the separating hyperplane problem, but by maximizing the margin between the two classes on the training data, this also leads to better classification performance on test data (Hastie et al., 2009). In 1995, Cortes and Vapnik (1995) have introduced the SVM in which the concept of a maximum margin classifier is used for detecting those data points lying on the maximum margin boundary. These data points determine the margin and are, therefore, called the support vectors of the maximum margin classifier. Consequently, the support vectors allow for determining the optimal boundary by only taking a small number of training data into account.

Before discussing non-linear separability, let's first go back to linear discriminant analysis which is visualized in Figure 1.5. The linear discriminant can be reshaped in terms of a maximum margin classifier and the resulting plot is shown in Figure 1.7. The left plot represents the

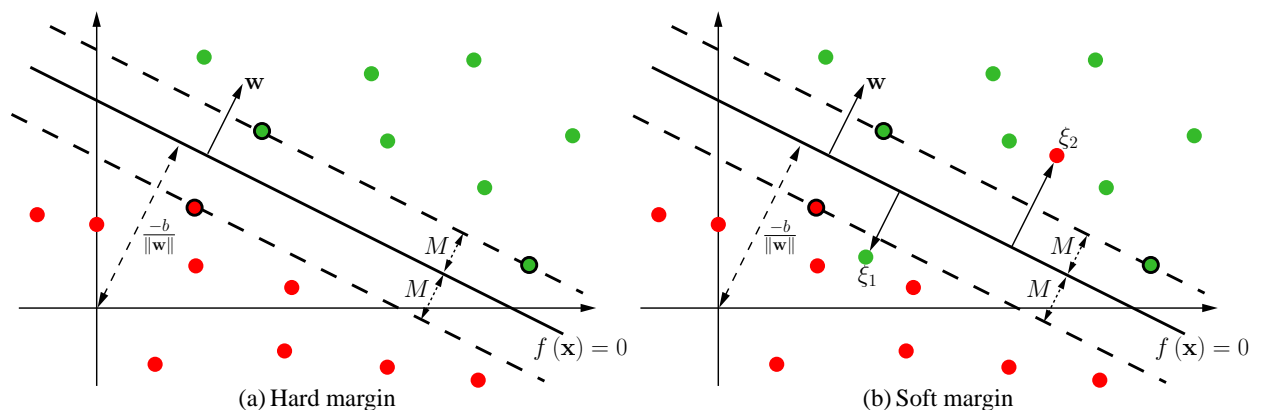


Figure 1.7: Maximum margin classifier. The linear decision boundary corresponding to $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ is visualized in a two-dimensional input space. Parallel to linear decision boundary, the two optimal boundaries are set as dashed lines and the distance $2M$ between both is called the margin. Data points lying on the dashed lines are called the support vectors and are encircled in black. The weight vector \mathbf{w} defines the orientation of the boundary, while the bias weight b defines the position of the boundary in terms of its perpendicular distance from the origin. The boundary separates the green data points (positive class) from the red data points (negative class). On the left, a linearly separable problem is visualized and the margin is denoted as ‘hard margin’. On the right, a non-linear separable problem is visualized. One point of each class is beyond its boundary and corresponds to slack variable ξ . By allowing but penalizing misclassifications in this case, the margin is denoted as ‘soft margin’.

linear separability case and the separating boundary or hyperplane is denoted by Eq. (1.3) (note that b is identical to w_0 in ANNs). For points lying on the boundary the equation $\mathbf{w} \cdot \mathbf{x} + b = 0$

is satisfied. As the classes are separable, it is possible to find a function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ with $y_i f(\mathbf{x}_i) > 0$, $y_i \in \{-1, +1\}$ the target of \mathbf{x}_i , $i = 1, \dots, N$ and N the number of data points. Thus, it is possible to determine the largest margin between the training points of both classes. As such, the margin is delineated by two hyperplanes parallel to the decision boundary and the distance between each hyperplane and the decision boundary is set to M . Do not confuse M with the same notation of the number of hidden neurons in the previous section. For this case, we are only interested in solutions for which all data points are correctly separated, meaning that $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 0$, \forall_i . Hence, it is possible to find the largest margin. The distance M of the closest point \mathbf{x}_t to the decision boundary equals

$$\frac{y_t(\mathbf{w} \cdot \mathbf{x}_t + b)}{\|\mathbf{w}\|} \quad (1.12)$$

As such, in the case of linear separability, all data points will satisfy the inequality $y_t(\mathbf{w} \cdot \mathbf{x}_t + b) \geq M$. Note that for any \mathbf{w} and b satisfying this inequality, any positive scaling satisfies them too. Hence, we can arbitrarily set $\|\mathbf{w}\| = \frac{1}{M}$. As an SVM aims to maximize the margin, this actually corresponds to

$$\min \frac{1}{2} \|\mathbf{w}\|^2, \quad (1.13)$$

$$\text{subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad (1.14)$$

This is a convex optimization problem which can be solved by quadratic programming. Herein, a quadratic function is minimized subject to a set of linear inequality constraints. With a convex optimization method, a global solution will be found (Burges, 1998; Bishop, 2006; Hastie et al., 2009).

Before further discussing the optimization problem of SVMs, let us first consider the non-linear separability case which is represented in the right plot of Figure 1.7. For overlapping class distributions, the above mentioned formulations need to be relaxed in some sense. Misclassification may be allowed but with a penalty that increases with the distance from the boundary. The hard margin is somewhat relaxed and is called ‘soft margin’ in this case. Slack variables $\xi_i \geq 0$, $i = 1, \dots, N$ are introduced. $\xi_i = 0$ for correctly classified data points, $\xi_i = 1$ for data points on the decision boundary, $0 < \xi_i \leq 1$ corresponds to points in the margin, and misclassified points correspond to $\xi_i > 1$. Regarding softening of the margin, the optimization problem as presented in Eq. (1.13) can be reformulated as

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \quad (1.15)$$

$$\text{subject to } \begin{cases} y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \forall_i \\ \xi_i \geq 0, \end{cases} \quad (1.16)$$

where the $C > 0$ is a user-defined cost parameter controlling the trade-off between the slack variable penalty and the margin. Note that for the separable case, C corresponds to ∞ . Because misclassified points correspond to $\xi_i > 1$, $\sum_{i=1}^N \xi_i$ is an upper bound on the number of

misclassified points. Hence, the parameter C is a regularization parameter controlling the trade-off between minimizing training errors and controlling model complexity (Cortes and Vapnik, 1995; Bishop, 2006; Hastie et al., 2009).

With the soft margin introduced, we can resume the discussion about the SVM optimization problem. The minimization can be reformulated as a Lagrangian formulation and Burges (1998) gives two advantages. First, the constraints in Eq. (1.16) are replaced by constraints on the Lagrange multipliers, which are easier to handle. Second, the training data will appear as dot products between vectors. So, the Lagrangian formulation, also the primal, equals

$$L_P = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i, \quad (1.17)$$

where $\alpha_i \geq 0$ and $\mu_i \geq 0$ are the Lagrange multipliers. Finding the optimal solution corresponds to minimizing L_P w.r.t. \mathbf{w} , b and ξ_i . Taking the respective derivations and setting these to zero results in

$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i, \quad (1.18)$$

$$0 = \sum_{i=1}^N \alpha_i y_i, \quad (1.19)$$

$$\alpha_i = C - \mu_i, \quad \forall_i. \quad (1.20)$$

Substituting Eqs. (1.18)-(1.20) into Eq. (1.17) results in the dual representation of the Lagrangian, also the Wolfe dual:

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j, \quad (1.21)$$

which gives a lower bound on the objective function by Eq. (1.13) and is a simpler convex optimization problem which can be solved with standard software. L_D should be maximized subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^N \alpha_i y_i = 0$. However, the solution must satisfy the set of constraints, also called the Karush-Kuhn-Tucker conditions, that for SVMs correspond to the linear constraints Eqs. (1.18)-(1.20) and

$$y_i (\mathbf{w} \cdot \mathbf{x}_i - b) - 1 + \xi_i \geq 0 \quad (1.22)$$

$$\xi_i \geq 0 \quad (1.23)$$

$$\alpha_i \geq 0 \quad (1.24)$$

$$\mu_i \geq 0 \quad (1.25)$$

$$\alpha_i (y_i (\mathbf{w} \cdot \mathbf{x}_i - b) - 1 + \xi_i) = 0 \quad (1.26)$$

$$\mu_i \xi_i = 0 \quad (1.27)$$

for $i = 1, \dots, N$. Notice that, except for the constraints, this dual formulation is identical

in both the linear and non-linear separability case (Burges, 1998; Bishop, 2006; Hastie et al., 2009).

From Eq. (1.18), it can be seen that the solution for \mathbf{w} is only comprised of non-zero coefficients α_i , that is, for those observations for which the constraints in Eq. (1.22) are exactly met, due to Eq. (1.26). These data points are called the ‘support vectors’ since α_i is only represented in terms of them alone (Hastie et al., 2009).

The support vector machine described above solves only an optimization for a linear decision function in the input feature space. However, for many real-world problems, the decision function is, however, a non-linear function of the data. So, how can the above-mentioned technique be generalized? Suppose a mapping from the input data $\mathbf{x} \in \mathbb{R}^D$ to another Euclidean space \mathcal{F} such that in this feature space \mathcal{F} the data points become linearly separable. We define the mapping function

$$\phi : \mathbb{R}^D \mapsto \mathcal{F}. \quad (1.28)$$

Thus, this mapping turns each input data point into a point in the space \mathcal{F} . In this sense, a data point is represented by its similarity to all other points in the input space. This idea of feature mapping is visualized in Figure 1.8 The optimization problem as stated in Eq. (1.21)

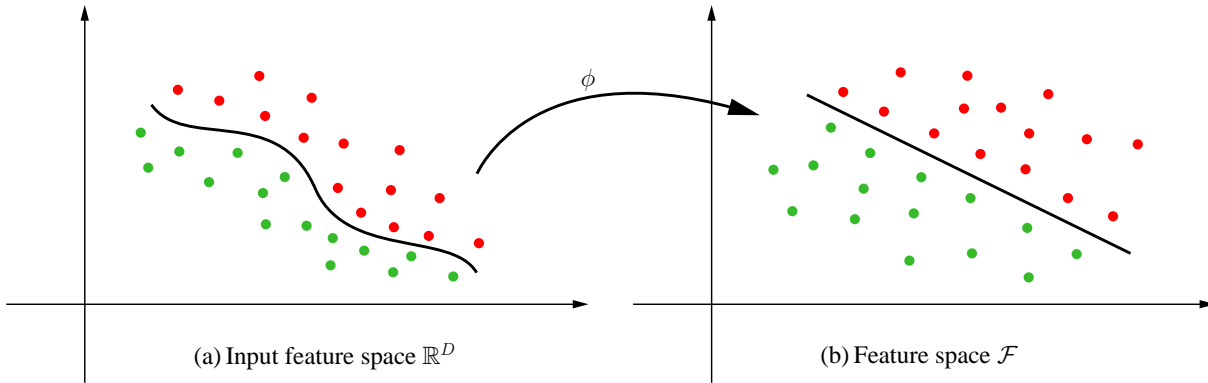


Figure 1.8: Concept of feature mapping. The function ϕ maps the non-linear separable input data into a feature space \mathcal{F} that allows linear separability.

can be represented in terms of the feature maps ϕ . But, how to find a good feature mapping function? To circumvent this problem, the so-called ‘kernel trick’ can be applied, by which a kernel function K is defined

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \quad (1.29)$$

that computes the inner products in the feature space \mathcal{H} . In feature space \mathcal{H} , the dual becomes

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j). \quad (1.30)$$

From Eq. (1.18), the decision boundary as presented in Eq. (1.3) can, ultimately, be written as

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w} \cdot \phi(\mathbf{x}) + b \\ &= \sum_{i=1}^n \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i), \end{aligned}$$

which indeed proves to be linear in feature space \mathcal{H} . Popular kernel functions are:

$$\begin{aligned} \text{Linear kernel:} \quad & K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j \\ d^{\text{th}} \text{ degree polynomial kernel:} \quad & K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \cdot \mathbf{x}_j)^d \\ \text{Radial basis function (RBF) or Gaussian kernel:} \quad & K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2} \\ \text{Neural network:} \quad & K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\kappa_1 \mathbf{x}_i \cdot \mathbf{x}_j + \kappa_2) \end{aligned}$$

Notice that the tanh kernel corresponds to a particular two-layer sigmoidal ANN. Note also that the role of the cost parameter C is more clear in feature space \mathcal{F} . A large value of C discourages any positive ξ_i and leads to an overfitting boundary in the input space. A small value of C will encourage a small value of $\|\mathbf{w}\|$ (due to smaller α_i values) which causes the boundary to be smoother (Burges, 1998; Schölkopf and Smola, 2002; Bishop, 2006; Hastie et al., 2009). In this study, we focused on the linear and the RBF kernel.

When finally arrived at the solution of SVM classification, the SVM can now easily be tested by determining on which side of the decision boundary a given test point lies and by labelling it with the corresponding label (Burges, 1998).

1.2.2.2 Towards the Multi-class Setting

In the previous section only two classes are considered while real-world data sets mostly cover multiple classes. In this section the multi-class SVM classification extension is reported. A general discussion is given that can also be used for many other classifiers than SVMs. Discussion is done in this section because, fundamentally, the SVM is a two-class classifier.

Different methods are proposed to solve multi-class classification problems (Bishop, 2006). Two main approaches are commonly used: a single optimization approach and the ensemble method (see also Subsection 1.2.3.1.1). In the first approach, a single objective function is trained for all classes based on maximizing the margin between each class and the other classes. The optimization problem in this approach is often complex and training is usually (very) slow. This approach is tackled by different researchers such as Weston and Watkins (1998), Crammer and Singer (2001), Hsu and Lin (2002) and Guermur (2007). The second approach of the ensemble method corresponds to combining different two-class or binary classifiers to build a multi-class classifier. Two popular methods are

1. one-versus-the-rest (also one-versus-others or one-versus-all): K binary classifiers f_k are trained such that each classifier f_k distinguishes class k from the other $K - 1$ classes, with $k = 1, \dots, K$ (Rifkin and Klautau, 2004). Considering the set of K estimated discriminant functions for prediction, this approach would, however, lead to regions of

the input space that are ambiguously classified. This problem can be resolved by deriving a global rule from the K classifiers stating that a test point is assigned the label of the class for which the highest value of the discriminant function is found, or

$$f(\mathbf{x}) = \arg \max_k f_k(\mathbf{x}). \quad (1.31)$$

2. one-versus-one: $\frac{K(K-1)}{2}$ different binary classifiers are trained on all possible pairs of classes (Hastie and Tibshirani, 1998; Fürnkranz, 2002). In this approach, test points are assigned the label of the class which results in the highest number of votes, also majority vote. This approach also results in ambiguities in the globally defined decision boundary and requires significantly more training and prediction time than the one-versus-others approach (Bishop, 2006). This problem can, however, be solved substantially faster by combining the $\frac{K(K-1)}{2}$ binary classifiers in a directed acyclic graph (DDAG) or by decomposing the multi-class classification task by means of binary tree classifiers (see also Subsection 1.1.2.2; Platt et al., 2000; Schwenker and Palm, 2001; Lee and Oh, 2003; Cheong et al., 2004; Fei and Liu, 2006; Xia et al., 2007).

A last approach is that of the construction of sets of binary classifiers by means of error-correcting output codes (Dietterich and Bakiri, 1995; Allwein et al., 2000; Crammer and Singer, 2002). A nice overview paper about SVM multi-class classification reviewing most of the above-mentioned papers is that of Hsu and Lin (2002).

1.2.2.3 Properties

SVMs are state-of-the-art in many application domains and witness a boost of number of applications. When focusing on the data and algorithms, the advantages of SVMs are somewhat the same as ANNs. Compared to ANNs, SVMs can also handle real- and binary-valued input data and training can also be parallelized. However, in contrast to ANNs, SVMs have several distinct advantages (Burges, 1998; Schölkopf and Smola, 2002):

- Complex nonlinear data relations can be expressed linearly in a high dimensional feature space, which allows for simple geometry and linear algebra
- The kernel trick allows to compute dot products in high dimensional feature spaces using simple functions defined on pairs of input data
- Due to the convex optimization, SVM training always finds a global solution, in contrast to the ANN where usually many local minima usually exist
- A cost parameter, kernel function and according parameters have to be chosen and optimized, in contrast to the many ANN parameters that need to be optimized
- The decision boundary can be reconstructed by only a small sample of training points
- SVMs can be regarded as a high performance classifier with an accuracy often higher than that obtained by ANNs

Also, due to the nature of kernels, very problem-specific kernels can be developed such as string kernels, text kernels, tree kernels, position-specific kernels, . . . (Vert, 2002; Leslie et al.,

2004; Shawe-Taylor and Cristianini, 2004), and multiple heterogeneous data sources can be combined in a straightforward way by combining the kernel functions defined on the different data types, which is known as data fusion (Lanckriet et al., 2004).

However, disadvantages also exist (Mitchell, 1997; Schölkopf and Smola, 2002; Bishop, 2006; Hastie et al., 2009):

- Limitations in speed can result from very large data sets that lead to a massive number of support vectors, which upscales quadratic programming. Different methods have been proposed to alleviate this problem
- High noise data leads to many support vectors, resulting in the previous disadvantage
- SVMs are fundamentally binary classifiers, implying that the multi-class classification task comes with an additional computational load
- SVMs are sensitive to overfitting when the parameter values are not properly optimized
- SVMs cannot handle categorical variables. Dummy features resulting from binary encoding may offer a solution here
- Standardization of the input values is necessary. It is best to standardize all inputs to have a zero mean and a standard deviation of one
- SVMs cannot handle missing values

1.2.2.4 Software

SVM implementation, training and testing is done using LibSVM 2.86 and BSVM 2.06 (Chang and Lin, 2001; Hsu and Lin, 2002). Wrapper code is written in JAVA for automation of data import and conversion, parameter optimization and statistical analysis of the classification results. LibSVM is based on the one-versus-others approach while BSVM implements the single optimization solution as proposed by Crammer and Singer (2001). In this work, the probability outputs were considered for further statistical analysis.

Numerous SVM software packages, toolboxes, wrapper codes, ... exist and SVMs are nowadays implemented in (almost) every statistical and mathematical software package. A list of all these implementations would also be a section on itself. A nice overview is given on the websites <http://www.kernel-machines.org> and <http://www.support-vector-machines.org>.

1.2.3 Random Forests

One of the techniques that is gaining a lot of popularity is random forests. While theoretical research has already been performed for many years, the number of RF applications is only increasing in the last ten years. In the first sections, the basic properties of RFs are discussed, followed by a description of the RF technique. For a more detailed reading I refer to the papers and webpage of Breiman (2001, 2002, 2004).

1.2.3.1 An Ensemble of Trees

1.2.3.1.1 Ensemble Methods

Ensemble methods, also called combining models or committees, are learning algorithms that construct a set or ensemble of many individual classifiers that are diverse and yet accurate (Hastie et al., 2001; Bishop, 2006). Or, the individual classifiers make different errors on new data points with an error rate that is better than random guessing. These classifiers, which are also called base learners, are combined to classify new data points. Classification decisions are obtained by taking a weighted or unweighted vote of their predictions. It is well known that ensembles are often much more accurate than the constituting individual classifiers. Ensemble methods differ in the way the base learner is learned and combined (Dietterich, 1998, 2000; Biau et al., 2008). Different explanations are given for the success of voting classifiers and their improved performance. One major explanation is based on the bias-variance decomposition (see also Subsection 1.1.1). The whole idea behind averaging over many classifiers is to reduce the variance term of the prediction error. Reducing the variance will consequently lead to a lower expected error (Schapire et al., 1998). Dietterich (2000) denotes that it is often possible to construct very good ensembles because of three fundamental reasons:

1. **Statistical.** A learning algorithm can be viewed as searching the hypothesis space \mathcal{H} to identify the best hypothesis. The statistical problem arises when the amount of training data is too small compared to the size of \mathcal{H} . Consequently, the learning algorithm finds many different hypothesis giving the same accuracy. By the construction of an ensemble of these classifiers, the algorithm can average their votes and reduce the risk of choosing the wrong classifier.
2. **Computational.** Many learning algorithms work by performing a local search with the risk of getting stuck in a local optimum. Examples are ANNs and decision trees. Building an ensemble by running a local search from many different starting points may provide a better approximation to the true unknown function f .
3. **Representational.** In many machine learning applications, the true function f cannot be represented by any of the hypotheses in \mathcal{H} . With a finite training sample, these algorithms explore only a finite hypothesis set and stop searching when a hypothesis is found that fits the training data. By weighted sums of hypotheses drawn from \mathcal{H} , it may be possible to expand the space of representable functions.

Ensembles can be constructed by many different methods (Dietterich, 2000). General methods are

- **Training set manipulation:** several base classifiers are constructed by a different subset of the training set. This technique works especially well for unstable learning algorithms, meaning that small changes in the training data results in major changes of the output. Examples of unstable learning algorithms are decision trees, neural networks and rule learning algorithms. Popular training set manipulation methods are bagging, cross-validation and boosting (Breiman, 1996a; Freund and Schapire, 1996).

- Feature manipulation
- Output target manipulation: or manipulating of the output values given by the learning algorithm. One example is the error-correcting output coding of Dietterich and Bakiri (1995) in which an ensemble of classifiers is constructed in which the l classes are randomly partitioned into two subsets. Each classifier predicts one of the two subsets and each class of the predicted subset receives a vote. The class with the highest number of votes is selected as the prediction of the ensemble.
- Randomness injection
Randomness can be injected into the classifiers in different ways: initial weights in ANNs (see also Section 1.2.1), random choice of features, noise on the data, ... (Dietterich and Kong, 1995; Raviv et al., 1996)
- Bayesian voting: this ensemble method consists of all the hypotheses in \mathcal{H} , each weighted by its posterior probability (Dietterich, 2000)

1.2.3.1.2 Decision Trees

Decision trees are tree-based methods which partition the feature space into a set of rectangles, which are aligned to the axes in feature space, and then fit a model in each one. Subdivision of the feature space is achieved by a sequential decision making process that, in the basic algorithm, corresponds to recursive binary partitioning. That split leading to the best fit is chosen as final split. Each split corresponds to one feature, a threshold and the training data associated with the node to evaluate. An example of a decision tree is visualized in Figure 1.9. A popular metric used by many decision tree algorithms is information gain or impurity. Specifically, the feature leading to the largest information gain or impurity decrease in the classification/regression framework is chosen as final splitting criterium. The tree node splitting process continues until a stopping rule is applied. Examples of stopping rules are: the evaluation of the tree performance on an independent test set, the application of a statistical test for expansion or pruning of the tree, a measure of the tree's complexity (e.g. as the presence of Q nodes in the tree), the inclusion of all features along a certain path through the tree, the correspondence of the training examples associated with the node having the same feature value, ... In general, a learned tree can be represented as a set of if-then rules, hereby improving human readability. When to halt splitting is a critical question in the decision tree method, as too many nodes may lead to overfitting and too few nodes to a high performance error and bad generalization. Therefore, the resulting tree can be pruned, which corresponds to collapsing internal nodes. Which nodes to prune is generally based on an independent validation set. Pruning allows to further balance the decision tree prediction error against a measure of model complexity (i.e. regularization). Ultimately, the tree classifier becomes more optimized. In their study, Dietterich and Bakiri (1995), however, reported that pruning does not necessarily improve multi-class classification performance and that the merit of pruning varies from one domain to another. Both the stopping rule, regarded as pre-pruning, and pruning allow the decision tree to overcome the problem of overfitting. For prediction, any test point is put down the tree at the root node and following a path towards a specific leaf node according to the decision criteria at each node. For

classification this will result in a class label, for regression in a numerical value (Breiman et al., 1984; Quinlan, 1993; Mitchell, 1997; Hastie et al., 2001; Bishop, 2006).

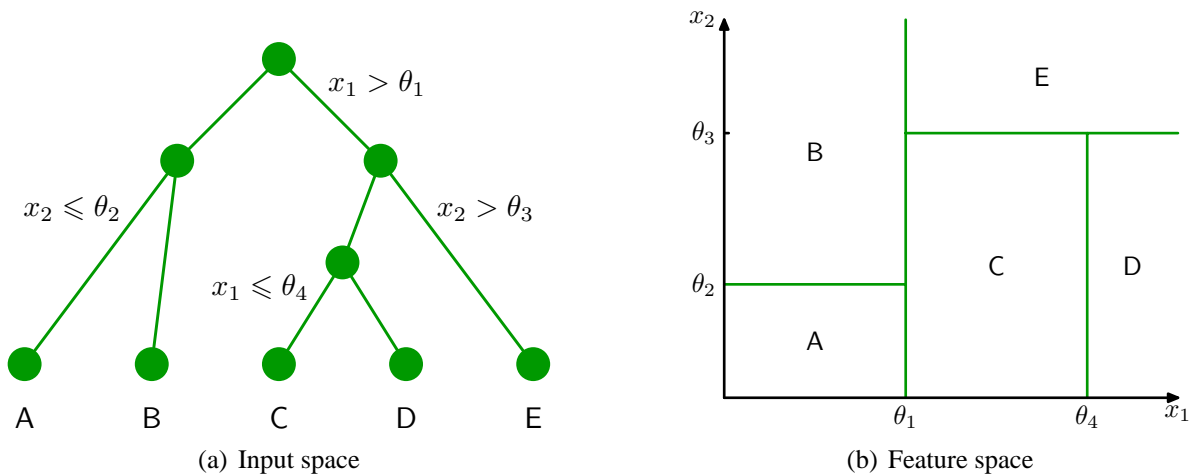


Figure 1.9: Decision tree in a setting with five classes A, B, C, D and E and two features x_1 and x_2 . Four thresholds θ_r are used as splitting criteria with $r = 1, \dots, 4$. The decision tree is visualized both in input space (left figure) and in feature space (right figure) (Bishop, 2006)

Decision trees are conceptually simple yet powerful. As mentioned above, one of the key advantages is their human interpretability. Other advantages are the ability of handling categorical features and missing values, their robustness to noisy data, etc. (Hastie et al., 2001; Bishop, 2006). Nonetheless, different problems are associated with decision trees. One major problem is their high variance. Often a small change in the data, even the addition or removal of a single data point, can result in a very different series of subsequent splits, leading to significant subtree structures below each node. Ultimately, the entire tree is altered. The main reason of the instability of decision trees is the hierarchical nature of the process: the effect of an error in the top split is propagated down to all splits below it (Dietterich and Kong, 1995; Hastie et al., 2001; Witten and Frank, 2005). The variance can possibly be reduced by pruning (see also above), by allowing soft thresholds for node splitting or by voting over an ensemble of decision trees (Dietterich and Kong, 1995). Next to this, the different splits are aligned with the axes of the feature space which may lead to suboptimal solutions. Solving this problem would involve a large number of additional splits (Bishop, 2006). Examples of decision tree algorithms are classification and regression trees (CART), ID3, C4.5, C5.0/See5 and ASSISTANT (Quinlan, 1993; Mitchell, 1997; Hastie et al., 2001; Bishop, 2006).

1.2.3.2 Random Forests

The concept of ensembles and decision trees converge in the RF technique, which is described in the following paragraphs. First, a short mathematical description is given, followed by the two main principles behind RFs: bagging and random split prediction. Finally, the algorithm behind RFs is explained.

1.2.3.2.1 A General Description

In 2001, Breiman (2001) has proposed a new machine learning technique consisting of an ensemble of classification and regression trees, better known as random forests. A RF classifier can be defined as a classifier consisting of a collection of tree-structured classifiers $f(\mathbf{x}, \Theta_v)$, $v = 1, 2, \dots, V$, with Θ_v independent and identically distributed (i.i.d.) random vectors, and where each tree casts a unit vote for the most popular class at input \mathbf{x} (Breiman, 2001, 2002).

The accuracy of RFs can be determined by a margin function and by the strength and correlation of the different base classifiers. A short mathematical representation is given for these three parameters. For a detailed mathematical description of the structure behind RFs, I refer to the paper of Breiman (2001).

Given an ensemble of classifiers $f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_V(\mathbf{x})$, with the training set D drawn at random from the distribution of the random vector Y , a margin M is defined by

$$M(D, Y) = \text{avg}_v I(f_v(D) = Y) - \max_{j \neq Y} \text{avg}_v I(f_v(D) = j) \quad (1.32)$$

where I is the indicator function. Or, the margin measures the extent to which the average number of votes for the right class exceeds the average vote for any other class. The larger the margin, the more confidence in the classification. As a result, the generalization error GE is given by

$$\text{GE} = P_{D,Y}(M(D, Y) < 0) \quad (1.33)$$

where the subscripts D, Y indicate the probability over the D, Y space. From the Strong Law of Large Numbers¹ and the tree structure follows that, for an increasing number of trees and for surely almost all vectors Θ_v , the generalization error converges to the probability

$$P_{D,Y}(P_{\Theta}(f(D, \Theta) = Y) - \max_{j \neq Y} P_{\Theta}(f(D, \Theta) = j) < 0) \quad (1.34)$$

Proof of the convergence is reported in Appendix I of the RFs paper (Breiman, 2001). This property shows that RFs do not overfit as more trees are added, but produce a bound on the generalization error. An upper bound of the generalization error can be derived, in terms of two parameters that are measures of the strength s of the individual classifiers and of the correlation ρ between them:

$$\text{GE} \leq \bar{\rho}(1 - s^2)/s^2, \quad (1.35)$$

with $\bar{\rho}$ the mean value of the correlation and s a measure of the margin. Thus, a RF classifier consisting of low-correlated trees with a high individual strength results in optimal generalization and a high accuracy (Breiman, 2001). As stated above, Breiman (2001) shows that the generalization error of RFs converges to a certain value. Nonetheless, Hastie et al. (2009) importantly remark that the limit of the number trees going to ∞ can, however, overfit the data as

¹If a certain chance experiment is repeated an unlimited number of times under exactly the same conditions, and if the repetitions are independent of each other, then the fraction of times that a given event A occurs will converge with probability 1 to a number that is equal to the probability that A occurs in a single repetition of the experiment (Tijms, 2004).

the average of the fully grown trees can result in too rich a model and can incur unnecessary variance. This can be addressed to the problem stated above that fully grown decision trees have a tendency to show large variances. An ensemble of these fully grown trees may, therefore, result in a classification model that is too complex, with a consequence of unnecessary variance. The authors state that the trade-off between the number of relevant and noise variables may lead to an increase of the misclassification error.

1.2.3.2.2 Bagging

When explaining bagging, bootstrapping needs to be touched first. Bootstrapping should be seen as a large class of methods that resample from the original data set. Therefore, these methods are also called resampling methods. Bootstrapping is a popular method for assessing the accuracy of a parameter estimate or a prediction. In 1979, Efron (Department of Statistics, Stanford University, USA) came up with the bootstrap for i.i.d. observations, which resamples the data with replacement (Efron and Tibshirani, 1993; Chernick, 2008). A repeated number of bootstrap samples are selected and form replicate data sets, each consisting of P data points, drawn at random with replacement from a given data set. As such, the distribution of the replicate set of bootstrap samples approximates the underlying distribution of the given data set and assigns to each sample a probability to be selected of $\frac{1}{P}$ (Efron and Tibshirani, 1993; Breiman, 2001; Hastie et al., 2001; Chernick, 2008). Also important to note is that the expected proportion of values not represented in a particular bootstrap sample equals $(1 - (1/P))^V$, with V the number of bootstrap samples. Considering a large V , this proportion approximates $e^{-1} \approx 0.368$, meaning that about approximately 36.8% of the data is left out in any bootstrap sample (Efron and Tibshirani, 1993; Hastie et al., 2001).

In 1996, Breiman (1996a) proposed a new machine learning method called bagging predictors. Bagging or bootstrap aggregation is a method to generate multiple bootstrap replicates of the original data set in order to construct different versions of a predictor, which are used to get an aggregated predictor. Bootstrap replicates are generated by uniformly drawing n examples from the data set with replacement (training data) and the remaining data points are called ‘out-of-bag’ (test data) and are used for predicting the corresponding class by a majority vote. Breiman suggested the bagged predictor because bagging averages the prediction over the collection of bootstrap replicates or ‘bags’, thereby reducing its variance. However, bagging only generates diverse classifiers if the learning algorithm is unstable, meaning that small changes in the training set cause large changes in the learned classifier (Breiman, 1996a; Hastie et al., 2001). In light of the variance reduction, it is important to note that bagging trades a light increase in bias for a major decrease in variance to yield significant improvement in performance (Dietterich and Kong, 1995). The prediction and generalization error of bagging is extensively studied in relation to the out-of-bag estimation and the bias-variance decomposition in different studies (Breiman, 1996b; Tibshirani, 1996; Wolpert and MacReady, 1999; Bylander, 2002).

Suppose we need to fit a model to our training data D , obtaining the prediction $f(\mathbf{x})$ at input \mathbf{x} . For each bootstrap sample D_v , $v = 1, 2, \dots, V$, we fit our model, giving prediction $f_v(\mathbf{x})$

(Hastie et al., 2001). The bagging estimate is defined by

$$f_{bag}(\mathbf{x}) = \frac{1}{V} \sum_{v=1}^V f_v(\mathbf{x}) \quad (1.36)$$

Bagging is used in RFs for two main reasons (Breiman, 2001):

- The use of bagging enhances accuracy when random features are used
- Bagging can be used to give estimates of the generalization error of the combined ensemble trees as well as estimates for the strength and correlation.

1.2.3.2.3 Random Split Prediction

As an approach to solve the high variance problem related to decision tree algorithms, Dietterich and Kong (1995) proposed an alternative approach for the generation of an ensemble by randomizing the internal decisions made by a base classifier. This method is also known as randomization and was first introduced in a simple form by Kwok and Carter (1990). In the case of decision trees, the best node split is randomly selected out of the z best splits. In their paper, Dietterich and Kong (1995) applied this procedure for the construction of 200 trees with $z = 20$ and found that, in their particular case, variance dropped while the bias remained unchanged. This, in contrast to the bagging procedure as proposed by Breiman (1996a). Dietterich (1998) further exploited this randomization approach by applying the technique on 33 UCI repository data sets. Performance evaluation was done by comparing randomization with bagging and boosting. Interestingly, when introducing classification noise (data examples with incorrect class labels), Dietterich reported that randomization gives similar results as bagging for the 33 cases considered. When dealing with noise, it is also easy to understand that bagging still performs very good as, next to the approximately 36.8% data left-out feature of bagging, classification noise will lead to large changes in the learned classifier and, thus, to more diverse classifiers, in contrast to randomization (Dietterich, 1998). In general, Dietterich and Kong (1995) concluded that randomization and bagging have resulted in similar performances with a light slight preference to randomization in low noise settings. With added classification noise, bagging showed to be the best method. It is also interesting to note that the effect of randomizing the number split variables does not depend as much on the training set size as does bagging. With an infinite large training set, randomization will still produce a diverse set of decision trees, while the effect of bagging on the error rate of the decision tree will be null (with an infinite sample size the tree algorithm will always grow the same tree) (Dietterich, 1998). Breiman integrated random split prediction in RFs to improve its accuracy. Hereby, the correlation between the different trees is minimized while maintaining strength. As a result, Breiman (2001) stated that RFs have an accuracy compared to that of the Adaboost algorithm of Freund and Schapire (1996).

1.2.3.2.4 A Look Inside the Forest

As stated above, RFs grow many classification or regression trees. The size of each tree in the forest is grown based on N training data points randomly sampled from the original data set and is evaluated by the remaining test data points. No pruning is performed and sampling of the training data is done by bagging. Each training set is about two-third of the size of the original data set. Evaluation of the accuracy of the grown tree by the out-of-bag data points results in the so-called out-of-bag error estimate. Randomly sampling data sets to grow the different trees of the forest is one aspect of the randomness of RFs, the second aspect is random split selection. When D features are present in the original data set, z features ($z \ll D$) are sequentially randomly sampled out of D to split each node of the tree. The final split is determined by the best split based on the z randomly sampled features. Note that z is held constant during the growth process of the forest. Each tree is maximally extended and no pruning is performed.

Evaluation of the classification accuracy of a RF classifier is based on the V out-of-bag data sets. Each out-of-bag data point \mathbf{x}_i ($i = 1, \dots, N$) resulting from randomly sampling for the construction of the v^{th} tree is put down the v^{th} tree to get a classification value, $v = 1, \dots, V$. Take J to be the class that got most votes every time case \mathbf{x}_i was out-of-bag. The proportion of times that J is not equal to the true class of n averaged over all cases is the out-of-bag error estimate, which is proved to be unbiased by many tests, given a large number of trees that have been grown Breiman (2001); Liaw and Wiener (2002). This algorithm is also described in Algorithm 2.

Algorithm 2 Random forests algorithm for classification

Require: Set the number of classification trees to be grown: T

Require: Set the number of features to split the tree nodes: Z

```

1: for  $t = 1 \rightarrow T$  do
2:   Draw a bootstrap sample from the original data set
3:   Grow an unpruned tree
4:   for all Tree nodes do
5:     1. Randomly sample  $Z$  features
6:     2. Choose the best split
7:   end for
8:   Predict out-of-bag data using the grown tree
9: end for
10: Aggregate the out-of-bag predictions:
11: for all data points  $\mathbf{x}_i$  in original data set ( $i = 1, \dots, N$ ) do
12:   for all  $O$  : out-of-bag data sets do
13:     if  $\mathbf{x}_i \in O$  then
14:       Determine predicted class label of  $\mathbf{x}_i$ 
15:     end if
16:   end for
17:   Final class label of  $\mathbf{x}_i$  determined by majority vote
18: end for
19: Calculation of error rate

```

1.2.3.3 Properties and Features

The two main properties of RFs are clearly the calculation of the out-of-bag error estimate that is an unbiased estimator of the classification error and the fact that RFs have low tendency to overfit. However, RFs have different other interesting features (Breiman, 2001, 2004; Hastie et al., 2009):

- Due to bagging together with the randomization effect in data selection and in the number of split variables, RFs gives excellent accuracy compared to many other algorithms
- RFs runs efficiently on large data sets
- Thousands of features can be handled. It is, however, important to note that this feature has to be interpreted carefully. Amaratunga et al. (2008) reported that, due to the small number of informative features in DNA microarray data, the performance of RF classifiers declines significantly. The main reason for this performance reduction can be assigned to the many low accuracy base classifiers in the forest. To tackle this problem, the authors have upgraded the random split prediction procedure of RFs by performing weighted random sampling of the different features at each node. Herein, less informative features are less likely to get selected, which increases the percentage of trees containing more informative features.
- The importance of each feature can be calculated. To estimate the importance of each variable, RFs looks how much the prediction error increases when the out-of-bag data for the corresponding variable is permuted while the other variables are left unchanged. Suppose D features. After each tree is constructed, the values of the d^{th} feature in the out-of-bag examples are randomly permuted and the out-of-bag data is rerun down the corresponding tree. This is repeated for $d = 1, \dots, D$. The majority of out-of-bag class votes for each data point with the noised d^{th} feature is compared with the true class label to calculate the misclassification rate. The percentage increase in misclassification rate is calculated and compared to the out-of-bag rate with all features intact. Next, a raw importance score and a z -score is calculated. By subtracting the number of votes for the true class in the out-of-bag data with feature d permuted from the number of votes for the correct class in the original out-of-bag data, the raw importance score for feature d is calculated by averaging the resulting difference over all trees in the forest. When dividing the raw importance score by its standard error, a z -score is retrieved. Recently, Strobl et al. (2008) reported, however, that variable importance is biased towards correlated predictor variables or features and proposed an alternative permutation strategy for variable importance evaluation.
- RFs include methods for
 - the estimation of missing values (while maintaining accuracy)
 - balancing the error in imbalanced data sets. RFs may suffer from the curse of learning from extremely imbalanced data sets (Chen et al., 2004). Different approaches exist to tackle the imbalance problem (see Subsection 1.1.3). Together with Breiman, Chen et al. (2004) proposed two ways to deal with this problem. A cost-sensitive approach by weighting the different classes with an according value

was compared with down-sampling of the majority class. They called the respective methods: weighted Random Forest and balanced Random Forest. The first method is currently integrated in the RFs software package. This feature is not used in the present study.

- RFs computes proximities between pairs of cases that can be used in clustering and to detect outliers. Proximity computation is performed following tree growth. All data, bag and out-of-bag, are put down the tree. If two different cases are in the same terminal node, their proximity is increased by one. At the end, the proximities are normalized by dividing them by the forest size. Hereby, it is possible to show how the different features relate to the classification. With respect to outlier detection, an outlier in class k can be seen as a case whose proximities to all other class k cases are small
- RFs can also handle unlabeled data, leading to clustering
- RFs can handle categorical values
- The interpretability of RFs is higher than that of ANNs and SVMs

1.2.3.4 Software

The RFs software version 5.1 is freely available under the GNU General Public License on the website of Breiman (Breiman, 2004). RFs is written in extended Fortran 77 and can be executed following compilation by the free `g77` compiler. A JAVA package was written for automation of data import and conversion, parameter optimization and statistical analysis of the classification results, and for automatic compilation and execution of the custom Fortran code.

RFs is also freely available in the `randomForest` R Package by Liaw and Wiener (2002) on the CRAN website (Liaw, 2009) and in the WEKA data mining software (Witten and Frank, 2005, 2009). Next to this, a parallel extension of the RFs algorithm is also developed (Topić, 2004). Even though the program is still freely available for download on the website of Breiman and in different data analysis packages, RFs is exclusively licensed to Salford Systems for the commercial release of the software. RFs is trademarked as `RFTM`, `RandomForestsTM`, `RandomForestTM`, `Random ForestsTM` and `Random ForestTM`.

1.3 Model Evaluation

Once a classification model has been trained, it should be evaluated for its prediction performance. To prevent biased outcomes, this evaluation should be performed with an independent test set. In many cases, the resulting estimates are continuous and express how well the respective test instance belongs to one of the implemented classes. In RFs, this measure is a probability estimates, while for ANNs and SVMs this is not the case. `LibSVM` and `BSVM` do, however, allow the calculation of probability estimates for SVMs, and the estimates of ANNs with a sigmoid activation function on the output neurons can also be regarded as a probability. A well-known problem in model evaluation is the question of how to analyze this kind of results. A very straightforward decision-making technique is thresholding where a threshold is set somewhere in the interval of the possible outcomes. A typical value used is the center of the

interval, for example 0.5 in the interval $[0,1]$ or 0 in the interval $[-1,1]$ (the latter for example in the case of an ANN with a tangent sigmoid activation function). However, when recalling to the section on the class imbalance problem (see also Subsection 1.1.3), it becomes clear that in case of skewed class distributions this threshold will result in biased conclusions. To illustrate this phenomenon, suggest a two-class data set including 10% positive data and 90% negative data. Depending on how well the data points of both classes are separated, a trained classifier will tend to classify positive examples as negative. Or, the outcomes of the positive examples will tend to have lower values that are near to the values of the negative examples. Consequently, setting a threshold in the center will result in a large number of false negative results. An alternative approach is the winner-take-all rule. By this rule, the test instances receive the class label corresponding to the highest output value. The disadvantage of this technique is, however, the loss of the output value ratios which can be very informative. Because we were confronted with data sets with multiple classes and skewed class distributions, it was hard to set convenient thresholds accounting for the skewness and setting thresholds at the center of the respective intervals will result in too biased results. For this reason, we further only considered the winner-take-all rule. In case of ties, one of the labels was chosen for further analysis but all labels were included in the final test reports.

1.3.1 Confusion Matrix

Constructing a confusion or contingency matrix is a very popular method for model evaluation. Construction of this matrix is based on the predicted class labels as set by a certain threshold or by the winner-take-all rule. First, we will focus on the two-class problem in which test instances are labelled either positive or negative. Hence, a confusion matrix summarizes the predictions by a classifier by reporting the number of

- true positives (TP): positive instances that are predicted as positive
- false positives (FP): negative instances that are predicted as positive
- true negatives (TN): negative instances that are predicted as negative
- false negatives (FN): positive instances that are predicted as negative

An example is given in Table 1.2. From this matrix different performance measures can be cal-

		Predicted class	
		positive	negative
True class	positive	TP	FN
	negative	FP	TN

Table 1.2: Confusion matrix. Two-by-two matrix summarizing the predictions of a two-class classification.

culated. A summary of the measures considered and the corresponding mathematical expressions are given in Table 1.3. A more complete overview is given in the cited papers (Fielding and Bell, 1997; Baldi et al., 2000). Note that when no TP and FP results are obtained, precision

Measure	Formula
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
Sensitivity True positive rate (TPR)	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
F-score	$\frac{2 \times \text{sensitivity} \times \text{precision}}{\text{sensitivity} + \text{precision}}$
False positive rate (FPR)	$\frac{FP}{TP+FP} = 1 - \text{specificity}$

Table 1.3: Performance measures. Overview of several performance measures as calculated from a two-class confusion matrix.

resolves in a value equal to inf. For the F-score, this case together with a denominator of zero will also resolve in a value of inf. Regarding accuracy, different formulations exist. For example, also the fraction $\frac{TP}{TP+FP+TN+FN}$ is known as accuracy. Note that this is also the way of how RFs calculate the out-of-bag error.

As mentioned before, in this study we were confronted with multi-class data sets. It is possible to extend the idea of a confusion matrix towards a multi-class setting. With the predicted labels of a multi-class test set, it is possible to construct a multi-class confusion matrix where the number of rows and columns correspond to the number of classes in the data set. On the main diagonal, the number of TP results are given while the non-diagonal cells contain the number of misclassifications. Each non-diagonal cell represents the number of FN (at row level) or FP (at column level), respective to the class under consideration. An example is given in Table 1.4. Different approaches can be used for the analysis of this multi-class confusion

		Predicted class		
		Class 1	...	Class K
True class	Class 1	TP	...	FN
	⋮	⋮	⋱	⋮
	Class K	FN	...	TP

Table 1.4: Multi-class confusion matrix. K -by- K matrix summarizing the predictions of a multi-class classification, with K the number of classes. The cells on the main diagonal represent the number of TP while the non-diagonal cells correspond to the errors made.

matrix. The accuracy measure can be calculated over the complete data set by the percentage of TP over the complete test set. The other measures defined in Table 1.3 cannot be calculated over the complete test set, though, these can be calculated for each class separately. The multi-class confusion matrix can be decomposed in K two-class confusion matrices in which each class is considered and is evaluated in a one-versus-others approach against all other classes. For each of these confusion matrices it is possible to calculate the performance measures as defined in Table 1.3. A measure of the classifier performance for the multi-class classification problem can simply be given by the average of a given performance measure over the different classes

(Hand and Till, 2001; Fawcett, 2006). Note that classes or species resulting in a precision and/or F-score with a value of inf are not taken into account for this averaging. As in this study each class was of even importance the use of a weighted average was not further considered.

1.3.2 ROC Curve

In the seventies, receiver operating characteristic (ROC) curves were introduced in signal detection theory for the analysis of the tradeoff between hit rates and false alarm rates of classifiers. In later years, ROC analysis found its way in medicine for the evaluation of diagnostic algorithms. About twenty years ago, ROC analysis was introduced in the machine learning community for the evaluation and comparison of algorithms as an alternative to the popular accuracy measure (Hanley and McNeil, 1983; Fawcett, 2006). Where decision making becomes problematic when an arbitrary threshold needs to be set on the continuous outputs of the test instances, ROC analysis circumvents this problem by providing a two-class classification evaluation measure following a simple algorithm. ROC curves are two-dimensional graphs in which the TPR is plotted on the Y axis and the FPR is plotted on the X axis. When defining a particular threshold for analyzing the test set predictions, a point can be drawn in this two-dimensional ROC space. Several points in ROC space are important to note. The lower left point (0,0) represents the strategy of never achieving a positive classification, while the opposite is true in the point (1,1). The point (0,1) represents perfect classification and the diagonal line $y = x$ represents random guessing. As mentioned before, a point corresponds to a particular threshold set on the continuous outputs corresponding with the different test instances. A ROC graph is easily constructed by simply sorting the test instances by their continuous outputs and by moving a threshold between each pair of output values. As a result, a ROC curve is generated between the two points (0,0) and (1,1) (Lasko et al., 2005; Fawcett, 2006). As a measure of classification performance, the use of the area under the ROC curve (AUC) was proposed by Bradley (1997). The AUC represents the probability that a classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. As the underlying probability distribution is not known, the AUC is calculated on a finite set of data points and is, thus, only an estimator of the true probability. Interesting, this probability of correct ranking is equivalent to the estimation by the Wilcoxon-Mann-Whitney test statistic (Bradley, 1997; Hand and Till, 2001). An example of a ROC curve for a two-class classification experiment is visualized in Figure 1.10 (see also Subsection 4.3.7.2). ROC analysis has some distinct advantages. ROC curves are insensitive to changes in the class distribution and they allow for easy comparison of the performance of different classifiers (Provost and Fawcett, 2001; Fawcett, 2006). Note that ROC curves could also be used for obtaining custom thresholds (Fawcett, 2006).

The preceding paragraph focuses on binary classification and an extension is possible for the multi-class setting. In general, we want to calculate a volume under the ROC surface instead of an area under the ROC curve (Flach, 2004; Fieldsend and Everson, 2005). Computationally, this is quite demanding and, therefore, in this study we extended the approach similar to the analysis described in the previous section. In a one-versus-others setting, an AUC value was calculated for the setting in which each class is discriminated from all other classes. Final averaging was

performed and the resulting AUC value is, thus, an approximation of the true AUC value. The resulting value shows how the discrimination of each class from all the other classes varies. This approach is somewhat similar to the method described by Hand and Till (2001), who average the AUC values over all pair of classes (one-versus-one approach). Another one-versus-all approach is given by Provost and Domingos (2001), where the multi-class AUC is calculated as a sum of the different AUC values weighted by the prevalence of the corresponding class in the data set.

Provost et al. (1998) and Fawcett (2006) describe two methods for averaging ROC curves from the perspective of cross-validation: vertical averaging and threshold averaging. The latter method was also handled for easy visualisation of the average ROC curve calculated in the one-versus-all model evaluation performed in this work. Herein, all points in ROC space defined a different threshold on the FPR axis. On each individual ROC curve, the data point corresponding with each threshold was determined. If this ROC point was not used for construction of the corresponding ROC curve, its value was calculated by interpolation. For each FPR threshold, an average ROC point was calculated over all corresponding ROC points and, ultimately, an average ROC curve was drawn by combining all average ROC points.

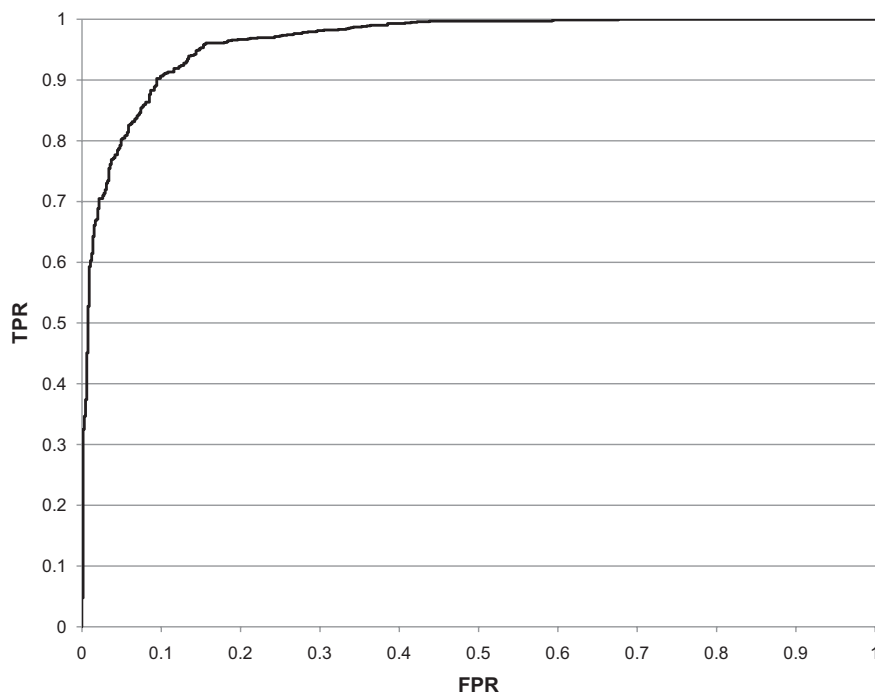


Figure 1.10: ROC curve. Example of a ROC curve for a two-class classification experiment. On the X axis, the false positive rate (FPR) is given, while the Y axis represents the true positive rate (TPR). The area under the curve is an estimation of the classification performance and equals here 0.965.

1.3.3 Wilcoxon Rank-Sum Statistic

The non-parametric Wilcoxon rank-sum test, also known as the Mann-Whitney U test or the Wilcoxon-Mann-Whitney test was used to test if two independent samples represent two

different distributions with respect to the rank-ordering of the observations in the two underlying population distributions. The test statistic is based on the assumptions that

- the observations are randomly selected from the population
- the two samples are independent of one another
- the observations are continuous
- the shapes of the underlying distributions are identical

The Wilcoxon rank-sum test is the nonparametric version of the two-sample t -test which assumes normal distributions of the two populations under the null hypothesis. Under the assumption of the null hypothesis no effect is present in the samples and both distributions are identical. The alternative hypothesis states that a distinct difference is present in the data, or it assumes a location shift of one of the distributions (Higging, 2004; Sheskin, 2004).

Assume M observations in the first population and N observations in the second population. Basically, the Wilcoxon rank-sum test makes the sum R of ranks of the observations of one of the two populations (arbitrary choice). The Wilcoxon rank-sum test is a two-sample permutation test based on R . Subsequent to the calculation of R , all possible permutations of the ranks are considered in which m ranks are assigned to first population and n ranks are assigned to the second population. For each permutation, the rank sum is calculated for the chosen population. For an upper-tail test, a p -value is calculated as

$$p_{\text{upper-tail}} = \frac{\text{number of rank sums } \geq \text{observed rank sum } R}{\binom{M+N}{N}} \quad (1.37)$$

Ties are commonly present in many data sets, meaning that identical observations are present. If ties were present, rank were averaged and a normal approximation of the p -value was calculated. Calculation of a normal approximation was also considered in the case of large sample sizes (Higging, 2004).

The Wilcoxon rank-sum statistic is equivalent to the Mann-Whitney statistic. In the calculation of the latter statistic, all pairs (x_i, y_j) are considered, with x_i an observation of the first population and $i = 1, \dots, M$, and with y_j observation j of the second population and $j = 1, \dots, N$. The upper tail statistic is subsequently defined as the number of pairs for which $x_i < y_j$ (Higging, 2004).

CHAPTER 2

Bacteriology

*What is a scientist after all? It is a curious man
looking through a keyhole, the keyhole of
nature, trying to know what's going on.*

JACQUES COUSTEAU

*Messieurs, c'est les microbes qui auront le
dernier mot.*

LOUIS PASTEUR

2.1 Introduction

This chapter deals with the result of over 3.5 billion years of evolution: the prokaryotic world we see, feel and experience today. Prokaryotes are detected from altitudes of 77 km in the atmosphere to depths of 2 km in the subsurface and have colonized soil, water, ice, air and eukaryotic organisms. The total number of individual prokaryotic cells on earth is estimated in the order of 5×10^{30} (Whitman, 2009). It is clear, the world of prokaryotes is worth looking at!

The founding father of bacteriology is Antonie van Leeuwenhoek who, in 1676, observed the first bacteria or 'animalcules' as described in his letters to the Royal Society of London. A first attempt towards a systematic rearrangement of microorganisms was done by Otto Müller at the end of the 18th century. The effective study of bacteria started somewhat later with the work of Louis Pasteur around 1862. Microbiology witnessed a real boost in the 1950s-1970s with the rise of the field of bacterial genetics and molecular biology and the following years were further characterized by molecular phylogenetic studies and numerical analysis. In the last decade, genomics has become 'the' hot-topic in microbiology (Sapp, 2005; Rosselló-Mora and Amann, 2001; Madigan et al., 2009). This very brief outline of the history of microbiology is a good starting point for this chapter as all items central in the different historical stages are touched in this study: systematics, genetics, molecular biology, phylogeny, genomics and numerical analysis. In order to find a way through all these elements, a short introduction is given to understand the context of the research performed.

2.2 A Taxonomic View on the World of the Bacteria

2.2.1 Introduction

The domain of the Bacteria is one of the three domains of life, next to the Eukarya and the Archaea, which was considered as the third domain of life only since the late 1970s (formerly kingdom Archaeobacteria) (Woese and Fox, 1977; Madigan et al., 2009). The Bacteria and the Archaea consist of prokaryotic organisms, while the Eukarya consist of eukaryotic organisms. The first two domains are typically referred to as ‘microbes’. However, microbial Eukarya do also exist and are called the protists. The main difference between prokaryotes and eukaryotes can be deduced from the latin connotation in their names (karyon) which refers to the presence of a nucleus (eukaryotes). The Archaea were originally regarded as aberrant members of the Bacteria because both groups share the same overall prokaryotic cell constitution (no membrane-bound compartments and no nucleus). Comparative ribosomal ribonucleic acid (rRNA) sequencing, however, revealed a closer genetic relationship between Archaea and Eukarya, besides a different deoxyribonucleic acid (DNA) packaging. Prokaryotic DNA is typically present in a single circular chromosome which contains most of the genes of the cell. All genes necessary for essential cell function, the house-keeping genes, are located on the chromosome. A minority of prokaryotes contains two or more chromosomes. Next to the chromosome, many prokaryotes also contain extrachromosomal circular DNA molecules or plasmids. These contain typically genes conferring additional properties (e.g. antibiotic resistance) (Madigan et al., 2009).

2.2.1.1 Towards Bacterial Classification

The study of the diversity of organisms and their relationships is called systematics, which links together phylogeny and taxonomy. Phylogeny is defined by the evolutionary relationships between organisms as deduced from the genetic information in nucleic acids and proteins. Taxonomy is the science of classification of organisms. Bacterial taxonomy consists of three interrelated fields: classification, nomenclature and identification. Classification is the arrangement of organisms into progressively more inclusive groups (taxa) on the basis of genotypic/phenotypic similarities or relationships. Herein, the genotype covers all genetic information of an organism coded in its DNA, while the phenotype corresponds to all observable characteristics of an organism and is regarded as the expression of the genotype. Important to mention, though, is that no official classification of the Bacteria exists because taxonomy remains a matter of scientific judgment and general agreement. At present, the best consensus is the most widely accepted and referenced taxonomic outline reported in Bergey’s Manual of Systematic Bacteriology, published by the Bergey’s Manual Trust (published first in 1923). Second, nomenclature is the assignment of names to the taxonomic groups according to the international rules defined in the Bacteriological Code - The International Code of Nomenclature of Bacteria (ICNB). A prokaryotic strain is given a genus name and species epithet. Note that the Bacteriological Code only deals with procedures for nomenclature and does not govern the

delimitation of taxa nor their relations. Third, identification is the practical use of a classification scheme to determine the identity of an isolate as a member of an established taxon or as a member of a previously unidentified species (Lapage et al., 1992; Sneath and Brenner, 1992; Vandamme et al., 1996; Euzéby, 1997; Brenner et al., 2005b; Madigan et al., 2009). We break the introduction here and return back to Chapter 1, Subsection 1.1.1 where the terms of classification and identification are defined in a machine learning perspective. In contrast to bacterial taxonomy, classification in machine learning terms corresponds to mathematically describing relationships between existing groups (in microbiology: taxa). Identification has the same connotation, that is, using a classification scheme to identify unknown samples (in microbiology: isolates).

Classification and an adequate description of bacteria require knowledge of their morphological, biochemical, physiological and genetic characteristics. Thus, taxonomy is a dynamic concept that changes on the basis of the available data. The formal taxonomic ranks or taxa following domain are: phylum, class, order, family, genus, species and subspecies. In this study, we focused on the rank of species, which is seen as the basic and most important taxonomic group in bacterial systematics (Brenner et al., 2005b; Madigan et al., 2009). Before going into more detail about taxonomy, it is important to first explain some necessary terms, e.g. what are strains and what is a type?

A bacterial strain is made up of the descendants of a single isolation in pure culture, and usually originates from a succession of cultures ultimately derived from an initial single bacterial colony. Concerning nomenclature, a bacterial strain can be designated in any manner, e.g. by the name of an individual, by a locality, or by a number. To make it more complex, a culture of bacteria is a population of bacterial cells in a given place at a given time and a clone is a population of bacterial cells derived from a single parent cell. Finally, a taxon is always associated with its nomenclatural type, referred to as ‘type’. The type is that element of the taxon with which the name is permanently associated. It is the name bearer and reference example of the taxon. The nomenclatural type is not necessarily the most typical or representative element of the taxon. A type strain of a taxon is one of the strains on which the author who first described a named organism based the description of the organism and which the author, or a subsequent author, definitely designated as a type. This naming can be extended towards higher taxonomic units. As such, the type of a genus or subgenus is the type species, that is, the single species or one of the species included when the name was originally validly published. More information about these nomenclature rules can be found in the Rules of Nomenclature with Recommendations of the ICNB (especially rules 15-22) (Lapage et al., 1992; Brenner et al., 2005b; Madigan et al., 2009).

2.2.1.2 What’s in a name? That which we call a species

A universally accepted concept of the term species does not exist for prokaryotes. By the term species concept, a framework is meant to understand how and why an observer can sort organisms into species. That is, what kind of unit do we think the term species embraces and what characteristics are shared between all members of a species. A lot of discussion and disagreements exists in this context and several species concepts are proposed. A general dis-

cussion on this topic can be found in the papers of Rosselló-Mora and Amann (2001), Cohan (2002), Gevers et al. (2005) and Doolittle and Papke (2006). As described in Bergey's Manual of Systematic Bacteriology a bacterial species is defined as 'a distinct group of strains that have certain distinguishing features and that generally bear a close resemblance to one another in the more essential features of organization'. For practical use, a species definition is proposed as a standard for how to assign isolates to a named species or to identify new species. In 1987, a first pragmatic definition was given by the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics (Wayne et al., 1987). In their report, the committee states that 'the phylogenetic definition of a species generally would include strains with approximately 70% or greater DNA-DNA relatedness and with 5°C or less ΔT_m (melting temperature). Both values must be considered'. Besides this, the committee stated that phenotypic characteristics should agree with this definition and would be allowed to override the phylogenetic concept of species only in a few exceptional cases. In 1994, Stackebrandt and Goebel showed that the 70% threshold in DNA-DNA hybridization (DDH) usually relates to more than 97% sequence identity in 16S rRNA gene. Note that the 16S rRNA gene contains conservative to hypervariable nucleotide regions. Important, the resolution power of 16S rRNA gene sequences is limited when closely related organisms are inspected, due to a limited amount of variation in the rRNA gene sequence. As an alternative, distinctions between (very) closely related bacteria can be revealed by using multiple genes whose sequences have diverged more than the 16S rRNA gene. This technique is known as multi-locus sequence typing (MLST) in which, commonly, six to eight housekeeping genes (core genes essential for functioning of the cell) from an organism are analyzed. The resolution of MLST allows for strain differentiation within a species. Because of progress in methodology (16S rRNA gene analysis, rapid DNA typing methods, MLST and genome analysis) and new insights in population structure, a new ad hoc committee was formed in 2002. This ad hoc committee for the re-evaluation of the species definition in bacteriology concluded that despite the drawbacks with respect to reproducibility, workability and rigid application of DDH values for species delineation, the presented system is sound. And, the current species definition as described by Rosselló-Mora and Amann (2001) can be considered pragmatic, operational and universally applicable: 'a species is a category that circumscribes a (preferably) genomically coherent group of individual isolates/strains sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardized conditions'. As a result, a DDH reassociation value of 70% is seen as the 'gold-standard' for species delineation. But, due to practical difficulties with DDH, the exponential growth of sequence databases and improvement in sequencing technology, the 97% 16S rRNA gene standard is widely used for species identification. Nonetheless, these methods were considered to be inadequate for defining a prokaryotic species and incapable of keeping pace with the levels of diversity that are being discovered in nature. Consequently, it is now accepted among microbial taxonomists that a prokaryotic species should be classified after the analysis and comparison of as many parameters as possible, combining phenotypic and genomic markers. This technique is also known as polyphasic taxonomy (Vandamme et al., 1996; Rosselló-Mora and Amann, 2001; Brenner et al., 2005b; Gevers et al., 2005; Madigan et al., 2009). In this last decade, computational genomics got into play in bacteriology, leading to a new approach for evaluat-

ing the current species delineation standards and the species concept. The number of genomes published online increases almost every day. At 04/08/2009, 874 complete bacterial genomes were published online and 1845 genomes were in progress (NCBI, 2009a). Pioneers in the computational analysis of bacterial genomes are Konstantinidis and co-workers who showed that the 70% DDH gold standard tightly corresponds to approximately 95% average nucleotide identity (ANI), which provides a robust representation of the phylogenetic relationships at the species level. This correspondence was experimentally verified by Goris et al. (2007). They also showed that, based on highly reliable markers, a genomic evaluation of the MLST method provides robust phylogeny of organisms sharing 70-95% ANI. This corresponds to closely or distantly related species of the same genus. Next to ANI, the MLST method has a clear taxonomic value (Konstantinidis and Tiedje, 2005a,b; Konstantinidis et al., 2006b,a; Goris et al., 2007; Konstantinidis and Tiedje, 2007).

Formal validation of taxonomic standing as a new species or genus requires the publication of a detailed description of the organism's characteristics and distinguishing traits, together with the proposed name, sequence accession numbers of one of the public sequence databases (GenBank, EMBL or DDBJ), and the deposit of the bacterial type strain(s) in at least two international culture collections in two or more different countries (IJSEM, 2009). Herein, an international recognised culture collection is regarded as a repository for cultures of microbial strains which are listed in a catalog and provided upon request. On 22/11/2009, 568 culture collections in 68 countries were registered (WFCC, 2009). For valid and official acceptance of the new genus and/or species, description needs to be published in the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM). Each month IJSEM publishes an approved list of novel validated names. Valid publication of a new name or combination is overseen by the International Committee on Systematics of Prokaryotes (ICSP) in collaboration with the Judicial Commission, an international body that is responsible for the correct application of the Bacteriological Code. From the IJSEM website (IJSEM, 2009), a formal description of a new taxon requires the following information:

1. A list of the strains included in the taxon
2. A statement or tabulation of the characteristics of each strain
3. A list of characteristics considered essential for membership in the taxon
4. A list of characteristics which qualify the taxon for membership in the next higher taxon
5. A list of diagnostic characteristics, i.e. characters which distinguish the taxon from closely related taxa
6. Designation of the type for that taxon
7. The reactions of the type strain of a new species
8. For all characteristics that vary among strains within the species the specific reaction of the type strain must be defined

Full description of this information can be found in Appendix 7 of the *International Code of Nomenclature of Bacteria* (1990 Revision) (Lapage et al., 1992). Different websites also provide a list of valid and approved bacterial names such as the *List of Prokaryotic Names with Standing in Nomenclature* (LSPN) of Jean Euzéby (Euzéby, 1997), the *Bacterial Nomenclature*

Up-to-Date of the German Collection of Microorganisms and Cell Cultures (DSMZ, 2009) and the NCBI Taxonomy Browser (NCBI, 2009b). On 03/11/2009, 7,995 bacterial species were validly published, covered by 1,561 validly published genera (Euzéby, 1997). The yearly number of valid bacterial species descriptions is visualized in Figure 2.1. At present, the number of prokaryotic species on earth cannot be estimated accurately. Theoretical estimates suggest that soil and deep sea contain more than 10^6 species (Whitman, 2009).

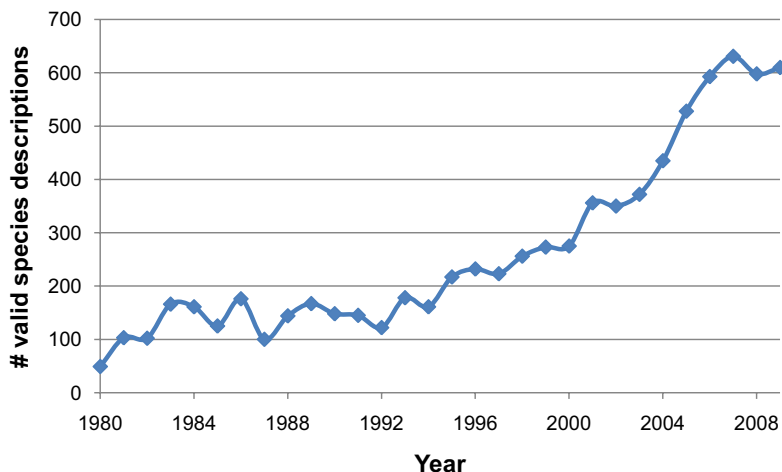


Figure 2.1: Trend in novel valid species descriptions. Number of valid bacterial species descriptions since 1980. For 2009, the number of validly published species is given as counted on 03/11/2009. Data captured from Euzéby (1997).

2.2.1.3 The Other Taxonomic Ranks and Subdivisions

The terms species and strain being clarified, let us now have a closer look at the other taxonomic ranks. First, all species are assigned to a genus which can be functionally defined as one or more species with the same general phenotypic characteristics, and which cluster together on the basis of 16S rRNA gene sequence data. Although most new genera are designated based on 16S rRNA gene sequence analysis, no formal definition for the rank genus exist. For almost all genera holds that they can be differentiated phenotypically. Following the binomial Linnaean nomenclature a bacterial strain is designated a genus and species name. Classification at the rank of family and higher levels are even less well defined and descriptions of the taxa are much more general. Mainly, consistency in both the genotype and the phenotype is widely considered a measure for describing these taxa. Below the rank of species, the lowest formal rank in bacterial taxonomy is that of subspecies. At present, also for this rank no definition or guideline is available. A subspecies is regarded as a subdivision of a species based on phenotypic variations or genotypic clusters of strains within the species. Besides a genus, species and/or subspecies name, a strain may also be denoted by an infrasubspecific term. This term is, however, not covered by the Rules of the ICNB as a formal taxonomic rank. Infrasubspecific ranking is based on a specific characteristic (e.g. pathogenic properties for a certain host) and different terms are defined, such as biovar, chemovar, cultivar, *forma specialis*, morphovar, pathovar, phagovar, phase, serovar and state (Lapage et al., 1992; Brenner et al., 2005b).

Let us finally go back to the second highest taxonomic rank, the phylum, that is also not covered by the Rules of the ICNB. A phylum is regarded to be a major evolutionary group (Lapage et al., 1992; Madigan et al., 2009). Importantly, the criteria as to what actually constitutes a phylum remain to be defined. Due to this absence of objective criteria, it is unclear how many phyla actually exist within the Bacteria, how to distinguish between them and how to discriminate them from their subdivisions (Gupta, 2005; Ludwig and Klenk, 2001). According to the second edition of the Bergey's Manual of Systematic Bacteriology 24 bacterial phyla exist (Brenner et al., 2005a). A 16S rRNA gene tree visualizing most phyla is visualized in Figure 2.2.

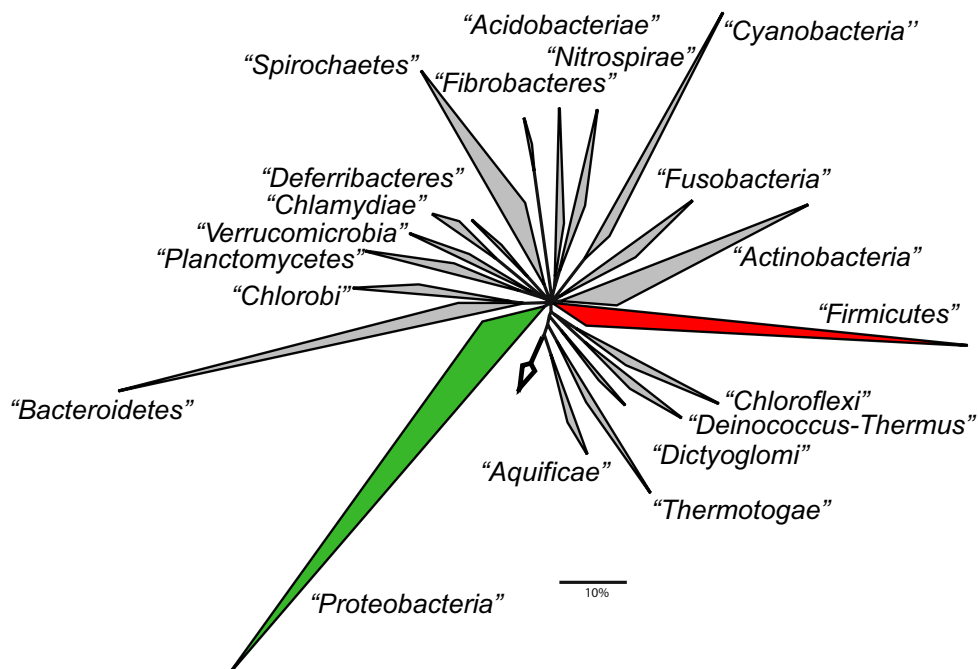


Figure 2.2: The major bacterial phyla. 16S rRNA gene-based phylogenetic parsimony tree showing the major bacterial phyla. The triangles indicate groups of related organisms, while the angle at the root of the group roughly indicates the number of sequences available and the edges represent the shortest and longest branch within the group. All available almost complete homologous sequences from *Archaea* and *Eukarya* were used as outgroup references to root the tree, indicated by the arrow (Ludwig and Klenk, 2001). In this study, we focused on two genera of the red phyla of the Firmicutes and on one genus of the green phyla of the Proteobacteria.

Besides this internationally accepted taxonomic ranking, different alternative groupings exist. A widely-used grouping of the Bacteria is that of the Gram-positive and Gram-negative bacteria. Herein, the distinction results from differences in the cell wall structure and is visualized by the Gram staining method. The cell wall of Gram-positives consists of a thick layer of about 90% of peptidoglycan, where the cell wall of Gram-negatives is multi-structured with the presence of an outer membrane, called the lipopolysaccharide (LPS) layer and a thin inner-membrane. The cell wall of Gram-negatives consists only of about 10% peptidoglycan. Other commonly used grouping approaches depend on growth conditions such as temperature, pH or the need for oxygen. In this latter case, one distinguishes between aerobes, microaerophiles, facultative (an)aerobes and anaerobes (Madigan et al., 2009).

2.2.2 Bacterial Identification

As stated in the previous subsection, identification of bacteria can only be achieved by relying on a specific classification scheme. Once this scheme is resolved, analysis of the genotype and/or phenotype can start. Remember that genotypic methods are directed towards DNA or RNA, while phenotypic methods focus on proteins, chemotaxonomic markers and a wide range of other expressed molecules, metabolic pathways and phenotypic characteristics such as motility, cell shape, etc. The main disadvantage of phenotypic analysis is that information embedded in the genome is only partly expressed, as gene expression is directly or indirectly related to environmental conditions. A lot of techniques exist for analyzing the massive amount of different molecules and the popularity of a technique depends heavily on its practical use, time-consumption and cost. Before molecular techniques were available, identification was solely based on morphology, physiology and growth conditions. From a taxonomic perspective, variability of a taxon is expressed in different molecules which underscores the use of a polyphasic approach in the classification and identification of bacteria. Nowadays, genotypic methods flourish and are commonly preferred over phenotypic techniques. Recall however, that in their species definition, the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics (Wayne et al., 1987) stated that ‘phenotypic characteristics should agree with this definition’, meaning that also phenotypic consistency is required in the definition of a new species (Vandamme et al., 1996; Madigan et al., 2009).

In this study, we focused on the phenotype as analyzed by chemotaxonomic methods. In contrast to classical phenotypic methods that focus on morphological, physiological and biochemical features, chemotaxonomy is referring to the application of methods for analyzing various chemical constituents of the cell, mainly lipids, proteins, amino acids and sugars. Notice that the measures of these features are regarded as a direct reflection of the expression of the genetic information. Therefore, it is important that the observed variation in chemical composition is the result of genetic differences and not due to culture and growth conditions. For reproducibility and comparative analysis of the results, highly standardized procedures are critical (Vandamme et al., 1996; Rosselló-Mora and Amann, 2001). A major part of the cell that is of high interest to many microbiologists are the bacterial membranes because of the high variability in their composition. Chemotaxonomy plays a central role in this research and we focused on the technique of whole-cell fatty acid analysis.

Many analytic methods rely on bacterial growth on certain nutrients. Solutions of these nutrients, culture media, are used for bacterial growth or cultivation. Most techniques, such as whole-cell fatty acid analysis, require the isolation of pure cultures, i.e. cultures containing only a single microorganism. Typically, pure cultures are obtained by enrichment procedures, e.g. the streak plate. Different types of culture media exist: defined media, complex media, selective media, solid versus liquid media, etc. (Madigan et al., 2009). For taxonomic purposes, the scope of culture-dependent analytical methods should be very large and, consequently, requires the growth of as many bacteria as possible. As highly standardized conditions are required to limit the variability in the respective features, the usefulness of these methods is restricted towards the fraction of bacteria that are able to grow on general media. Note that a large group of bacteria is

simply unculturable. For example, it is estimated that only 0.3% of the bacteria in soil, 0.25% of the bacteria in sediments and even 0.001-0.1% of the bacteria in seawater are culturable (Amann et al., 1995). This clearly shows the disadvantage of culture-dependent analytic techniques, although more emphasis is put on stimulating research towards the improvement of cultivation methodologies.

2.2.3 A General Focus on the Genus

In the following section, the general characteristics of the three genera considered in this research are briefly described. These genera have already been studied for decades and, thus, a massive amount of information is available from the literature. The following descriptions rely on the genus and species descriptions in Bergey's Manual of Systematic Bacteriology (Palleroni, 2005; Logan and De Vos, 2009; Priest, 2009) and I refer to this excellent manual for detailed reading.

2.2.3.1 *Bacillus*

In the domain of the Bacteria, the genus *Bacillus* is classified in the domain of *Bacteria*, phylum of *Firmicutes*, class of *Bacilli*, order of *Bacillales*, family of *Bacillaceae* (Brenner et al., 2005a). The type species is *Bacillus subtilis* as described by Cohn in 1872. At 21/11/2009, 157 *Bacillus* species were validly published. The monthly changes in the taxonomy of the genus since January 2006 are visualized in Figure 2.3. On 21/11/2009, for 10 valid species one or more complete genome sequences were available online (NCBI, 2009a).

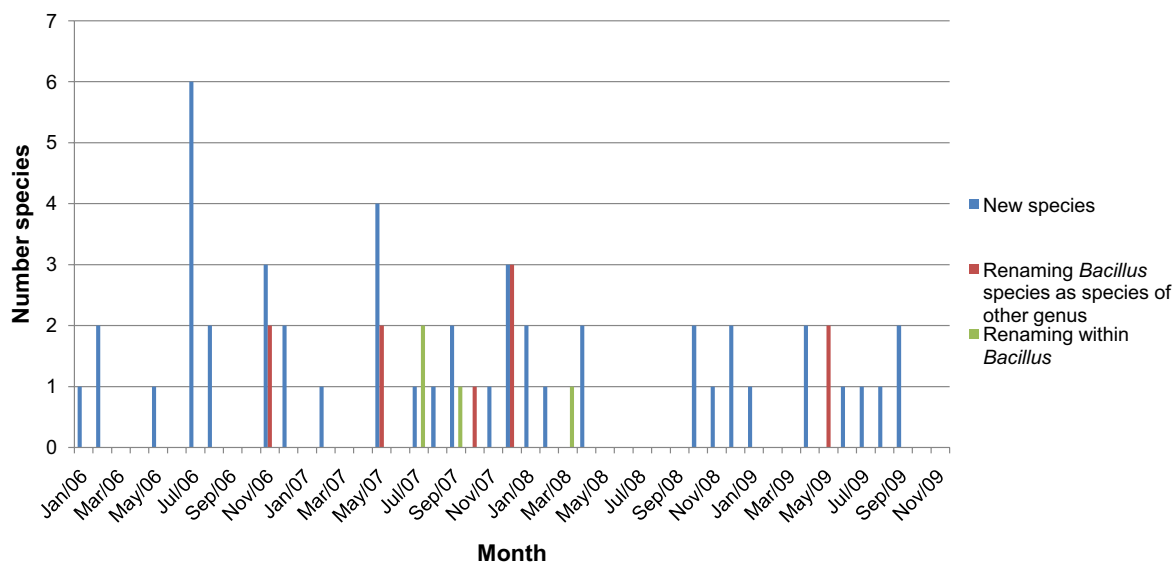


Figure 2.3: Monthly nomenclatural changes in the bacterial taxonomy of the genus *Bacillus* as published by the IJSEM between January 2006 and November 2009. The number of novel described species is given, together with the number of species renamed within the genus *Bacillus* as the number of valid *Bacillus* species renamed to a new species outside the genus. Information extracted from the List of Prokaryotic Names with Standing in Nomenclature (Euzéby, 1997).

Bacillus belongs to the Gram-positive bacteria and strains are typically rod-shaped, straight or slightly curved cells. Aerobe, facultative anaerobes and anaerobes are described and most

species grow on routine media. *Bacillus* strains are mostly isolated from soil, but are also present in water, food and clinical specimens. A main characteristic is the formation of endospores, that turns *Bacillus* into an important contaminant of food and industrial/medical sites. Most species have little or no pathogenic potential although a few important exceptions exist. These all belong to a phenogenetically distinct group. *Bacillus anthracis* causes the anthrax disease, *Bacillus thuringiensis* is pathogenic to invertebrates and *B. cereus* is a food spoiler that may produce various toxins responsible for food-borne diseases. Because of its pathogenicity towards invertebrates, *Bacillus thuringiensis* is of main interest in agriculture for the development of pesticides and for genetic engineering for the creation of transgenic crops. *Bacillus* is one of the model genera in bacteriology (Logan and De Vos, 2009).

The genus *Bacillus sensu lato* has known a drastic rearrangement in the last decades, giving rise to several new genera following splitting of one or more *Bacillus* species. Examples are *Alicyclobacillus* (1992), *Aneuribacillus* (1996), *Brevibacillus* (1996), *Geobacillus* (2001), *Gracilibacillus* (1999), *Marinibacillus* (2001), *Paenibacillus* (1993), *Ureibacillus* (2001) and *Virgibacillus* (1998) (Euzéby, 1997; Berkeley, 2002; Kämpfer, 2002).

Two *Bacillus* groups, each containing closely related species, are particularly taxonomically challenging: the *Bacillus cereus* group and the *Bacillus subtilis* group. First, the *Bacillus cereus* group that consists of *B. cereus*, *B. thuringiensis*, *B. anthracis*, *B. mycoides*, *B. pseudomycoides* and *B. weihenstephanensis*. This group is discussed because of the relevance of its species differentiation, especially concerning the valid species *B. cereus*, *B. anthracis* and *B. thuringiensis*. The main points of discussion concern the weakness of species discriminations that are based on phenotypic and pathogenic characteristics (Drobniewski, 1993; Bavykin et al., 2004; Priest et al., 2004; Rasko et al., 2005; Tourasse et al., 2006; Vilas-Bôas et al., 2007). The second group is the *Bacillus subtilis* group that consists of *B. subtilis*, *B. amyloliquefaciens*, *B. atrophaeus*, *B. licheniformis*, *B. mojavensis*, *B. sonorensis*, *B. pumilus* and *B. vallismortis*. This group forms a very tight phylogenetic cluster on the basis of 16S rRNA gene sequencing, although the *gyrB* gene has a higher resolution for species discrimination. Furthermore, the classification achieved by *gyrB* gene sequence analysis is in agreement with the results obtained from DDH. In 2006, five new species were described which show high similarities to the *Bacillus subtilis* group in 16S rRNA gene sequence while DDH reassociation percentages lower than 70% were found: *B. aerioides*, *B. aerophilus*, *B. altitudinis*, *B. stratosphericus* and *B. tequilensis* (Chun and Sook Bae, 2000; Gatson et al., 2006; Hutsebaut et al., 2006; Shivaji et al., 2006; Wang et al., 2007a). These results were confirmed by a maximum likelihood tree following 16S rRNA gene sequence analysis based on the validly published taxonomy of May 2008 (Euzéby, 1997). This tree of 147 valid *Bacillus* species is shown in Figure 2.4 (Chun and Sook Bae, 2000; Gatson et al., 2006; Hutsebaut et al., 2006; Wang et al., 2007a).

2.2.3.2 *Paenibacillus*

In 1993, Ash et al. proposed the new genus *Paenibacillus* following 16S rRNA gene sequence analysis. The same higher taxonomic ranking is identical to that of *Bacillus*, except for the rank of family which is defined as *Paenibacillaceae* (Brenner et al., 2005a). The type

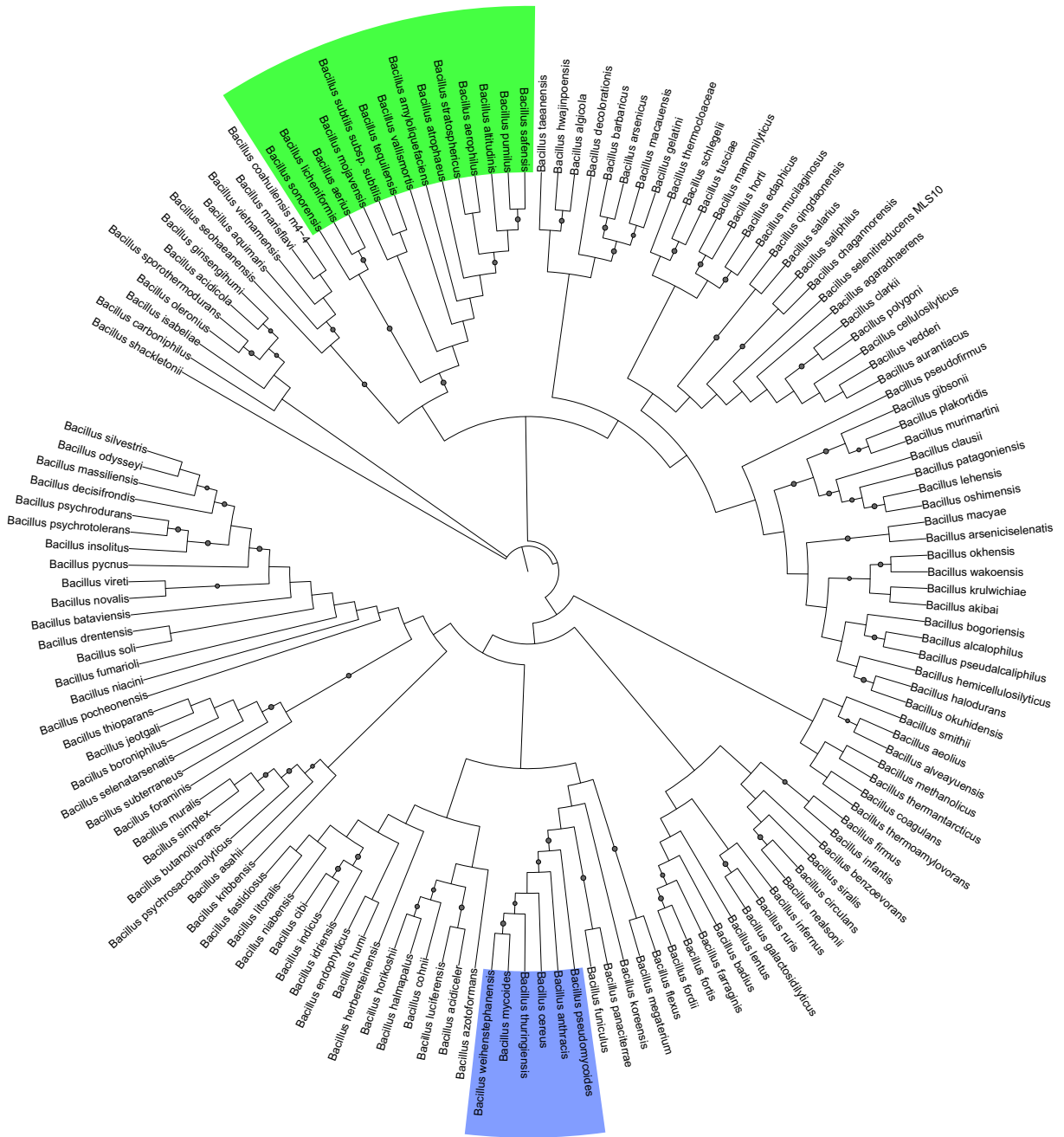


Figure 2.4: 16S rRNA gene sequence-based maximum likelihood tree of the genus *Bacillus*. The 147 species included correspond to the validly published taxonomy of May 2008. One high-quality 16S rRNA sequence per species is selected from the SILVA database (Pruesse et al., 2007). The phylogenetic tree is built by the maximum likelihood algorithm as implemented in the RAxML software (based on 1000 bootstraps) and is visualized with the iTol webtool (Stamatakis, 2006; Letunic and Bork, 2007). The tree branch lengths are ignored because of outlier species that make the use of branch length pointless. The green group corresponds to the *Bacillus subtilis* group, while the blue group corresponds to the *Bacillus cereus* group. Dotted branches correspond to bootstrap values larger than 75%.

species is *Paenibacillus polymyxa*. On 21/11/2009, the genus contained 103 validly published species. The monthly changes in the taxonomy of the genus since January 2006 are visualized in Figure 2.5. No genomes of valid *Paenibacillus* species have been completed and published so far.

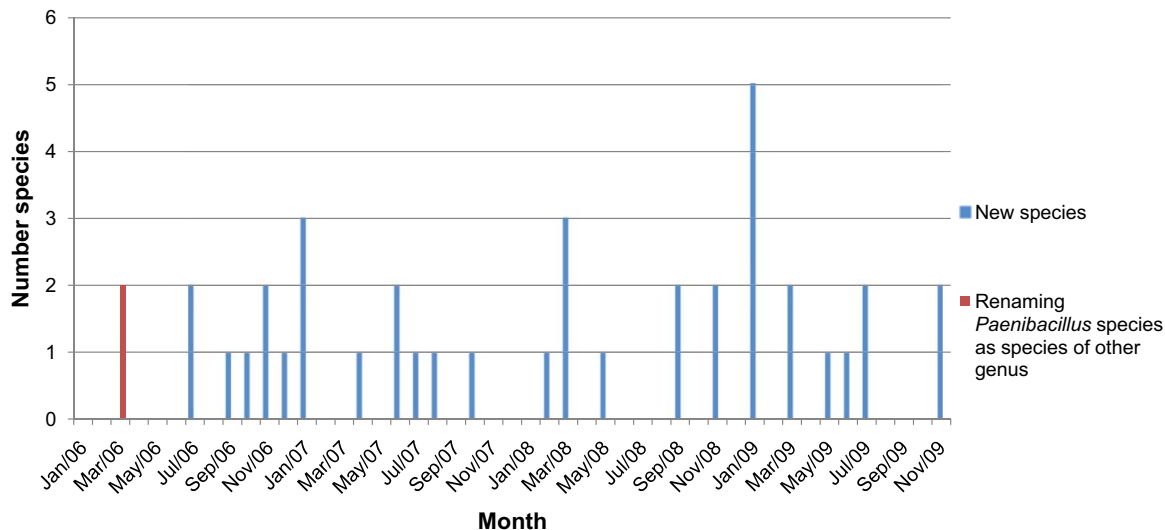


Figure 2.5: Monthly nomenclatural changes in the bacterial taxonomy of the genus *Paenibacillus* as published by the IJSEM between January 2006 and November 2009. The number of novel described species is given, together with the number of valid *Paenibacillus* species renamed to a new species outside the genus. Information extracted from the List of Prokaryotic Names with Standing in Nomenclature (Euzéby, 1997).

Paenibacillus strains are rod-shaped, Gram-positive and also show endospore formation. Paenibacilli are facultative anaerobic or strictly aerobic, and most of them grow on nutrient agar at neutral pH. Some strains are pathogens of insects such as. *Pa. larvae* and *Pa. poppilliae*. The main habitat of *Paenibacillus* species is also soil. Besides DDH and 16S rRNA sequence analysis, species discrimination is also achieved by MLST with the housekeeping genes *rpoB*, *gyrA*, *gyrB*, *recA*, etc. (Priest, 2009). A maximum likelihood tree of the *Paenibacillus* genus based on 16S rRNA gene sequences is visualized in Figure 2.6. This tree is based on the validly published list of bacterial species of May 2008.

2.2.3.3 *Pseudomonas*

In the domain of the Bacteria, the taxonomic ranking of the genus *Pseudomonas* is as follows: domain of *Bacteria*, phylum of *Proteobacteria*, class of *Gammaproteobacteria*, order of *Pseudomonadales* and family of *Pseudomonaceae* (Brenner et al., 2005a). The type species is *Pseudomonas aeruginosa*, originally discovered by Schroeter in 1872. The genus was, however, proposed by Migula in 1894. On 21/11/2009, it contained 126 validly published *Pseudomonas* species. The monthly changes in the taxonomy of the genus since January 2006 are visualized in Figure 2.7. On 21/11/2009, for 6 valid species one or more complete genome sequences are available online (NCBI, 2009a).

Pseudomonas members are straight or slightly curved motile rods with merely polar flagella. *Pseudomonas* strains belong to the Gram-negative bacteria and are respiratory but never

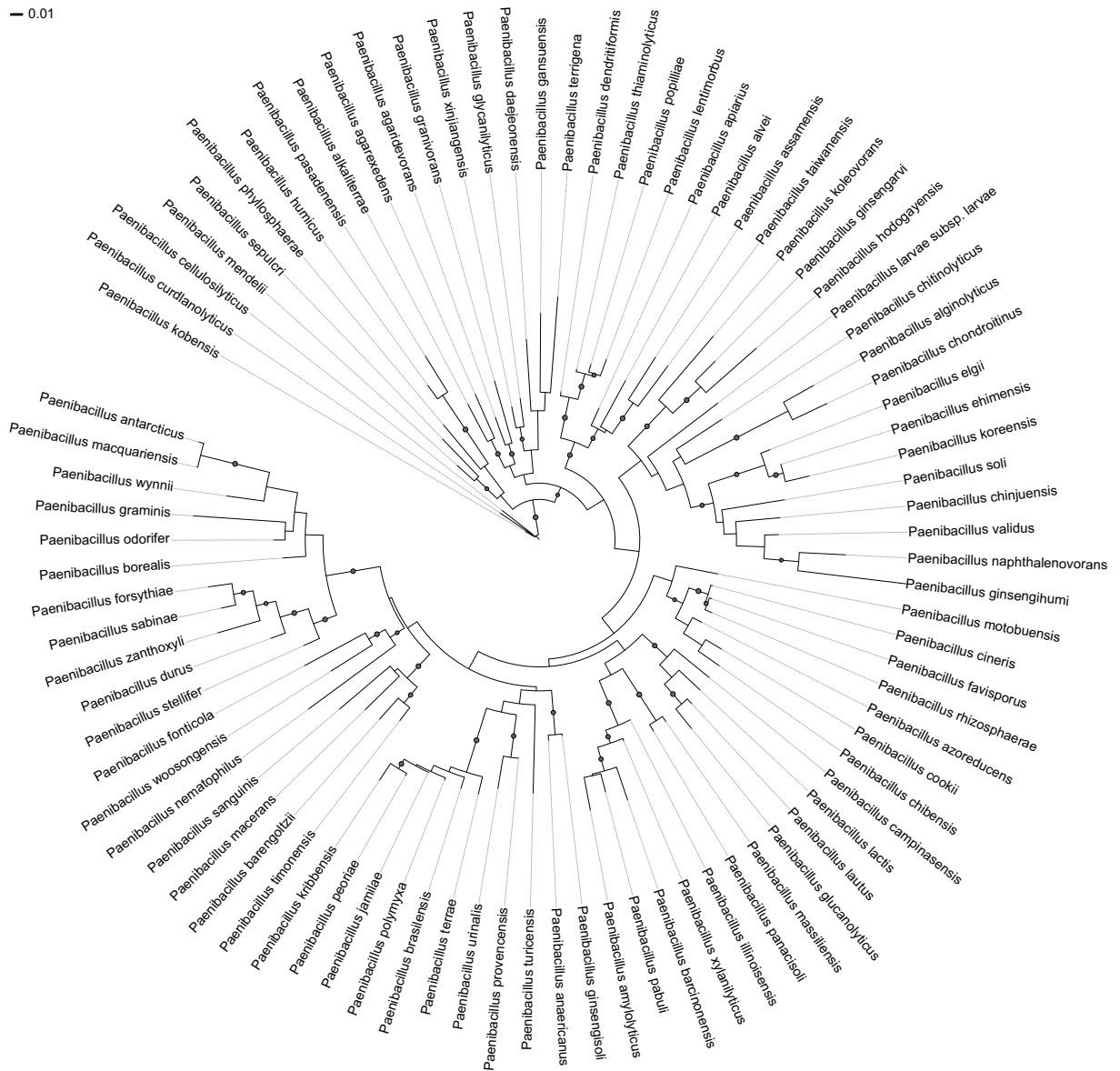


Figure 2.6: 16S rRNA gene sequence-based maximum likelihood tree of the genus *Paenibacillus*. The 101 species included correspond to the validly published taxonomy of May 2008. One high-quality 16S rRNA sequence per species is selected from the SILVA database (Pruesse et al., 2007). The phylogenetic 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 (Stamatakis, 2006; Letunic and Bork, 2007). Dotted branches correspond to bootstrap values larger than 75%.

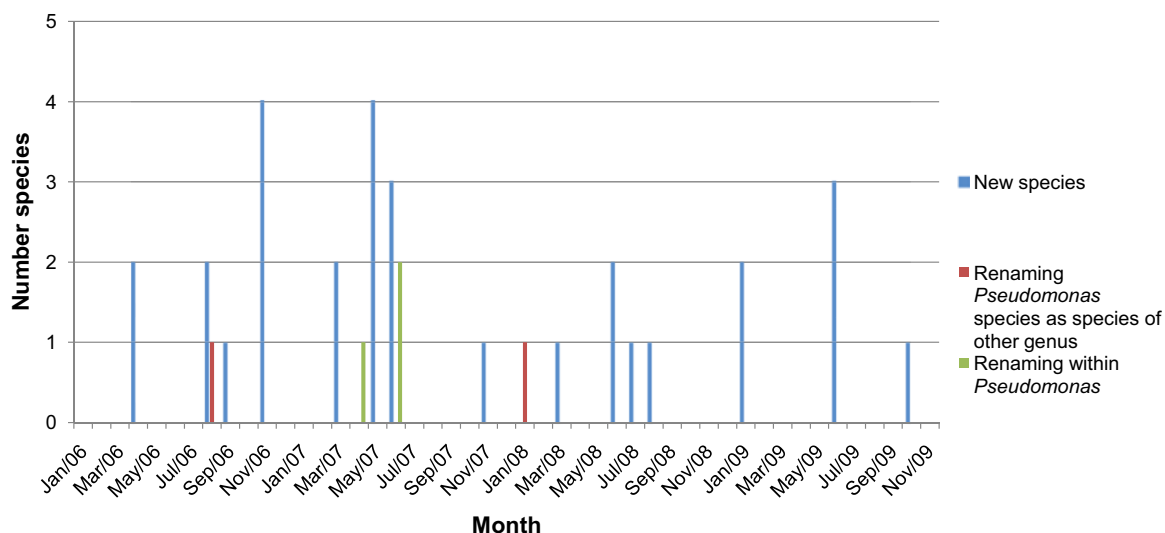


Figure 2.7: Monthly nomenclatural changes in the bacterial taxonomy of the genus *Pseudomonas* as published by the IJSEM between January 2006 and November 2009. The number of novel described species is given, together with the number of species renamed within the genus *Pseudomonas* as the number of valid *Pseudomonas* species renamed to a new species outside the genus. Information extracted from the List of Prokaryotic Names with Standing in Nomenclature (Euzéby, 1997).

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 pathogenicity. 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 classification of the genus *Pseudomonas* are Palleroni and co-workers (University of California, Berkeley, USA), who in 1973 described an initial grouping of five discrete *Pseudomonas* clusters based on rRNA-DNA hybridization (Palleroni et al., 1973; Palleroni, 1984). Since then, the taxonomy of aerobic Pseudomonads underwent a series of rearrangements based on rRNA gene similarity groups and the genus as described today corresponds to rRNA group I. This implies that numerous species previously assigned to the genus *Pseudomonas sensu lato* were transferred at the generic or suprageneric ranks, mainly residing in the α -, β -, and γ Proteobacteria classes. Examples are *Acidivorax*, *Aminobacter*, *Brevundimonas*, *Burkholderia*, *Comamonas*, *Halomonas*, *Methylobacterium*, *Ralstonia*, *Sphingomonas*, *Xanthomonas*, etc. cited in Kersters et al. (1996) and 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). In 2000, Anzai et al. (2000) reevaluated 128 valid and invalid *Pseudomonas* species based on 16S rRNA sequence data and further reassigned several species to other genera. The complexity of the taxonomy of the present genus *Pseudomonas* is also demonstrated by *rpoB* gene analysis (Tayeb et al., 2005). How-

ever, validation of the groupings still needs DDH data and extensive phenotypic analysis before emendations at the species level can be proposed. Regarding plant pathogenicity, a multitude of pathovars are described within the species *P. syringae*. A DDH study regarding the different *P. syringae* pathovars showed the existence of nine discrete genomospecies (Gardan et al., 1999). We followed these genomospecies classifications as if they were formal species. A maximum likelihood tree of the genus *Pseudomonas* based on 16S rRNA gene sequences is visualized in Figure 2.8. This tree is based on the validly published list of bacterial species of May 2008.

2.2.4 Where Machine Learning Meets Bacteriology

The first studies reporting a computational approach to bacterial identification were based on multi-criteria decision making (Fichefet et al., 1984; Butler et al., 1992). In later years, the use of machine learning techniques in bacteriology was dominated by the application of artificial neural networks (ANNs). ANNs were applied on different data types for the identification of numerous bacterial species and groups. The main research was done by only a small number of research groups, e.g. Ruggiero and co-workers (University of Genova, Italy) and Goodacre and co-workers (University of Wales, UK). The first group applied ANNs mainly on marine and environmental bacteria by using gas chromatographic FAME data (Ruggiero et al., 1993; Bertone et al., 1996; Giacomini et al., 2000, 2004). This is also the only research group so far using machine learning techniques on gas chromatographic FAME data for the identification of bacteria. However, they focused on the discrimination of species of different genera and, thus, not on intra-genus species identification. The second group focused more on bacteria of clinical interest by using pyrolysis mass spectral data and Fourier transformed infrared spectroscopy (FTIR) (Freeman et al., 1994; Goodacre et al., 1996a,b, 1998a,b). Different other papers were published reporting ANN applications in bacteriology. The main research goal concerned achieving an improved identification of bacterial species with clinical or food safety importance. Though different types of data were used, research was mainly directed to the analysis of spectral data. Examples are pyrolysis mass spectrometry or gas chromatography (Chun et al., 1993; Donohue and Welsh, 2004; Voisin et al., 2004), FTIR (Quinteiro Rodríguez, 2000; Udelhoven et al., 2000; Harrington et al., 2001; Mouwen et al., 2006; Rebuffo et al., 2006; Dziuba et al., 2007; Rebuffo-Scheer et al., 2007; Bosch et al., 2008), MALDI and SELDI TOF (Bright et al., 2002; Lancashire et al., 2005; Schmid et al., 2005; Yang et al., 2009), FAME MS (Xu et al., 2003), genetic fingerprints (REP-PCR, RAPD, 16S rDNA) (Tuang et al., 1999; Moschetti et al., 2001; Iversen et al., 2006), protein fingerprints (electrophoresis, SDS-PAGE) (Yong et al., 2002; Piraino et al., 2006), electronic sensor of volatile metabolites (Dutta et al., 2004, 2002; Moens et al., 2006) and biochemical test kits (Iversen et al., 2006). In the last years, the first papers dealing with random forests (RFs) and support vector machines (SVMs) for bacterial identification were published. The respective RF studies were mainly addressing MALDI TOF data (Satten et al., 2004; Hettick et al., 2006; Moura et al., 2008; Williamson et al., 2008) while SVMs were mainly used on mass and Raman spectrometry (Satten et al., 2004; Rösch et al., 2005; Gaus et al., 2006).

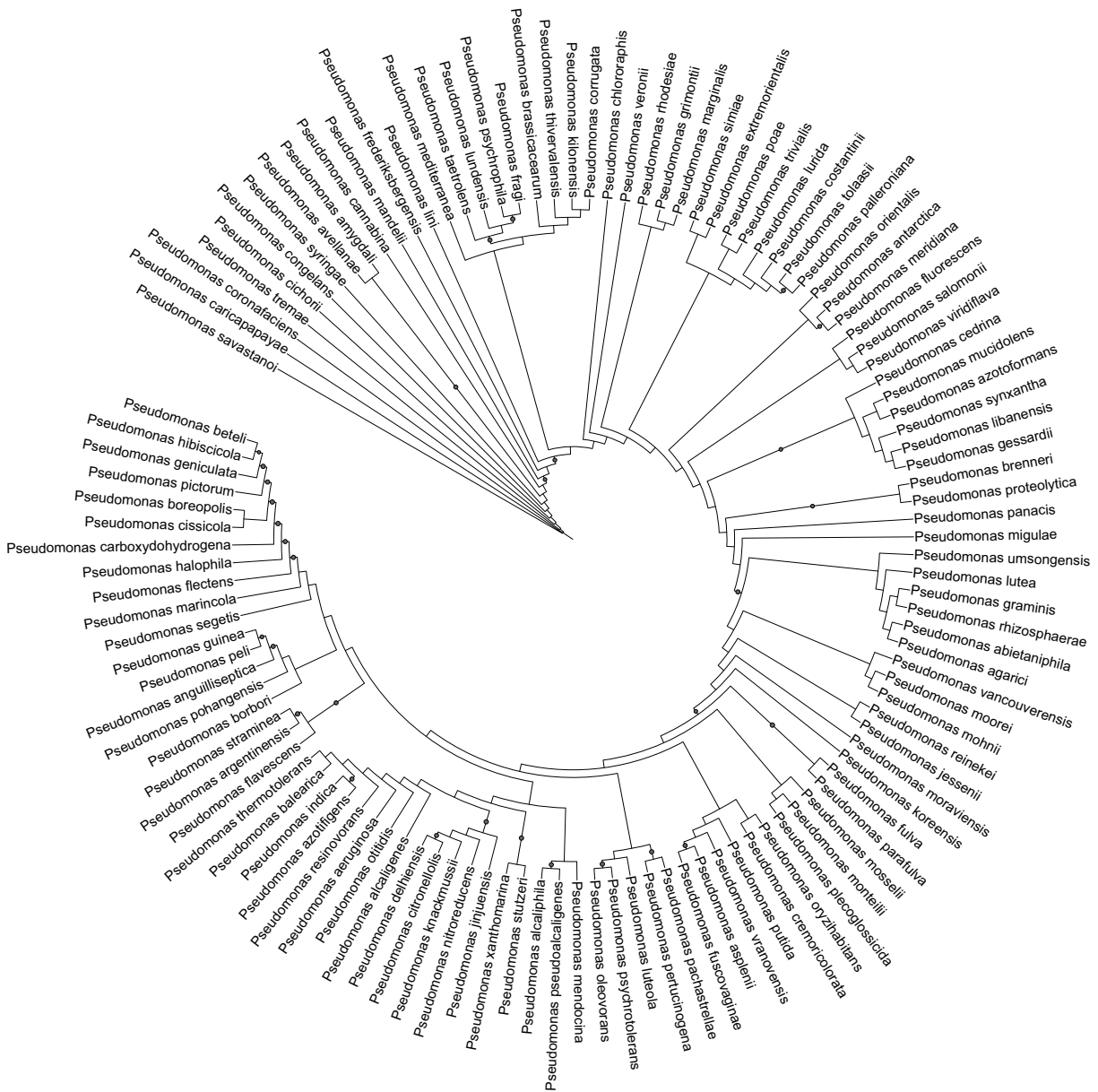


Figure 2.8: 16S rRNA gene sequence-based maximum likelihood tree of the genus *Pseudomonas*. The 117 species included correspond to the validly published taxonomy of May 2008. One high-quality 16S rRNA sequence per species is selected from the SILVA database (Pruesse et al., 2007). The phylogenetic tree is built by the maximum likelihood algorithm as implemented in the RAxML software (based on 1000 bootstraps) and is visualized with the iTol webtool (Stamatakis, 2006; Letunic and Bork, 2007). Tree branch lengths are ignored because of outlier species, that make the use of branch lengths pointless, and dotted branches correspond to bootstrap values larger than 75%.

2.3 Bacterial FAME Analysis

Following the introduction of gas chromatography by James and Martin in 1952, gas chromatographic (GC) fatty acid analysis of bacteria started around the year 1960 by investigation of *Bacillus subtilis* and a species of the genus *Sarcina* (Akashi and Saito, 1960; Saito, 1960a,b; Kaneda, 1963a). The first evidence that GC whole-cell fatty acid analysis could be used for the classification of bacteria was given by Abel et al. in 1963. Important to mention is that the researchers concluded that the advantages of speed and simplicity make lipid analysis a practical method for the classification of bacteria. A review paper on fatty acid analysis of bacteria in that period is due to O'Leary (1962). In 1991, Welch published a review concerning applications of cellular fatty acid analysis in which he concluded that GC FAME profiling offers considerable power for microbial identification because characteristic patterns of cellular fatty acids can be defined to the species level and results are rapidly achieved.

Through the last decades, different improvements for sample preparation and GC analysis were introduced (Moss et al., 1974, 1980; Moss, 1981; Miller, 1982; Lambert and Moss, 1983; Welch, 1991; Buyer, 2002a,b, 2003, 2006; MIDI, 2009a). Some of the first researchers suggesting automated GC FAME analysis for bacterial identification were Eerola and Lehtonen (1988). In 1991, the company Microbial ID Inc. (MIDI, Newark, Delaware, USA) launched a commercial bacterial identification system based on fatty acid profiling: the Sherlock Microbial Identification System (MIS). At present, Sherlock MIS is the reference system for bacterial FAME profiling. This system was evaluated for its identification power of different bacterial species and groups in different research papers (Stead et al., 1992; Steele et al., 1997). As bacterial FAME analysis became more popular, more and quite diverse bacterial groups were investigated (Mukwaya and Welch, 1989; Kotilainen et al., 1991; Osterhout et al., 1991; Welch, 1991; Stead et al., 1992; Kämpfer and Kroppenstedt, 1996; Steele et al., 1997; Heyrman et al., 1999; Song et al., 2000; Čechová et al., 2004; Pineiro-Vidal et al., 2008) and numerical analysis of the fatty acid profiles was introduced in later years (O'Donnell et al., 1985; Eerola and Lehtonen, 1988; Kämpfer, 1994; Kämpfer and Kroppenstedt, 1996). In the perspective of this work, it is important to mention that the prediction of bacteria based on FAME data by means of machine learning techniques was first described in 1993 and was further investigated by the same research group (Ruggiero et al., 1993; Bertone et al., 1996; Giacomini et al., 2000, 2004). While the researchers focused on identification at the genus level, they concluded that, as FAME data yields information at the species level, it would be worthwhile to build a FAME-based bacterial species identification system.

Interestingly, FAME profiling is also used beyond taxonomy. Keeping the focus on microbiology, the technique is also used for bacterial community typing (Haack et al., 1994; Glucksman et al., 2000; Quezada et al., 2007), microbial source tracking (Seurinck et al., 2005) and bacterial spore typing (Song et al., 2000).

2.3.1 Towards FAME Profiling

2.3.1.1 The Nature of Fatty Acids

In whole-cell fatty acid profiling, the main focus is set on the components of any cellular lipid with a carbon chain length of 9 to 20 atoms. This includes the majority of fatty acids located in the cell membrane as glycolipids and phospholipids, and the fatty acid constituents of lipopolysaccharides (in case of Gram-negative bacteria). The primary source of fatty acids in microbial cells is the cell membrane, with the LPS layer an additional main source in Gram-negative bacteria. The biosynthesis of fatty acids in bacteria is accomplished by the type II fatty acid synthetase system which relies on a highly conserved collection of enzymes. Main players are the molecule coenzyme A which esterifies the fatty acids and the acyl carrier protein (ACP). Most bacteria synthesize fatty acids with a chain length of 10 to 19 carbon atoms, and the most prevalent fatty acids are those with 16 to 18 carbon atoms. A usual whole-cell fatty acid profile constitutes 5 to 15 fatty acids. Branched-chain fatty acids predominate in some Gram positive bacteria, while short-chain hydroxy acids often characterize the lipopolysaccharides of the Gram negative bacteria (Sasser, 1990; Welch, 1991; Rock and Jackowski, 2002). Based on a plasmid and mutagenesis study, it is suggested that the fatty acid composition is highly conserved genetically and that significant changes take place only over considerable periods of time (Kunitsky et al., 2006). More than 300 fatty acids and related compounds have already been found in bacteria. The wealth of information contained in these compounds can be estimated by considering not only the presence or absence of each acid, but also by using the data in a quantitative fashion. The theoretical ability to differentiate amongst 2^{300} different combinations makes FAME analysis impractical. However, due to the non-random distribution within groups of bacteria, the huge number of fatty acids increases the power for describing an increasing number of bacterial taxa (Sasser, 1990). In his review paper, Welch (1991) states that FAME analysis allows for genus discrimination and characteristic FAME patterns can be found at the species level. This is underscored by the observation that FAME profiles at the genus level mostly show qualitative differences (peak presence) while at the species level mostly quantitative differences are found (peak ratios). In this perspective, machine learning is an excellent option when aiming at an improved bacterial identification. When focusing on lower taxonomic levels, typing (or subgrouping) depends on the species in question. Those with more complex profiles are more amenable for typing than those with low fatty acid variability (Welch, 1991; Kunitsky et al., 2006).

Where fatty acids were initially named to reflect the nature of their source, e.g. sarcinic acid originates from a *Sarcina* species (Akashi and Saito, 1960), currently, a nomenclature is used for naming fatty acids. Fatty acid names are based on the number of carbon atoms, the type of functional groups and the double-bond locations present in the molecule structure. The systematic name is simplified by writing a C followed by the number of carbon atoms to the left of a colon and the number of double bonds on the right. The letter ω indicates the double-bond position from the hydrocarbon-end of the chain (ω end), while the letters *c* and *t* indicate cis and trans configurations of the hydrogen atoms. The carboxyl-end (COOH) of the chain

is called the α end. Numbering of branched-chain, cyclopropane-containing and hydroxy fatty acids typically starts from the carboxyl end of the molecule, but as an alternative can also start from the α end. Iso- and anteiso-branched fatty acids are methyl-branched fatty acids at the second and third carbon from the ω -end of the carbon chain. These fatty acids are indicated by the prefixes iso and anteiso (Welch, 1991; Kämpfer, 2002). Other branched fatty acid structures also exist. Some examples are given in Figure 2.9

2.3.1.2 Culture and Growth Conditions

The influence of different conditions on the relative composition of the FAME profile is reported in different papers. The importance of culture and growth conditions was already pointed out by Abel et al. (1963) who stated that the chemical composition could provide a basis for classification given that the bacteria are grown under defined conditions. Kaneda stressed that the relative proportions of the fatty acids can vary depending on physiological and environmental/culture conditions (Kaneda, 1966a, 1967, 1971, 1977). He also showed that the profiles vary during the growth phase of bacteria. The effect of varying temperature and/or growth medium composition on the bacterial fatty acid content is studied by several researchers (Marr and Ingraham, 1962; Drucker and Veazey, 1977; Rilfors et al., 1978; Chung et al., 1993; Juneja and Davidson, 1993; Huys et al., 1997). For temperature, it is shown that the ratios between fatty acids changed with varying temperatures. Also, important to note is that Huys et al. (1997) recommended a thorough evaluation of the growth medium when designing a standardized FAME protocol. In general, the composition of the fatty acid profile varies quantitatively (peak area) according to growth medium, incubation duration, incubation atmosphere, temperature and chromatographic equipment. Furthermore and importantly, these factors have little impact on the qualitative fatty acid composition (peak presence) (Welch, 1991; Kämpfer, 2002). For the purpose of comparing fatty acid profiles from different strains, it is important that culture media and growth conditions are identical (Kämpfer, 2002). From these papers, it is clear that identifying bacteria based on FAME data requires standard culture and growth conditions, both for reproducibility of the FAME profiles as for comparative studies. In this work, we followed the protocol defined by the commercial Sherlock MIS, which is extensively described in the next subsection.

2.3.1.3 The Sherlock Microbial Identification System

2.3.1.3.1 Introduction

In 1987, Myron Sasser gained the rights to the fatty acid-based technology and created MIDI and the Sherlock MIS in 1991. The main fields of interest to MIDI are environmental and clinical microbiology. In 1998, the U.S. Centers for Disease Control and Prevention recognized the Sherlock MIS as an official method for aerobic bacterial identification. Sherlock MIS is the only system cleared by both the U.S. Department of Homeland Security and U.S. Food and Drug Administration for *B. anthracis* (anthrax disease) confirmation (MIDI, 2009b,c). In 2007,

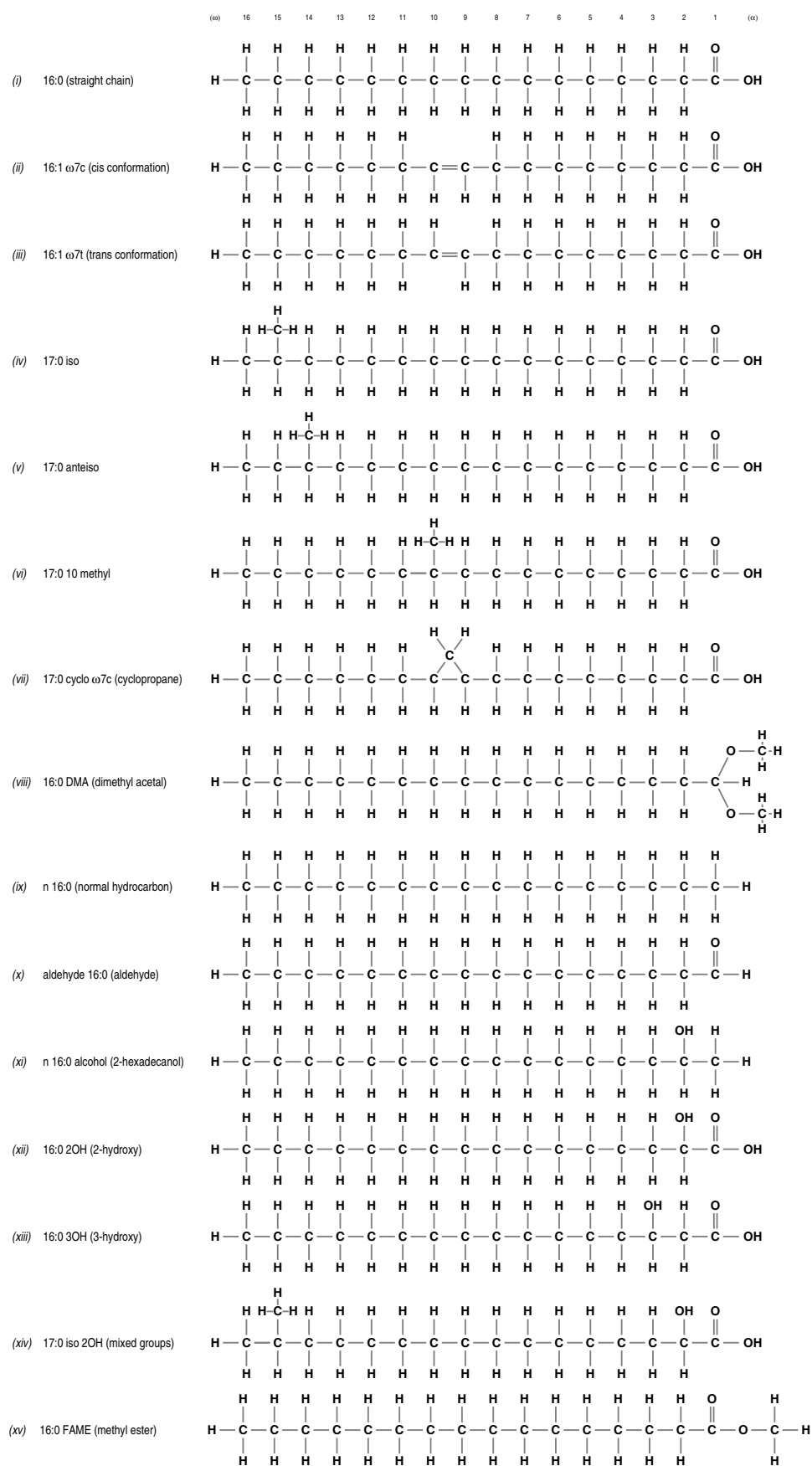


Figure 2.9: Nomenclature of fatty acids (Dawyndt, 2004).

MIDI Inc. came up with a new sample preparation method which allows identification in only 15 minutes time (MIDI, 2009a).

2.3.1.3.2 The Sherlock MIS Package

The Sherlock MIS consists of a HP 6890A gas chromatograph (GC; Hewlett-Packard Co., Avondale, Pennsylvania, USA) equipped with a flame ionization detector, a fused-silica capillary column (25m by 0.2 mm) coated with 5% phenyl methyl silicone (film thickness 0.33 μm ; HP Ultra2), automatic sampler and computer. Hydrogen is the carrier gas, nitrogen is the 'make-up' gas, and air is used to support the flame. Sherlock MIS uses the Agilent Chemstation (version 4.02, Hewlett-Packard) software for controlling sampling, analysis and integration of the chromatographic samples. Typical operating parameters of the system for fatty acid chromatography are as follows: injector temperature, 250°C; detector temperature, 300°C; and oven (column) temperature, regulated by a computer-controlled program which increases the temperature from 170 to 300°C at 5°Cmin⁻¹ and holds it at 300°C for 5 min prior to recycling. The flame ionization detector allows for a large dynamic range and provides good sensitivity. The electronic signal from the GC detector is passed to the computer where integration of the peaks is performed. Peak naming and identification library matching is performed by the Sherlock MIS software. Sherlock MIS has the distinct advantages that it is a sensitive and automated identification system, allowing high-throughput analysis, and sample preparation and GC analysis is cheap and rapid. Currently, the cost of FAME analysis by MIDI Inc. is \$60-75 per sample. (Sasser, 1990; Osterhout et al., 1991; Welch, 1991; Vancanneyt et al., 1996; MIDI, 2009b).

2.3.1.3.3 Sample Preparation Protocol

As mentioned before, changing culture and growth conditions can result in drastic changes of the FAME profile composition. To minimize these effects, standard protocols need to be followed accurately in order to achieve the highest stability and reproducibility possible in order to allow comparative analysis of the FAME profiles. The Sherlock MIS focuses on different microbial niches for identification of a wide range of bacteria e.g. clinical, environmental, industrial, veterinary, drinking/waste water, food, etc. Based on these niches, different identification libraries were developed for which reference strains were cultured and processed under controlled conditions. Examples are the TSBA and CLIN libraries for aerobes and clinical aerobes, the BHIBLA and MOORE libraries for anaerobes, and the YST, YSTCLN, FUNGI and ACTIN for the identification of yeasts, fungi and actinomycetes (MIDI, 2005b). Considering the genera *Bacillus*, *Paenibacillus* and *Pseudomonas* implies the aerobic environmental niche and the corresponding TSBA50 identification library. For this library, a specific sample preparation protocol is accurately followed. Most aerobic bacteria grow well on the Trypticase Soy Broth Agar (TSBA), which consists of 30 gl⁻¹ Trypticase Soy Broth (BBL) supplemented with 15 gl⁻¹ of Bacto Agar (Difco) (Vancanneyt et al., 1996). Sherlock MIS recommends the streaking plate method for culturing of bacteria. Herein, four quadrants with decreasing cell densities are created on the agar plate. The standard incubation conditions for aerobes 28°C and 24h. In

the case of slow-growing organisms, extended incubation may be necessary to obtain quantitative reproducibility and to achieve sufficient cell mass for analysis (MIDI, 2005b). For bacteria not showing an optimal growth, the conditions may be altered given that this deviation from the standard protocol is described in detail in a log book or in a database. In most cases, this concerns the afore-mentioned elongated incubation duration or an adapted incubation temperature. An example of this latter case is the use of a higher temperature e.g. 52°C for the growth of thermophiles.

Following growth of the bacteria, fatty acids are extracted for subsequent GC analysis. The usual preparation of extracts consists of hydrolysis of the whole-cell fatty acids and subsequent methylation of the fatty acid esters to make them volatile in the gas chromatograph. Extraction and derivation of the different fatty acids from the bacterial cells is achieved by the method as described by Miller (1982). Briefly, GC ready extracts are prepared in the following five steps. Approximately 40 to 50 mg (wet weight) of bacterial cells is harvested from the streaked plate, and placed into a tube (13 by 100 mm) with a Teflon-lined screw cap. The most stable fatty acid compositions are obtained from cultures in the late log phase of the growth and this corresponds typically to organisms present in quadrant three of the plate. Next, cells are saponified by heating them at 100°C for 30 min following the addition of 1.0 ml of 15% NaOH (w/v) in 50% aqueous methanol (v/v). This kills and lyses the bacterial cells, and liberates the fatty acids from cellular lipid. Subsequently, the hydrolysate is cooled to ambient temperature, 3.25 N of HCl in 45.8% methanol is added, and the mixture is heated at 80°C for 10 min (this step is critical in time and temperature). As a result, the pH of the solution drops below 1.5 and causes methylation of the fatty acid, required for an increased volatility in a partially polar column. Next, the methylated fatty acids are quickly cooled down to ambient temperature and extracted through the addition of 1.25 ml of hexane and methyl tertiary butyl ether (1:1 v/v), after which the tubes are capped and gently mixed for 10 minutes. This step removes the fatty acid methyl esters from the acidic aqueous phase and transfers them to an organic phase. Subsequently, the lower aqueous phase is pipetted out and discarded. Finally, in order to reduce contamination of the injection port liner, the column and the detector, the sample is washed by adding 1.2% of dilute NaOH (w/v) to the remaining organic layer. This base washing removes free fatty acids and residual reagents from the organic extract, which will degrade the column and distort the peak shape of hydroxy fatty acids in subsequent runs. Approximately two-thirds of the organic layer containing the fatty acid methyl esters (FAMES) is then transferred to a septum-capped sample vial for GC analysis. The entire sample preparation process takes about 1h (Sasser, 1990; Osterhout et al., 1991; Welch, 1991; Dawyndt, 2004; MIDI, 2005b). This procedure is illustrated and summarized in Figure 2.10. Following GC analysis, the extracted FAMES are named using the Sherlock MIS peak naming software.

In this perspective, it is important to mention that this protocol clearly shows the main disadvantage of FAME-based bacterial identification. Bacterial strains are required to grow on plates following specific culture and growth conditions, even though MIDI Inc. states that the Sherlock MIS libraries were developed by selecting conditions that are most favorable for a majority of microorganisms. It is clear that this restricts the bacterial scope of the identification technique drastically. Nonetheless, in this work, we dealt with bacterial strains that allow

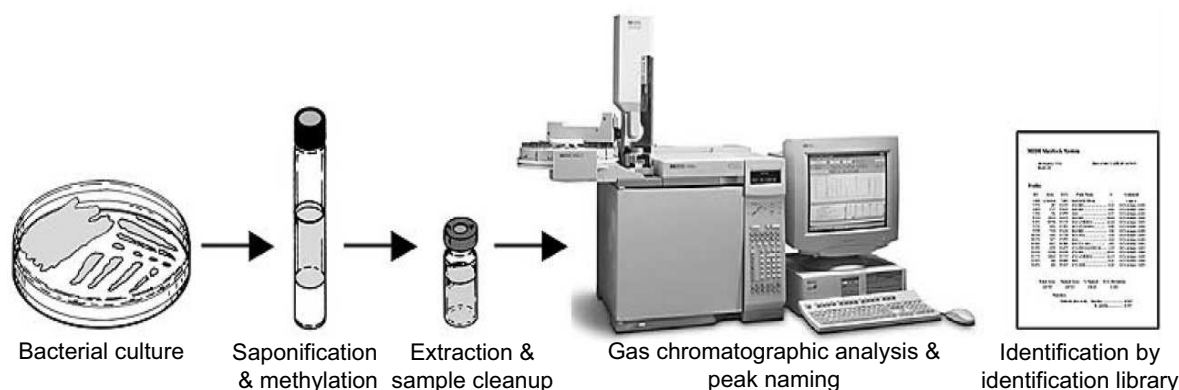


Figure 2.10: The Sherlock Microbial Identification System workflow (Sasser, 1990).

identification following the described growth and sample preparation protocol.

2.3.1.3.4 Calibration and Peak Naming

Good laboratory practice regarding culture/growth conditions and sample preparation, and proper GC operation may still lead to an unsuccessful FAME profiling. For instance, defects or disorders in equipment can still result in distorted profiles. The peak naming methodology of Sherlock MIS uses the composition of a calibration standard to continually monitor the health of the system. The standard is a mixture of the straight-chain saturated fatty acids from 9 to 20 carbons in length ($C_{9:0}$ to $C_{20:0}$) and five hydroxy acids. All compounds are added quantitatively so that the GC performance may be evaluated by the software each time the calibration mix is analyzed. The hydroxy compounds are especially sensitive to changes in pressure/temperature relationships, to contamination in the injection port liner and to column degradation. These can result in poor peak shape (peak tailing) or in a loss of the hydroxy acid peak area. As a result these compounds function as quality control checks for the system. When a calibration analysis is completed, the computer checks the results against the peak naming table for a specific number of peaks and a pattern of retention times and area percent amounts. Deviations from the expected values result in a failure to calibrate, and a warning message. A calibration runs twice at the beginning of every batch and is automatically reanalyzed after every 11th (default) sample injection. Each sample batch also contains a positive control, and a reagent blank (containing no bacteria) as a negative control. As positive control for the TSBA library, MIDI Inc. recommends the strain *Stenotrophomonas maltophilia* LMG 958^T because the strain has a complex fatty acid profile that is diagnostic for problems throughout all stages of sample preparation. These two quality control samples are analyzed after calibration and before any other batch samples. The negative control is diagnostic for reagent contaminants (Sasser, 1990; MIDI, 2005b; Kunitsky et al., 2006; MIDI, 1990).

FAMES are identified on the basis the so-called equivalent chain length (ECL) units. This value is a representation of a fatty acid's retention time which is related to a set of straight-chain saturated FAMES ($C_{9:0}$ to $C_{20:0}$) present in the calibration mixture. The ECL value is equal to the number of carbon atoms present in a straight-chain saturated fatty acid, e.g. $C_{11:0}$ has an

ECL value of 11.000. So, a second and quantitative function of the calibration standard is to provide accurate retention times for these straight-chain FAMES. Based on the ECL values of the peak naming table entries for all peaks in the calibration mix, a 'nominal' retention time for each peak is calculated. Ultimately, the ECL value of each FAME peak x in each batch sample can be calculated by linear interpolation of its elution time in relation to the elution times of the reference FAMES as given by

$$\text{ECL}_x = n + \frac{\text{RT}_x - \text{RT}_n}{\text{RT}_{n+1} - \text{RT}_n}, \quad (2.1)$$

where RT_n is the retention time of the straight-chain saturated FAME with n carbon atoms, preceding x in the calibration mix. RT_{n+1} is the retention time of the straight-chain saturated FAME eluting after x in the calibration mix. Sherlock MIS allows detection at 0.010 ECL units, which ensures a great precision in the resolution of fatty acid isomers (i.e. compounds with the same formula but with a different molecular structure) (Osterhout et al., 1991; MIDI, 2005b; Kunitsky et al., 2006; MIDI, 1990).

ECL calculation is followed by a match of the ECL values against the naming windows of a peak naming table. The Sherlock software compares the ECL of each peak in the batch sample with the expected ECL of the fatty acids in the peak naming table. Peaks that do not correspond to ECL values of known fatty acid peaks are left unnamed and are not further considered. In this perspective, it is important to remark that the accuracy of naming fatty acid peaks by comparing retention times with those of a known mixture is high but definitive identification can only be made by mass spectrometry. Practical constraints like column length or limited run time force acceptance of a less than perfect chromatography. Thus, we are not dealing with an ideal situation in which all peaks are clearly resolved and no data is lost due to some inability of the chromatographic separation process. Also, some peaks are not clearly delineated from one or more neighbour peaks, resulting in overlapping naming windows. The Sherlock approach defines a so-called summed feature wherever imperfect peak separation occurs. For further data analysis and comparison, this summed feature is regarded as an entity equivalent to clearly delineated FAME peaks. Note also that the names of a small number of FAME peaks have not yet been resolved by mass spectrometry, leading to the peak naming table entry 'unknown' followed by the corresponding ECL value. After naming the peaks in an unknown sample, Sherlock MIS compares the ECL values for the most stable FAMES to the peak naming table's theoretically perfect values and may recalibrate internally if sufficient differences are detected (Sasser, 1990; Welch, 1991; Dawyndt, 2004; MIDI, 2005b, 1990).

Once the naming process of the different chromatographic peaks is finished, only the named FAME peaks are further considered. For each batch sample, a composition report is generated covering several parameters of all peaks named in the chromatogram. For each of these peak, the retention time, area, area/height, response factor, ECL value, peak name, the relative amount and some calibration information is listed. When one or more summed feature are present, the names of the constituting compounds are reported as a comment. Also, the error between the actual ECLs and the expected ECLs is reported (denoted as ECL deviance), which is a measure of the system accuracy in naming peaks. The last parameter is the Reference ECL shift which

reports the drift from the last calibration and is a measure of the chromatographic stability. An example of this report with the chromatogram is given in Figure 2.11. The area under the peak reflects the relative amount of the individual fatty acids. The amount of fatty acids is calculated as a percentage of the total amount. Calculation of this relative percentage is, however, not that simple. Peaks in the early part of the analysis are more affected by GC oven temperatures and those later in the analysis are more severely impacted by carrier gas flow rates. The use of an electronic pressure controller for achieving constant flow minimizes this error. This implies that the first peaks will somewhat be underestimated, while peaks at the end of the chromatogram are slightly overestimated. To obtain an objective approximation of the relative fatty acid amount a_i^r of the i^{th} named peak, a weighted percentage is calculated by the formule

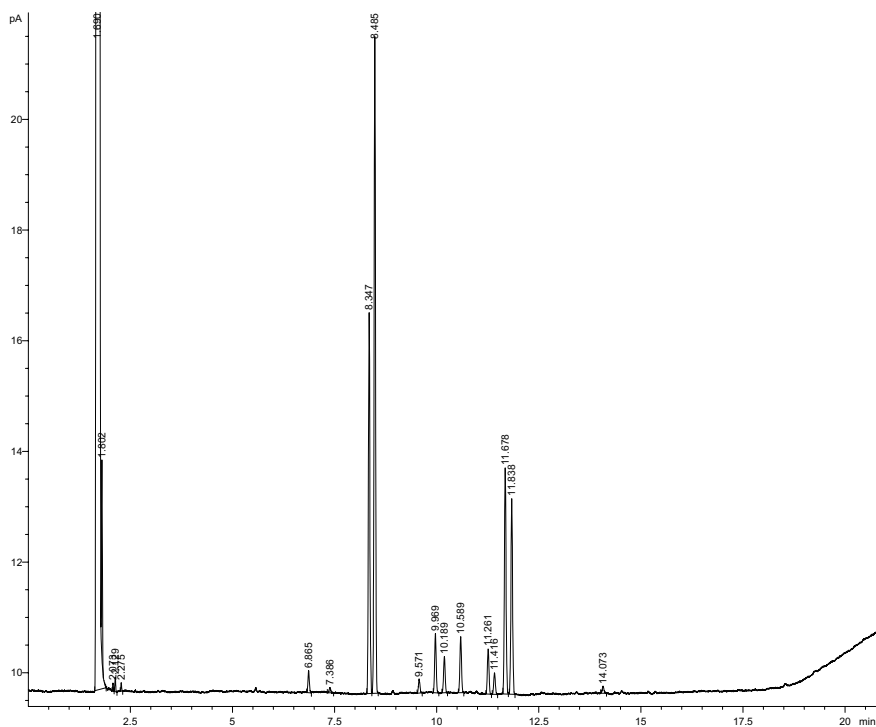
$$a_i^r = \frac{r_i a_i^a}{\sum_{j \in \mathcal{N}} r_j a_j^a}, \quad (2.2)$$

where \mathcal{N} represents the set of named peaks of the profile, a_j^a is the absolute area of the j^{th} peak and r_j is the weight factor assigned according to the ECL position of the j^{th} peak. In the FAME profiles report, these weights are denoted as the response factors and their value is derived from the calibration. As such, the response factor adjustment corrects the area counts for long-term drift and instrument-to-instrument variation. Note that from Eq. (2.2), it follows that the sum of the relative fatty acid amounts in a given profile equals 1. Besides these parameters, the profile report also informs the user about the total area count (denoted as total response) of peaks eluting at or between C_{9:0} and C_{20:0}, relating to all extracted fatty acids, the total area of all named peaks (denoted as total named), the percentage named (denoted as percent named) and the final total amount of named peak (denoted as total amount). A final remark to be made is that the fatty extraction procedure may carry over sterols and other non-fatty acid materials. So, electronic noise may result in transient spikes, which might interfere with the chromatographical process. Fatty acid peaks are report to have area/height ratios greater than 0.017 and less than 0.070, making it possible to set thresholds at these levels. Electronic noise spikes typically correspond to area/height ratios less than 0.017 and non-fatty acids peaks (e.g. sterols) usually correspond to ratios greater than 0.070, allowing rejection of these artifacts (Sasser, 1990; Welch, 1991; Dawyndt, 2004; MIDI, 2005b, 1990).

In this work, the TSBA50 peak naming table was used which consists of 135 naming windows. By the definition of 7 summed features, these windows cover 117 FAME peaks. The same number of windows, summed features and FAME peaks is present in the earlier version of this peak naming table, the TSBA40 peak naming table. An overview of the entries of the TSBA50 peak naming table is reported in Appendix B.1.

2.3.1.3.5 Identification Library

Once a microbial strain has been cultured, processed and analyzed by the Sherlock MIS, the fatty acid fingerprint can be matched with a specific identification library. Due to quantitative and qualitative shifts in the FAME profile by changing growth conditions, only comparisons between FAME profiles resulting from the same conditions make sense. Therefore, it is impor-



E059054.61A [19147] BACI-SUBTI(LMG7135T/B407/P37)

Volume: DATA2 File: E059054.61A Samp Ctr: 15 ID Number: 19147
 Type: Samp Bottle: 7 Method: TSBA50
 Created: 9/5/2005 4:47:56 PM
 Sample ID: BACI-SUBTI(LMG7135T/B407/P37)

RT	Response	Ar/Ht	RFact	ECL	Peak Name	Percent	Comment1	Comment2
1.690	5.042E+8	0.028	----	6.992	SOLVENT PEAK	----	< min rt	
1.802	10846	0.025	----	7.211		----	< min rt	
2.073	420	0.023	----	7.738		----	< min rt	
2.129	755	0.022	----	7.848		----	< min rt	
2.275	391	0.020	----	8.133		----	< min rt	
6.865	1891	0.038	0.986	13.619	14:0 ISO	1.19	ECL deviates 0.000	Reference -0.001
7.386	504	0.045	0.978	13.999	14:0	0.32	ECL deviates -0.001	Reference -0.002
8.347	34537	0.040	0.967	14.623	15:0 ISO	21.37	ECL deviates 0.000	Reference -0.001
8.485	59681	0.040	0.966	14.713	15:0 ANTEISO	36.88	ECL deviates 0.000	Reference -0.001
9.571	1318	0.039	0.958	15.387	16:1 w7c alcohol	0.81	ECL deviates 0.000	
9.969	5663	0.041	0.955	15.627	16:0 ISO	3.46	ECL deviates 0.000	Reference -0.001
10.189	3569	0.043	0.954	15.759	16:1 w11c	2.18	ECL deviates 0.002	
10.589	5583	0.042	0.952	16.000	16:0	3.40	ECL deviates 0.000	Reference -0.001
11.261	4550	0.045	0.949	16.389	ISO 17:1 w10c	2.76	ECL deviates 0.001	
11.416	2149	0.044	0.949	16.479	Sum In Feature 4	1.30	ECL deviates 0.003	17:1 ISO I/ANTEI B
11.678	23156	0.044	0.948	16.630	17:0 ISO	14.04	ECL deviates 0.000	Reference -0.002
11.838	19488	0.043	0.947	16.722	17:0 ANTEISO	11.81	ECL deviates -0.001	Reference -0.002
14.073	781	0.050	0.942	18.000	18:0	0.47	ECL deviates 0.000	Reference -0.003
----	2149	---	----	----	Summed Feature 4	1.30		17:1 ISO I/ANTEI B

ECL Deviation: 0.001
 Total Response: 162871
 Percent Named: 100.00%

Reference ECL Shift: 0.002 Number Reference Peaks: 9
 Total Named: 162871
 Total Amount: 156304

Matches:

Library	Sim Index	Entry Name
TSBA50 5.00	0.873	Bacillus-subtilis

Figure 2.11: Sherlock MIS example report. Report of the analysis of the whole-cell FAME composition of *Bacillus subtilis* LMG7135^T. The report consists of three main parts: the chromatogram with the different FAME peaks, a report with detailed peak information and an identification report. In this report, only one entry was given for the identification of *Bacillus subtilis* LMG7135^T.

tant that identifications are only performed by a library built upon the same culture and growth condition protocol. In this work, we used the TSBA50 identification library. The Sherlock MIS identification libraries consist of more than 100,000 analyses of strains obtained from experts and from culture collections. The cultures were collected from around the world to avoid a potential geographic bias. However, the scope of the Sherlock MIS libraries is limited due to MIDI's inability to obtain adequate numbers of strains. To provide normal species variability, where possible, 20 or more strains of a species or subspecies were analyzed to make the entry. When, due to high intra-taxon variability, chromatographic subgroups (so-called GC groups) were found within a taxon, more strains were obtained to delineate each GC group. Each group is considered as a separate library entry. Therefore, MIDI Inc. states that its libraries are carefully developed to take inter-strain and experimental variation into account. In view of identification by machine learning techniques, this is a quite important remark to make. Due to variations in the FAME profiles, generalization is critical for an adequate identification. Despite MIDI Inc. ensures identification is based on normal species variability and that 20 or more strains of a species are integrated, however, users are not informed about this critical inter- and intra-species variability (Sasser, 1990; MIDI, 2005b; Kunitsky et al., 2006). In the combined Sherlock libraries, there are nearly 2,000 microbial species, including 700 environmental aerobic species, 620 anaerobic species and 200 species of yeasts. However, when looking more into detail to the genera covered in this study, no major update was seen in the libraries. The upgrade from the TSBA40 (783 entries) to the TSBA50 (888 entries) identification library showed an increase of 3, 5 and 10 new species in the genera *Bacillus*, *Paenibacillus* and *Pseudomonas*, respectively, and a removal of 5 *Bacillus* species and 2 *Pseudomonas* species (MIDI, 2003, 2005c). The upgrade from the TSBA50 to the TSBA6 identification library showed only an increase of 3 new *Bacillus* species and the removal of 1 *Pseudomonas* species (MIDI, 2005a,c). As the upgrade of TSBA40 to TSBA6 took 2 years and as the bacterial landscape is monthly changing (see also Figures 2.3, 2.5 and 2.7), it can be concluded that, from a taxonomic perspective, no major changes were implemented for the tree genera. Note also that not all library entries correspond to one species (GC groups and species groups) and that not all included species are valid according to the List of Prokaryotic names with Standing in Nomenclature (LSPN) (Euzéby, 1997). An overview of the entries for the genera *Bacillus*, *Paenibacillus* and *Pseudomonas* is given in Appendix B.2. Of course, one straightforward solution is to create personal custom identification libraries. MIDI Inc. allows users to do this by their Sherlock MIS software by the Library Generation System package. Also here, it is important that the FAME profiles of all included strains should be generated following the same culturing method and sample preparation protocol (Kunitsky et al., 2006).

Identification of unknown samples by the Sherlock MIS is based on a so-called Similarity Index (SI). The SI is a numerical value in the interval [0,1], which expresses how closely the fatty acid composition of an unknown sample matches with the mean fatty acid composition of the strains used to create the library entry listed as its match. Thus, the SI is not a probability or percentage but an expression of the relative distance of the unknown sample from the population mean. An SI value of 1.0 indicates a perfect match with the taxon associated with the library entry. The SI assumes that fatty acid distributions for species of microorganisms are normally

distributed and that the mean of the population characterize the taxon represented by the entry. As each fatty acid varies from the mean percentage, the SI will decrease proportionally to the cumulative variance between the composition of the unknown sample and the library entry (Kunitzky et al., 2006). The SI of a particular fatty acid x compared to the population mean of the library entry A is given by

$$\text{SI}(x, A) = e^{-(\alpha d)^2}, \quad (2.3)$$

where $\alpha = 3$ (corresponding to an SI of 0.600) and d is a particular distance. In calculating a distance between fatty acid profiles two important facts have to be considered. First, FAMES do not necessarily have the same variances, which makes, for instance, the use of the Euclidean distance unsuitable. As an example, assume that the patterns in a library entry show relative areas that are densely concentrated around the value t_k for fatty acid k , while for fatty acid l the same patterns are within a much wider interval around the value t_l . Then the distance d between x_k and t_k is much more significant than the same distance measured between x_l and t_l . Euclidean distance does not take into account this possible asymmetry. Another problem with Euclidean distance and related measures is a bias in the similarity value towards the major fatty acids, as these will have a larger impact on the global similarity or dissimilarity than the minor fatty acids. The latter fatty acids may, however, also have a large discriminatory power. A solution for this problem is to normalize the distance with respect to the variance. Second, fatty acids are not independent features. Fatty acids are synthesized by a biosynthesis pathway in which fatty acids are converted to other fatty acids (due to growth and/or a temperature shift, e.g. $C_{16:0}$ to $C_{16:1}$ due to the action of a desaturase). These dependencies can be described by an $m \times m$ covariance matrix which captures the mole-to-mole relationship of the conversion of one fatty acid to another, with m the number of fatty acids present in the peak naming table. Finally, a solution accounting for both the variances and the covariances of fatty acids is the Mahalanobis distance. This distance between the fatty acid profile x of the unknown sample and the mean fatty acid profile μ of the library entry can be expressed using the standard formula for multivariate Gaussian (or normal) distance given by

$$d^2(x, \mu, \Sigma) = (x - \mu)\Sigma^{-1}(x - \mu)^T, \quad (2.4)$$

where Σ is the respective covariance matrix and Σ^{-1} its inverse. $(x - \mu)^T$ is the transposed column vector of $(x - \mu)$. If fatty acids would be independent of each other, the non-diagonal elements (covariances) of the covariance matrix would be zero and the Mahalanobis distance would become the normalized Euclidean distance. When the variances of the fatty acids would also be equal, the Mahalanobis distance would be reduced to the Euclidean distance. One major problem is related to the calculation of the Mahalanobis distance: the distance only exists if Σ^{-1} can be calculated, i.e. if Σ is non-singular (or, the determinant is not zero). To overcome this problem, Sherlock MIS uses a technique based on an eigenvalue-eigenvector analysis of the covariance matrix. Herein, a small value is added to the eigenvalues to avoid that some eigenvalues are zero or close to zero. To get the inverted matrix Σ^{-1} , the adjusted eigenvalues are inverted and rotated by the eigenvector matrix (Sasser, 1990; Osterhout et al., 1991; Dawyndt, 2004).

MIDI defines three interpretation guidelines when using their Sherlock MIS system. Samples with an SI of 0.5 or higher and with a separation of 0.1 between the first and the second match are considered good library comparisons. If the SI is located in the interval [0.3,0.5] and is well separated from the second match, it may be a good match but an atypical strain. Values lower than 0.3 suggest that the organism is not a species in the library. Of course, if matches are reported, these indicate the most closely related entries in the identification library (MIDI, 2005b; Kunitsky et al., 2006). An example of an identification report is given in Figure 2.11.

2.3.1.4 In-house FAME Database

Since 1989, Sherlock MIS has been used for FAME profiling at the Laboratory of Microbiology (Ghent University, Belgium) and the BCCM™/LMG Bacteria Collection (Ghent University, Belgium). Following culturing, sample preparation, GC analysis, peak naming and identification, FAME profiles are stored in an Oracle database management system (Oracle Corporation, Redwood Shores, CA, USA). At present, twenty years of FAME analysis have resulted in more than 71,000 FAME profiles. The evolution in the number of generated FAME profiles is visualized in Figure 2.12. However, no internal quality control system is implemented on top of the database, meaning that not all data are suitable for data analysis. All information of the Sherlock MIS peak naming tables and identification libraries is also stored in this in-house database. Where Sherlock MIS is quite restrictive for extensive data analysis, this approach allows the use of third-party data analysis and data mining software packages and, thus, allowing a broader range of possibilities for FAME data analysis. For computational data analysis, the software package BioNumerics (Applied Maths, Sint-Martens-Latem, Belgium) is in use for many years. This software package was chosen because it allows easy connection with database management systems by the Open Database Connectivity (ODBC) protocol and offers a wide range of data mining possibilities. Regarding machine learning, the used software version implemented only very basic machine learning techniques (clustering and ANNs). Nonetheless, for a machine learning approach in computational FAME analysis this software package offered a good starting point in the selection of FAME profiles and the creation of large data sets.

2.3.2 FAME Analysis of Species in the Genera *Bacillus*, *Paenibacillus* and *Pseudomonas*

2.3.2.1 *Bacillus*

Analysis of the fatty acid content of the members of the genus *Bacillus* started somewhat in the same year as Abel et al. (1963) suggested the practice of fatty acid analysis in bacterial classification. One of the first research papers on fatty acid analysis in the genus *Bacillus* concerned *B. subtilis* (Saito, 1960a). Later, one of the main players in the research on the fatty acid content of *Bacillus* species was Kaneda (Alberta Research Council, Edmonton, Canada). His work is mainly described in the series ‘Biosynthesis of Branched-chain Fatty Acids’ (Kaneda, 1963a,b, 1966a,b) and ‘Fatty Acids in the Genus *Bacillus*’ (Kaneda, 1967, 1968), and in his later review papers (Kaneda, 1977, 1991). Further research concerned the fatty acid content of an increasing number of species together with the effect of bacterial growth conditions (Shen et al.,

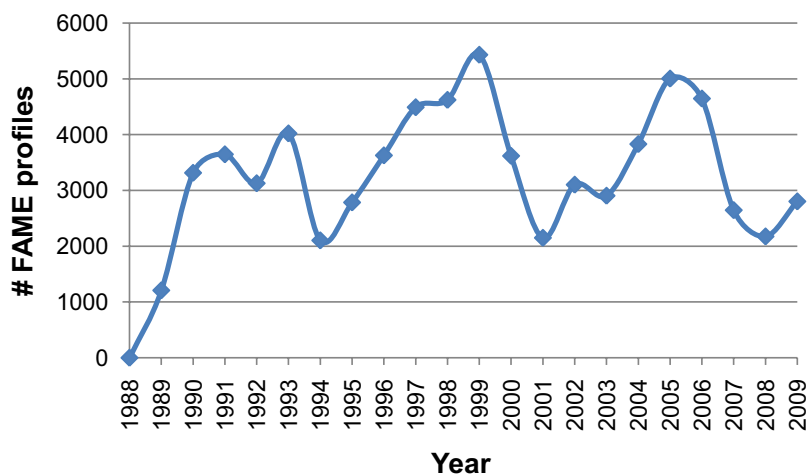


Figure 2.12: Trend in whole-cell FAME analysis. Number of whole-cell FAME profiles generated at the Laboratory of Microbiology and BCCMTM/LMG Bacteria Collection (Ghent University, Belgium) from the start in 1989 to 21/11/2009. The total number of FAME profiles equals 71,267.

1970; Kaneda, 1972; Rilfors et al., 1978). Numerical analysis of fatty acid data was compiled in a review paper on about 30 *Bacillus* species (Kämpfer, 1994). In this work, an enormous heterogeneity in the fatty acid profiles was demonstrated and the genus *Bacillus* could be subdivided in seven FAME clusters with even several fatty acid biotypes within one species. Even though the genus *Bacillus* has known substantial changes in its taxonomy (splitting off into new genera and a huge increase in the number of new species), since the paper of Kämpfer (1994) no genus-wide study of the fatty acid content of the genus *Bacillus* has been performed. In the perspective of the systematics of the genus *Bacillus*, one of the latest reviews on whole-cell fatty acid analysis is also given by Kämpfer (2002). Predominant fatty acids in *Bacillus* are C_{14:0}, iso and anteiso C_{15:0} and anteiso C_{17:0} (Logan and De Vos, 2009). More detailed information about the fatty content of the different *Bacillus* species can be found in the references of the papers mentioned above, in the different papers with species descriptions and in Bergey's Manual for Systematic Bacteriology (Logan and De Vos, 2009).

2.3.2.2 *Paenibacillus*

The genus *Paenibacillus* was split off from the genus *Bacillus* (Ash et al., 1993). According to the description of *Paenibacillus* gen. nov., fatty acids are primarily long-chain cellular fatty acids of the straight-chain saturated, anteiso- and iso-branched types with a predominance of anteiso C_{15:0} generally comprising around 55% but ranging between 34%-80%. Iso C_{15:0}, iso C_{16:0} and C_{16:0}, anteiso C_{17:0} generally comprise the remainder of the fatty acids (Ash et al., 1993; Priest, 2009). Heyndrickx et al. (1996) performed a numerical FAME analysis of 11 species of the genus *Paenibacillus*. Also for this genus, a clustering of FAME profiles resulted in the distinction of several species groups. The major groups corresponded well with the species composition of two major amplified ribosomal DNA restriction analysis (ARDRA) clusters. For more detailed information about the FAME content of the different *Paenibacillus* species see also the corresponding species description and to Bergey's Manual for Systematic Bacteriology (Priest, 2009).

2.3.2.3 *Pseudomonas*

Pseudomonas belongs to the Gram-negative bacteria. This implies that initial research on the fatty acid in this genus was focused on the corresponding LPS layer. As mentioned before (see Subsection 2.3.1.1), this LPS layer is responsible for an important discriminatory fraction of hydroxy fatty acids. It was initially shown for *Pseudomonas aeruginosa* by several researchers that the major fatty acid fraction of the LPS layer was constituted 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* (often referred to as the "Pseudomonads") was performed at the laboratory of Moss (Center for Disease Control, Atlanta, USA; (Moss et al., 1972; Moss and Samuels, 1974; Dees and Moss, 1975; Moss and Dees, 1975; Moss, 1975; Moss and Dees, 1976; Dees et al., 1979; 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 on improvements 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 and Komagata (1983). In the early 1990s, the first evaluations of Sherlock MIS were also performed with *Pseudomonas* species (Osterhout et al., 1991; Stead, 1992; Stead et al., 1992), of which 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). Importantly, 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 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-OH, C_{12:0} 2-OH and C_{12:0} 3-OH are predominant (Palleroni, 2005). For more detailed information about the FAME content of the different *Pseudomonas* species see also the corresponding species descriptions and to Bergey's Manual for Systematic Bacteriology (Palleroni, 2005).

PART II

DATA MINING
&
MACHINE LEARNING

CHAPTER 3

Data Analysis

The beginning of knowledge is the discovery of something we do not understand

FRANK HERBERT

3.1 Introduction

With an eye on performing a more extensive computational analysis such as machine learning experiments, it is always good to have a first look at the data at hand. Different questions can initially be asked. How do the features look like and how do they relate to each other? It is not only interesting to look within each species but also to analyze how the data differs between different species. Therefore, as a first step in our computational analysis, we focus on the extraction of this basic knowledge from the data sets.

We first describe our in-house FAME database and how the different FAME profiles were selected and exported from this database. Next, the different data sets are discussed. The main part of this section encompasses the results of our experiments with four basic data analysis techniques. With the calculation of average FAME profiles, the data is analyzed at species level. Clustering of the data shows relations between profiles, species and peaks. A TaxonGap analysis is done to visualize distance between species based on FAME data. And, finally, a principal component analysis is performed to visualize the data in a lower dimensional space and to further evaluate the variability in the data.

3.2 Data Selection

The joint in-house FAME database of the Laboratory of Microbiology (Ghent University, Belgium) and the BCCM™/LMG Bacteria Collection currently contains more than 71,000 FAME profiles. From this database, using the BioNumerics (Applied Maths, Sint-Martens-Latem, Belgium) data analysis software (versions 4.6 and 5.1), FAME data sets were created for the three genera *Bacillus*, *Paenibacillus* and *Pseudomonas*. In order to start from high-quality data sets, the sampled data sets needed a subsequent manual inspection. This was mainly necessary due to the presence of FAME profiles of low quality, FAME profiles of non-public LMG strains and FAME profiles of non-validly described species. A customized selection strategy was designed, which comprises the following points:

1. For identification purposes, it is essential to integrate only validly described bacterial species and to aim at a genus-wide scope. From the list of validly described bacterial species, all *Bacillus*, *Paenibacillus* and *Pseudomonas* species were selected. In this study, we used the bacterial Nomenclature Up-to-Date of the German Collection of Microorganisms and Cell Cultures (DSMZ, 2009), which corresponds to the List of Prokaryotic Names with Standing in Nomenclature (Euzéby, 1997) and the NCBI Taxonomy Browser (NCBI, 2009b).
2. All FAME profiles corresponding with the selected valid species were selected and exported from the database using the BioNumerics software package. This selection comprises FAME profiles of both public and non-public LMG strains, and research strains. Research strains are strains that are not deposited in the BCCMTM/LMG Bacteria Collection and are denoted in the FAME database by the letter 'R', a hyphen and a unique number.
3. Because of the above-mentioned problems, further fine-tuning of the FAME profile selection was required. The following criteria were imposed:
 - Removal of those FAME profiles that were not generated in accordance with the growth and culture conditions as defined by a certain peak naming table. In this study, we used the Sherlock MIS TSBA50 peak naming table. The corresponding conditions are a growth incubation temperature of 28°C, a growth duration of 24h, an aerobic growth atmosphere and the TSA growth medium. The BCCMTM/LMG Bacteria Collection defines the TSA growth medium as LMG Medium 185. Nonetheless, some species do not grow under these standard conditions. This is mostly due to a too low temperature and/or short duration. Thus, exceptions were allowed for some species. A nice example is the moderate thermophile *Bacillus thermoamylovorans* that has an optimal growth temperature of 50-52°C (Combet-Blanc et al., 1995). Only FAME profiles of this species were sampled corresponding with growth at 52°C. Other examples are some *Paenibacillus* species that require a growth duration of 48h in order to have a sufficient amount of biomass for good GC analysis. Of course, deviations from the standard conditions need to be explicitly described and reported, as only unknown FAME profiles grown under the same conditions will become validly identified. In the database, each FAME profile is also linked to a description field, reporting additional information regarding the adopted growth and culture conditions. However, not all FAME profiles are additionally described by this field. Based on this information, profiles corresponding with aberrant conditions could also be removed. Note that most FAME profiles generated before 2001 are not adequately annotated and are considered to be standard profiles. Based on the additional description, also in this case, aberrant profiles could be removed.
 - Check of consistency in growth and culture conditions for each species separately. Because of quantitative variations with changing environmental conditions, only one single value or type of temperature, duration, atmosphere and medium was allowed. Again, this was a requirement for making valid identification possible.

- Based on the strain catalog of the BCCMTM/LMG Bacteria Collection, only those profiles corresponding with public LMG strains were retained. This rule was imposed by the Bacteria collection for not publishing non-public strain information. As we were, however, confronted with species for which very little data is available, a bypass operation was allowed in which non-public LMG strains were enclosed as research strains.
- A threshold of three FAME profiles per species was set. This in view of learning with a training, validation and test set. To enable validation of model parameters and testing of the prediction performance of the final identification model, at least one profile was included in each subset.
- Logically, empty FAME profiles and profiles comprising only one peak with a 100% relative peak percentage were removed. These aberrant profiles mainly result from gas-chromatographic problems or incorrect sample preparation.
- Removal of FAME profiles with less than three FAME peaks. This rule was imposed as such a profiles have a high probability of being erroneous. As stated in Subsection 2.3.1.1, a sound FAME profile consists of 5 to 15 FAME peaks.
- Subspecies were not further considered and, thus, were enclosed within the corresponding species. Intrasubspecific annotations were only considered in the genus *Pseudomonas* where the subdivision proposed by Gardan et al. (1999) was followed.
- A visual inspection of the profiles was finally required in order to remove outliers. Outliers are regarded as FAME profiles that are qualitatively and quantitatively distinct from the majority of FAME profiles of the same species or strains. Logically, these outliers can only be detected within species and strains corresponding with a large number of FAME profiles.

It is clear that this procedure corresponds to a lot of manual and tedious work. As an example, it approximately took 1.5 days for generating the *Bacillus* data set. Because of this manual creation and inspection, the presence of minor errors remained possible. This was mainly the case for profiles generated before 2001. Nonetheless, this whole sampling process can be automated but will still need a good logging system and ultimate visual inspection for detecting any sampling errors or FAME profile outliers.

This data selection procedure was mainly executed four times during this work. First, a *Bacillus* FAME data set was created according to the validly published taxonomy of November 2006. Second, *Bacillus* and *Pseudomonas* data sets were created according to the validly published taxonomy of October 2007. These data sets were created for presentation of the machine learning work at the BioMicroWorld congress in Seville in November 2007. As the identification results of this sampling approximated the results of the first and third sampling, these data sets are not further considered in this work. The third sampling resulted in three FAME data sets, according to the validly published taxonomy of March 2008. These data sets were also transformed into other data set types. The three genera were merged into a genera data set. FAME profiles annotated by genus and species name as well as annotations only by the genus name were considered. From the *Bacillus* data set, a small 15 species data set was created.

And, from the *Pseudomonas* data sets, two data sets were created regarding plant pathogenesis (more information below). Fourth, a final update of the third data selection was performed according to the validly published taxonomy of May 2008. In this latter case, a full genus-wide scope was aimed. Due to delays in delivery of the bacterial strains by several culture collections (world-wide) and different authors, this work is still ongoing. For each of these data sets, several hundreds of additional FAME profiles were generated by the Laboratory of Microbiology so that at least three FAME profiles were available for each species. An overview of some data set statistics is given in Table 3.1. Tables A.1 and A.2 report the included strains of the 2006 and 2008 data sets, respectively, together with the total number of corresponding FAME profiles. This dissertation mainly handles research performed by the data sets generated in March 2008 and this is also the case for the data analysis reported in this chapter.

Year	Data Set	# Species	# Strains	# Profiles	# FAME peaks
2006	<i>Bacillus</i>	82	477	1,077	79
2007	<i>Bacillus</i>	76	502	1,045	87
	<i>Pseudomonas</i>	89	566	1,466	97
2008	<i>Bacillus</i>	74	436	961	71
	<i>Paenibacillus</i>	44	189	378	46
	<i>Pseudomonas</i>	95	667	1,673	94
	<i>Genera</i>	213	1,292	3,012	105

Table 3.1: Statistics of the generated data sets. For each data set year, type of data set, and number of corresponding species, strains, FAME profiles and FAME peaks are reported.

From Table 3.1 some interesting facts can be deduced. For instance, from 2006 to 2008 the number of validly described *Bacillus* species included in the FAME database decreased with seven species, as a result from species renamed within the genus or to species of another genus. This is nicely visualized in Figure 2.3. From 2006 to the end of 2008, ten species were renamed outside the genus *Bacillus* and four *Bacillus subtilis* group species were enclosed in another species of the group: *B. axarquiensis* and *B. malacitensis* were enclosed in the species *B. mojavensis* (Wang et al., 2007b), while *B. velezensis* was re-evaluated as *B. amyloliquefaciens* (Wang et al., 2008). Besides this, a remarkable decrease in the number of *Bacillus* strains is noticed. This is due to the species *B. circulans*, for which the strains were restricted toward the so-called *B. circulans sensu stricto*. As a consequence, 68 of the 72 strains were removed from the data set. Many questions and a lot of confusion still exist about the taxonomic position of these strains. It is clear that this species needs a thorough further revision.

In the case of the genus *Pseudomonas*, a lot of microbial research focuses on the many plant-pathogenic strains present in the genus. Different species comprise pathogenic strains such as *P. syringae*, *P. savastanoi* and *P. marginalis*. Especially the species *P. syringae* is widely studied because of the large number of plant-pathogenic strains. An important study is performed by Gardan et al. (1999) who analyzed 48 *P. syringae* pathovars and 8 related species by DNA-DNA hybridization. In the according study, nine genomospecies are proposed. A short overview of these genomospecies is given. Genomospecies 1 corresponds to *P. syringae sensu stricto* and includes the *P. syringae* pathovars *syringae*, *aptata*, *lapsa*, *papulans*, *pisi*, *atrofaciens*, *aceris*,

panici, *dysoxyl*i and *japonica*. Genomospecies 2 includes the *P. syringae* pathovars *phaseoli*-*cola*, *ulmi*, *mori*, *lachrymans*, *sesami*, *tabaci*, *morsprunorum*, *glycinea*, *ciccaronei*, *eriobotryae*, *mellea*, *aesculi*, *hibisci*, *myricae*, *photinae* and *dendropanacis* and *Pseudomonas savastanoi*, *Pseudomonas ficuserectae*, *Pseudomonas meliae* and *Pseudomonas amygdali*. This genomospecies is given the name of *P. amygdali*, as this is the earliest valid name. Genomospecies 3 includes the *P. syringae* pathovars *tomato*, *persicae*, *antirrhini*, *maculicola*, *viburni*, *berberidis*, *apii*, *delphinii*, *passiflorae*, *philadelphii*, *ribicola* and *primulae*. *P. syringae* pv. *tomato* serves as the type strain. Genomospecies 4 includes "*Pseudomonas coronafaciens*" (not validly described yet) and *P. syringae* pathovars *porri*, *garcae*, *striaefaciens*, *atropurpurea*, *oryzae* and *zizaniae* and corresponds to "*P. coronafaciens*". Genomospecies 5 includes *P. syringae* pv. *tremae* and is named *Pseudomonas tremae*. Genomospecies 6 includes *Pseudomonas viridiflava* and the strains of *P. syringae* pv. *ribicola* and *P. syringae* pv. *primulae* and is named *P. viridiflava*. Genomospecies 7 includes *P. syringae* pv. *tagetis* and *P. syringae* pv. *helianthi*. *P. syringae* pv. *tagetis* serves as reference. Genomospecies 8 includes *P. syringae* pv. *theae* and *Pseudomonas avellanae* and thus corresponds to *P. avellanae*. Genomospecies 9 includes *P. syringae* pv. *cannabina* and corresponds to *Pseudomonas cannabina*. As DNA-DNA hybridization is the reference for species delineation, this genomospecies (sub)division is followed in this study. Hence, for this case of plant-pathogenic *Pseudomonas* species, two separate FAME data sets were created: a plant-pathogenic *Pseudomonas* data set covering 25 species and the complete *Pseudomonas* data set with the species labeled as being plant-pathogenic or not. Again, a *Pseudomonas* species was considered to be plant-pathogenic when at least one of its strains is known as a pathogen of either plants or mushrooms. An overview of the considered plant-pathogenic species is given in Table 3.2, together with the different host(s) and corresponding references. These species are also denoted by superscript 'p' in Table A.2. Remark that *P. fluorescens* is not included in the list of species, while some biovars of this species are pathogenic to plants (Gardan et al., 2002).

Remark that some species are generically misnamed and are or need to be transferred to another genus. Examples are *P. beteli* and *P. hibiscicola* that are heterotypic synonyms of the species *Stenotrophomonas maltophilia* (Van Den Mooter and Swings, 1990), and *P. geniculata* that should be transferred to the same genus (Anzai et al., 2000). *P. pictorum* is a close relative of the genus *Stenotrophomonas*. The same problem is true for the species *P. cissicola* that should be transferred to the genus *Xanthomonas*, with *P. boreopolis* being a close relative (Hu et al., 1997; Anzai et al., 2000; Palleroni, 2005) A similar anomaly holds for *P. flectens* that based on 16S rRNA gene sequence analysis clusters in the family of the *Enterobacteriaceae* (Anzai et al., 2000; Palleroni, 2005). In this study, we further denote these species as *P. beteli* group, except for *P. flectens*. However, for most species no valid renaming has been done yet (Euzéby, 1997).

The scope of the data sets created in this study is, logically, dependent on the number and type of research projects performed at the Laboratory of Microbiology during the last 20 years, and of the service demands directed to the BCCMTM/LMG Bacteria Collection by companies, research institutes, etc. Thus, not all bacterial species are of equal importance and interest in both cases, ultimately resulting in data sets covering species with a different number of

Species	# FAME profiles	Host(s)	Reference(s)
<i>P. agarici</i>	8	Mushroom, <i>Agaricus bisporus</i>	(Höfte et al., 2007)
<i>P. amygdali</i>	111	Almond, <i>Prunus amygdalus</i>	(Höfte et al., 2007)
<i>P. asplenii</i>	13	Bird's-nest fern, <i>Asplenium nidus</i>	(Gardan et al., 2002)
<i>P. avellanae</i>	4	Hazelnut, <i>Corylus avellana</i>	(Janse et al., 1996)
<i>P. beteli</i>	6	Betel, <i>Piper betle</i>	(Van Den Mooter and Swings, 1990)
<i>P. cannabina</i>	7	Hemp, <i>Cannabis sativa</i>	(Gardan et al., 1999)
<i>P. caricapapayae</i>	5	Papaya, <i>Carica papaya</i>	(Höfte et al., 2007)
<i>P. cichorii</i>	32	Wide host range	(Smith et al., 1988; Höfte et al., 2007)
<i>P. cissicola</i>	8	<i>Cissus japonica</i>	(Hu et al., 1997)
<i>P. coronafaciens</i>	44	Oat, <i>Avena sativa</i>	(Gardan et al., 1999)
<i>P. corrugata</i>	22	Tomato, <i>Lycopersicon</i> ; also <i>Chrysanthemum</i> , <i>Geranium</i> , <i>Medicago</i> , pepper	(Catara et al., 2002; Gardan et al., 2002; Höfte et al., 2007)
<i>P. constantinii</i>	6	Mushroom, <i>Agaricus bisporus</i>	(Munsch et al., 2002; Höfte et al., 2007)
<i>P. flavescens</i>	7	Walnut, <i>Juglans regia</i>	(Hildebrand et al., 1994)
<i>P. flectens</i>	6	<i>Phaseolus vulgaris</i>	(Gardan et al., 2002)
<i>P. fuscovaginae</i>	45	Rice, <i>Oryza sativa</i> ; also <i>Allium</i> , <i>Secalotriticum</i> , <i>Triticum</i> , grass	(Miyajima et al., 1983)
<i>P. hibiscicola</i>	6	<i>Hibiscus rosa-sinensis</i>	(Van Den Mooter and Swings, 1990)
<i>P. marginalis</i>	63	<i>Pastinaca sativa</i> ; also <i>Allium</i> , <i>Cichorium</i> , <i>Medicago</i> , <i>Phaseolus</i> , <i>Phragmipedium</i>	(Gardan et al., 2002; Höfte et al., 2007)
<i>P. mediterranea</i>	10	Tomato, <i>Lycopersicon</i> ; also pepper	(Catara et al., 2002; Höfte et al., 2007)
<i>P. salomonii</i>	10	Garlic, <i>Allium sativum</i>	(Gardan et al., 2002)
<i>P. syringae</i>	88	Wide host range	(Gardan et al., 2002)
<i>P. syringae</i> genomospecies 3	38	Wide host range	(Gardan et al., 1999)
<i>P. syringae</i> genomospecies 7	8	<i>Helianthus annuus</i> , <i>Tagetes erecta</i>	(Gardan et al., 1999)
<i>P. tolaasii</i>	53	Mushroom, <i>Agaricus bisporus</i> ; also <i>Agaricus bitorquis</i> , <i>Allium sativum</i> , <i>Pleurotus ostreatus</i> , <i>Pleurotus eryngii</i>	(Munsch et al., 2002; Höfte et al., 2007)
<i>P. tremae</i>	8	<i>Trema orientalis</i>	(Gardan et al., 1999, 2002)
<i>P. viridiflava</i>	17	Wide host range	(Höfte et al., 2007)

Table 3.2: Overview of the considered plant and mushroom pathogenic *Pseudomonas* species. For each species, the number of FAME profiles, the host(s) and corresponding reference(s) are given.

corresponding FAME profiles. Moreover, not all species have an equal number of strains and not all strains are of equal importance. This further led to an additional imbalancing of the data sets. This can clearly be noticed in the respective strain tables A.1 and A.2. The imbalanced nature of the data sets created in March 2008 is illustrated in Figures 3.1 and 3.2. In view of data analysis and machine learning research, it is important to keep this imbalanced nature in mind during statistical analysis. Also important to consider, especially in view of identification by machine learning models, is the very large number of species together with the very small number of FAME profiles per species. Depending on the separability of the species by FAME data, the accuracy of species prediction could be seriously confined.

From these figures, it is immediately clear that the majority of classes only contained a small number of FAME profiles. For each of the three genera, one can also notice which species were and still are of main research, service interest and importance to both the Laboratory of Microbiology and the BCCM™/LMG Bacteria Collection. Core examples for the genus *Bacillus* are *B. cereus* (food-poisoning and opportunistic human/animal pathogen) and *B. subtilis* group species (type species; food, clinical, veterinary and environmental importance); for the genus *Paenibacillus*: *Pa. larvae* (insect pathogen), *Pa. polymyxa* (type species; plant rhizosphere-associated) and *Pa. thiaminolyticus* (large number of strains); and for the genus *Pseudomonas*: *P. aeruginosa* (type species; opportunistic human and plant pathogen), *P. amygdali* (plant pathogen), *P. fluorescens* (food spoilage and plant pathogen), *P. marginalis* (plant pathogen), *P. putida* (common soil species, bioremediation) and *P. syringae* (plant pathogen).

3.3 Data Analysis and Visualization

In this section, data analysis of the 2008 data sets is described. Different methods were used and a genus-wide analysis was pursued. From literature, the latest genus-wide studies concerning FAME analysis in the genera *Bacillus*, *Paenibacillus* and *Pseudomonas* date from the study of Kämpfer (1994), Heyndrickx et al. (1996) and Vancanneyt et al. (1996), respectively (see also Subsections 2.3.2.1, 2.3.2.2 and 2.3.2.3). It is clear that an updated revision of the genera was needed, which is described in the following subsections. Analysis focused on the peak level (average and peak distribution), on the species level (distance calculation and clustering) and on dimensionality reduction (PCA).

3.3.1 Average FAME Profile

First, it is always intriguing to study how feature values vary over the different classes, which in this study correspond to bacterial species. Especially in the framework of the present study, this type of analysis should be preferred over a global analysis of the complete data set, i.e. without regarding the species labels of the profiles. As the sampled data sets were imbalanced, this type of data analysis would result in biased conclusions. Thus, average FAME profiles were calculated for each species and the average peak values of the major fatty acids and standard deviations are reported in Appendix A.2. For the genera *Bacillus* and *Paenibacillus*, only peaks with a prevalence in more than ten species are considered, while in the genus

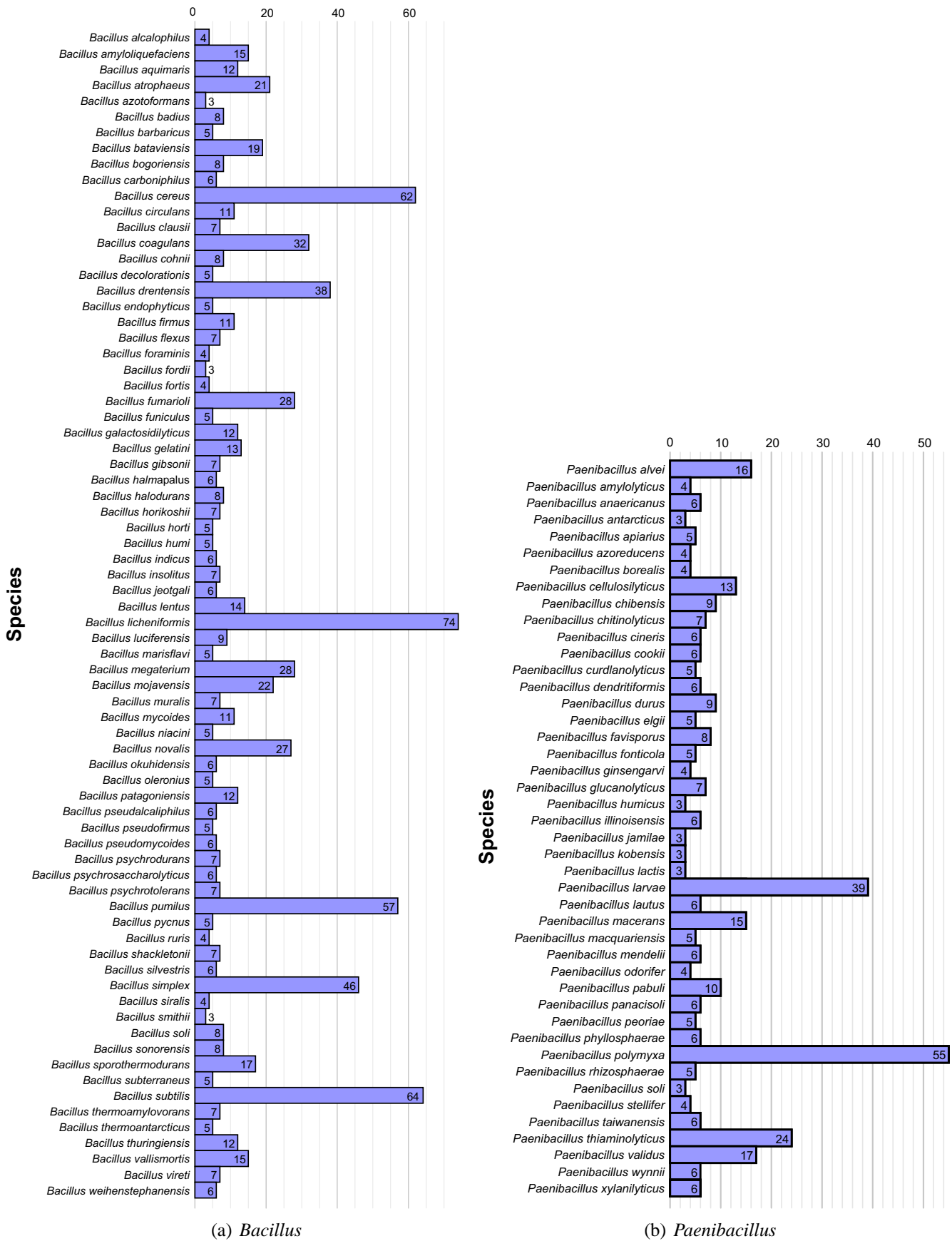


Figure 3.1: Number FAME profiles per *Bacillus* and *Paenibacillus* species.

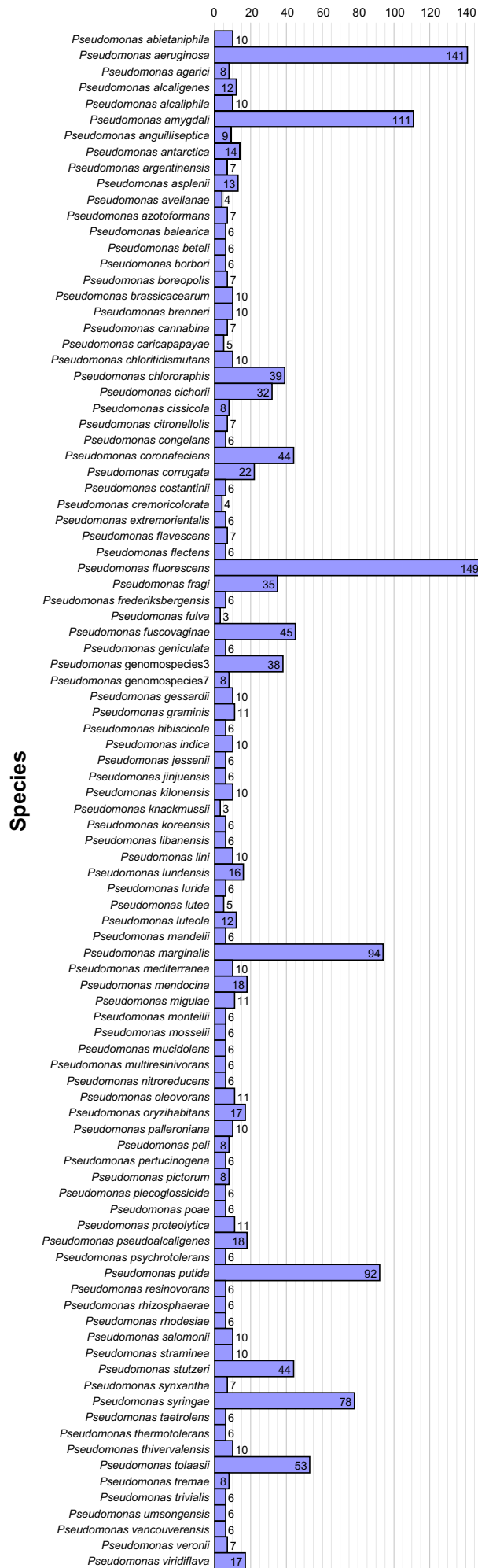


Figure 3.2: Number FAME profiles per *Pseudomonas* species.

Pseudomonas the prevalence cut-off is set at twenty species (due to a higher number of included species and peaks). These average FAME profiles were further analyzed by looking at the peak averages and standard deviations, and at prevalences over the averaged data set.

71 FAME peaks were found in the *Bacillus* data set, comprising 74 species. The major fatty acids of the average profiles are reported in Table A.3. Major fatty acids in the genus *Bacillus* were C_{15:0} anteiso and C_{15:0} iso. Smaller amounts of C_{14:0}, C_{14:0} iso, C_{16:0}, C_{16:0} iso, C_{16:1} w11c, C_{17:0} anteiso, C_{17:0} iso and iso C_{17:1} w10c were found. These findings corresponds to the study of Kämpfer (1994) and the description of Logan and De Vos (2009) (see also 2.3.2.1). The standard deviations were small to moderate, except for *B. azotoformans* for which only three profiles were included, originating from two strains (LMG 9581^T and LMG 15443) and with two profiles corresponding to the type strain. Distinct differences were seen in the fatty acid profiles of both strains. The distribution of the peak values of these average FAME profiles as averaged over all species is visualized in Figure 3.3. It is clear that two C_{15:0} fatty acids were predominant with quite large standard deviations. This was also true for the eleven fatty acids that corresponded to smaller peak values. Thus, a large variability was present in the major peaks of the data set, being advantageous for qualitative and quantitative discrimination purposes. A peak prevalence distribution over the average profiles of all species is illustrated in Figure 3.4. From this plot, it could be concluded that seven peaks were present in each *Bacillus* species and could, thus, be regarded as core-*Bacillus* peaks. A lot of peaks only occurred in one to ten species, and could be regarded as species- or even strain-specific.

In the genus *Paenibacillus* 44 species covered 46 FAME peaks. Table A.4 reports the main average FAME profile of each species. The predominant fatty acid was C_{15:0} anteiso. Smaller peak areas were detected for C_{14:0}, C_{14:0} iso, C_{15:0} iso, C_{16:0}, C_{16:0} iso, C_{16:1} w11c, C_{17:0} anteiso and C_{17:0} iso. This corresponds to the conclusions of Heyndrickx et al. (1996) and the description in Bergeys Manual of Systematic Bacteriology (Priest, 2009). Standard deviations denoted also quite stable FAME profiles. Differences were seen in the six FAME profiles of *Pa. cineris*. These profiles originated from two strains (LMG 18439^T and LMG 21976) showing distinct differences in almost all major fatty acids. The distribution of the peak values of these average profiles as averaged over all species is visualized in Figure 3.5. The standard deviation of the predominant peak C_{15:0} anteiso showed that this fatty acid is a quite stable core-*Paenibacillus* peak. As a result, species discrimination based on this fatty acid will be confined. Eight fatty acids corresponded to smaller peak amounts and showed quite large standard deviations, meaning that discrimination is possible in both a qualitative and quantitative manner. Also a peak prevalence distribution was plotted for the average profiles of all species. Figure 3.6 shows that eight more core-*Paenibacillus* peaks were present. Moreover, also in this data set a lot of peaks had a prevalence of only one species, allowing qualitative discrimination. In between, some fatty acids occurred in about half of the species but in small amounts. An example is FAME iso C_{17:0} w10c. In this case, quantitative as well as qualitative discrimination becomes possible.

The 95 *Pseudomonas* species covered 94 FAME peaks. The main average profile of each species is reported in Table A.5. Predominant fatty acids were C_{16:0}, C_{18:1} w7c and summed feature 3. Smaller peak areas were found for the FAMEs C_{10:0} 3OH, C_{12:0}, C_{12:0} 2OH and C_{12:0} 3OH. Note the difference with the two other data sets where hydroxy fatty acids (related to

the LPS layer of Gram-negatives) are not present as major fatty acids, in contrast to this data set. These findings correspond to the study of Vancanneyt et al. (1996) and the description in Bergey's Manual of Systematic Bacteriology (Palleroni, 2005). As in the *Bacillus* data set, standard deviations were also high. Interestingly, distinct average fatty acid profiles were seen for the species *P. beteli*, *P. boreopolis*, *P. cissicola*, *P. flectens*, *P. geniculata*, *P. hibiscicola* and *P. pictorum*. As stated above these species were actually generically misnamed within the genus *Pseudomonas sensu stricto*. But, from this FAME study, it was also clear that this group of species is an outgroup within this genus.

The distribution of the relative peak area values of the average FAME profiles as averaged over all species is visualized in Figure 3.7. The standard deviations of the major fatty acids were quite large, allowing again for a quantitative as well as a qualitative discrimination. The majority of peaks had a very low average, implying the possibility of only a qualitative discrimination. Nonetheless, from the distribution plot of peak prevalence in the average FAME profiles of all species (see Figure 3.8), half of the number of peaks was present in more than ten species. Interestingly, only four fatty acids occurred in all average profiles, though twelve fatty acids occurred in more than 80 of the 95 species (84%). Thus, this data set with a larger number of species, strains and profiles related to only a small number of FAME profiles with moderate to high average peak values. This makes the presence of FAME peaks with small peak area percentages, or thus qualitative discrimination, more important for prediction purposes, when compared to the *Bacillus* and *Paenibacillus* data sets.

3.3.2 Clustering

Next to the calculation of an average FAME profile for each species present in the data set, it is also interesting to look at how the data initially group or cluster together. For easy visualization, a heatmap of the data set was created. Initially, only peaks were clustered. In this case, the different data set instances were alphabetically ordered by species name. Subsequently, also clustering was performed at the species level. Clustering was performed with the statistical software R and the algorithm `heatmap.2` function of the `gplots` package. In calculating distances between fatty acid profiles, it is important to consider that minor fatty acids can have a large discriminatory power. Therefore, a good alternative to the commonly used Euclidean distance and to the Mahalanobis distance, as used by the Sherlock MIS system, is the Canberra metric that is defined as

$$d_{\text{canb}}(x, y) = \frac{1}{n} \sum_{i=1}^n \frac{|x_i - y_i|}{|x_i| + |y_i|}. \quad (3.1)$$

The Canberra metric calculates the sum of a series of fractions between pairs of data points. Thus, this metric does not only take into account the distance between two points but also their distance to the origin (Dawyndt, 2004). This distance is very sensitive to small changes when the two considered points are close to the origin. In other words, minor fatty acids will have a larger contribution to the distance, when for instance compared to the Euclidean distance. Note, however, that the covariances between the fatty acids are not considered in this distance and that

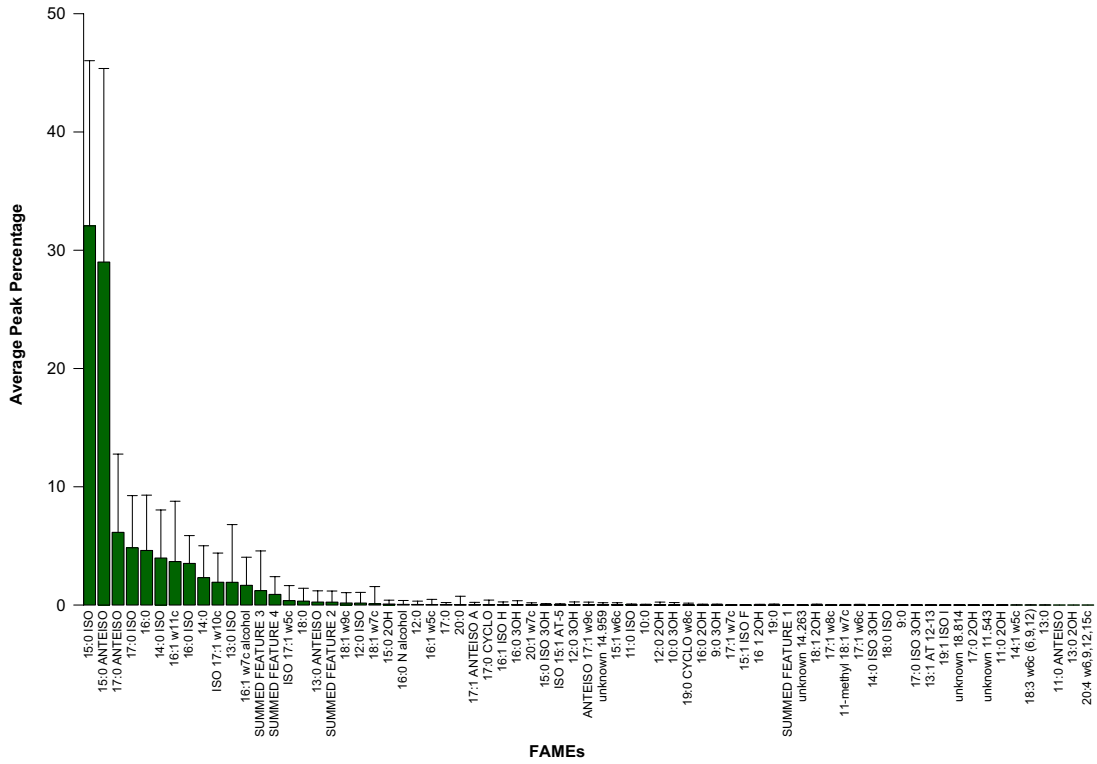


Figure 3.3: Average of peak percentages in the genus *Bacillus*

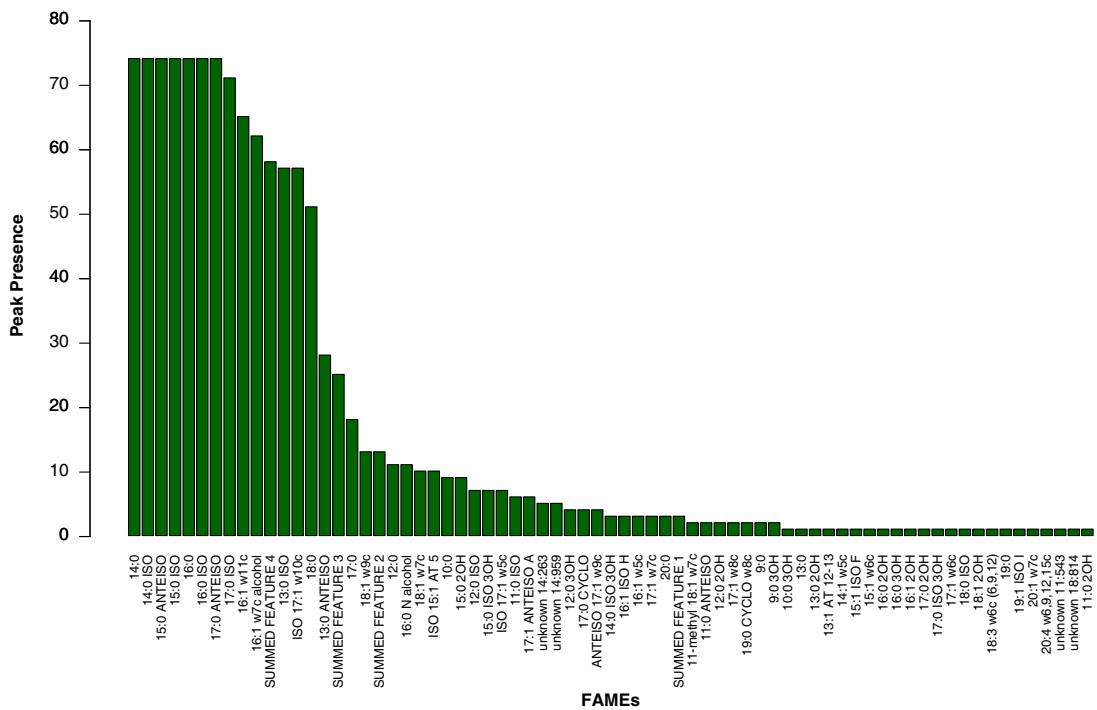


Figure 3.4: Peak distribution in average species profiles the genus *Bacillus*

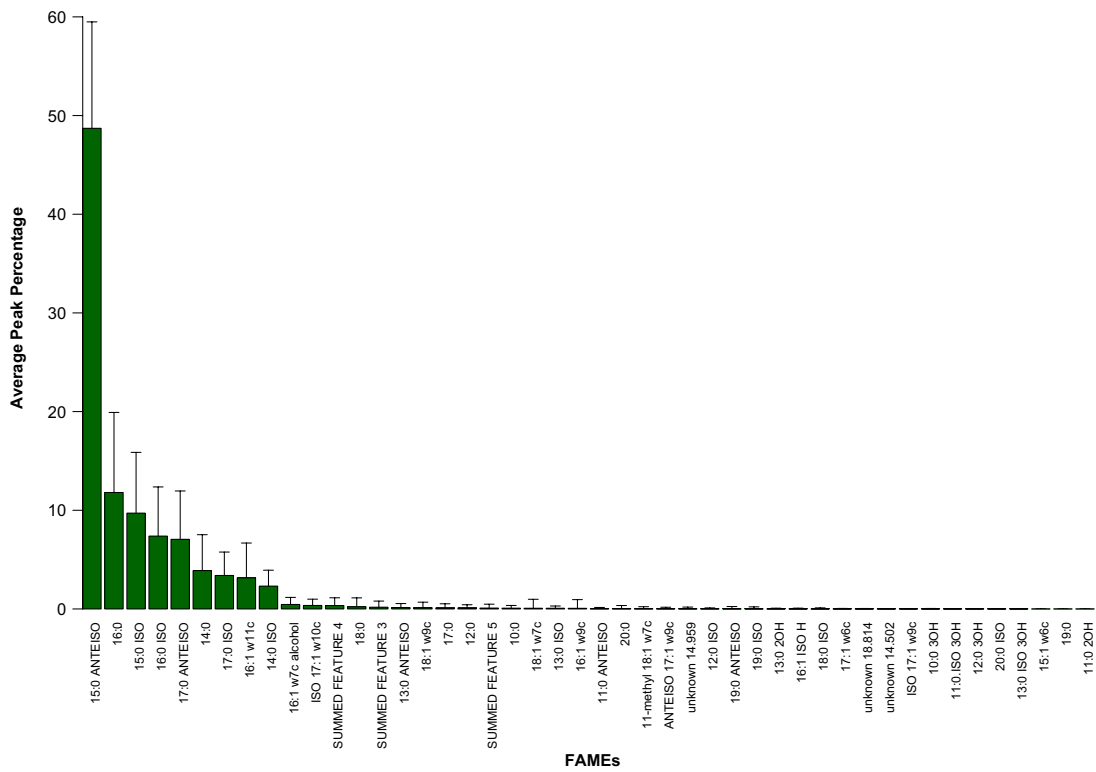


Figure 3.5: Average of peak percentages in the genus *Paenibacillus*

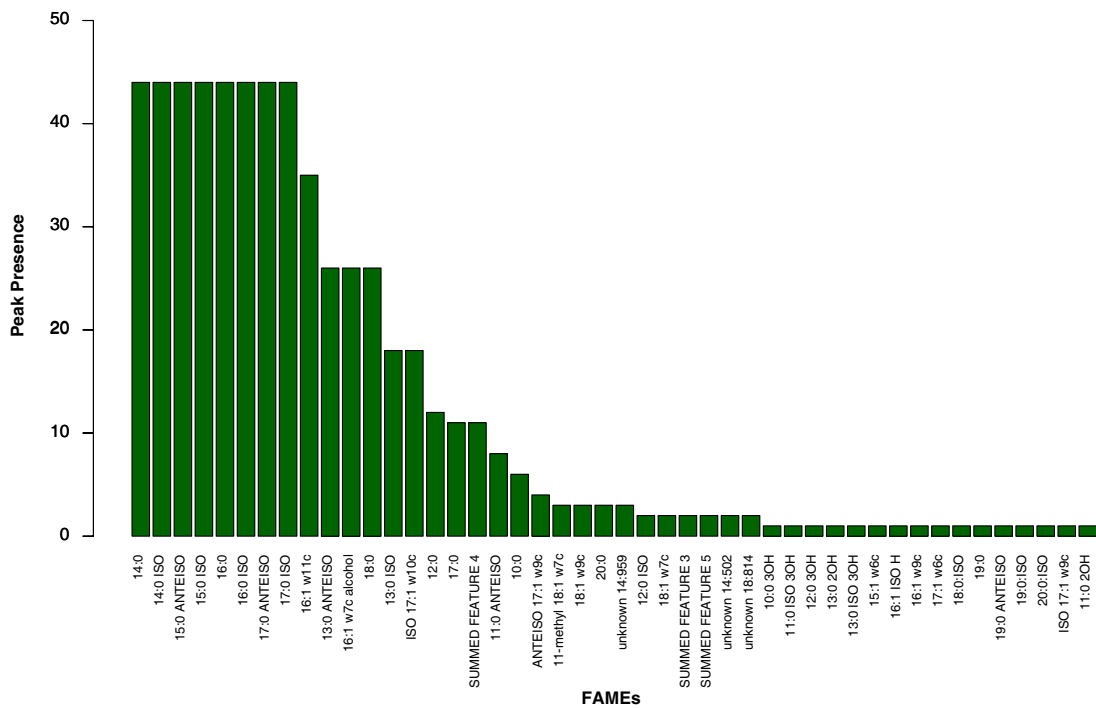


Figure 3.6: Peak distribution in average species profiles the genus *Paenibacillus*

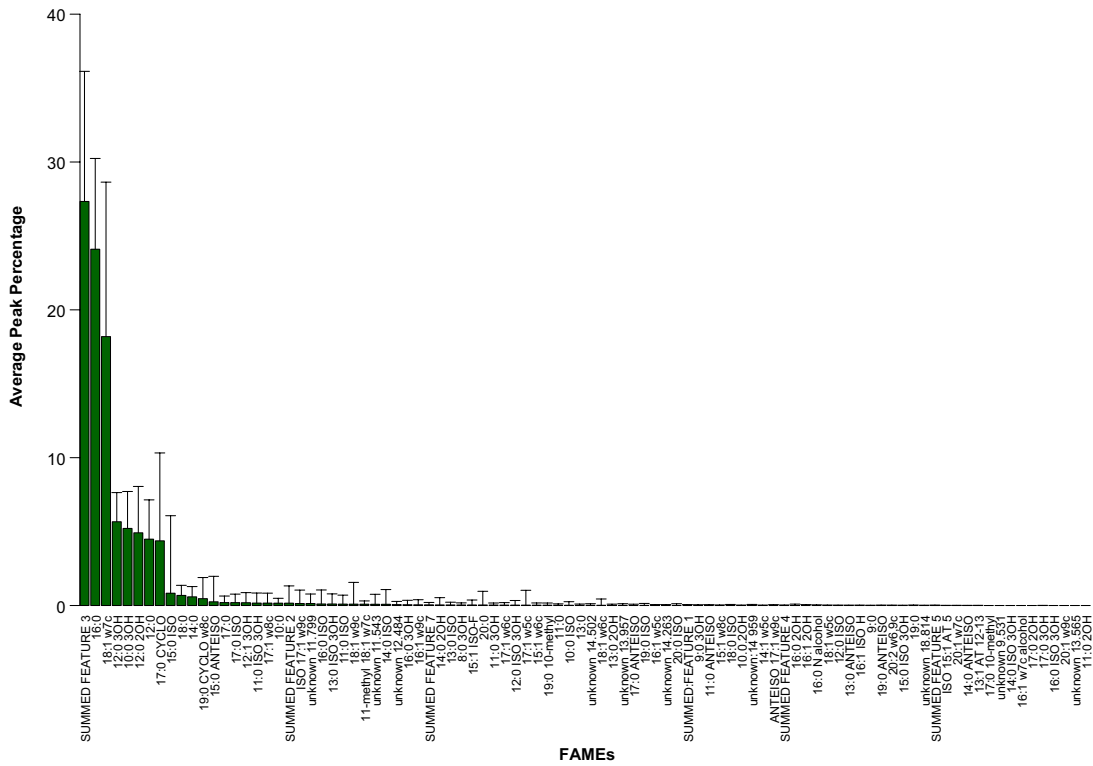


Figure 3.7: Average of peak percentages in the genus *Pseudomonas*

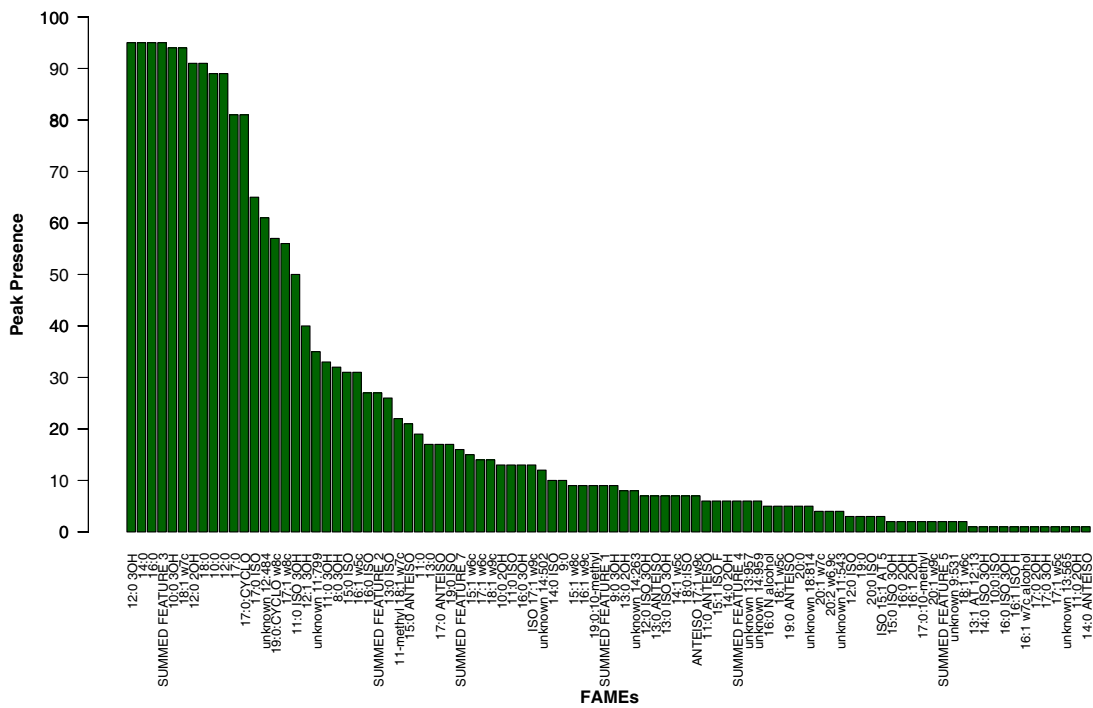


Figure 3.8: Peak distribution in average species profiles of the genus *Pseudomonas*

all FAMEs are regarded as independent features. The two clustered data heatmaps are visualized for the genera *Bacillus*, *Paenibacillus* and *Pseudomonas* in Figure 3.9– 3.11, respectively. The [0,1]-interval of FAME values is visualized by a green-black-red colour range. Also, at the left of each heatmap, each species is coded by a different colour. In the peak-clustered heatmaps, FAME profiles are alphabetically ordered according to their corresponding species name. A dendrogram is drawn for each hierarchical clustering, using the average distance.

Similar to the findings in the previous section, also from these figures it is immediately clear that the FAME profiles of the species of the genera *Bacillus*, *Paenibacillus* and *Pseudomonas* comprised a small number of major fatty acids, while the FAME scope of the different data sets is pretty large. From the peak-clustered heatmaps, different FAME groups could be distinguished. Now, when also clustering of the data instances was considered, it became clear that the FAME profiles of different species showed to be quite similar and that different species were represented by several FAME groups. These similarities and FAME subgrouping of the different species clearly makes the identification task based on similarity calculation a hard job, but an interesting challenge when considering machine learning. In the case of *Bacillus*, certain species clearly cluster together and some distinct species clusters could be distinguished. The same, but better, was also true for the *Paenibacillus* data set. It could also be noticed that in the case of *Pseudomonas* this subgrouping was enormous, making the prediction task very hard. Note however, that a different number of species was present in each data set with a different number of strains and profiles. However, this is not necessarily a constraint for good data separation.

3.3.3 TaxonGap

3.3.3.1 A Visualization Tool for Intra- and Inter-Species Variation among Individual Biomarkers

3.3.3.1.1 Introduction

Selection of optimal biomarkers for the identification of different operational taxonomic units (OTUs, i.e. leaves of a phylogenetic tree) may be a hard and tedious task. This is especially the case when phylogenetic trees for multiple biomarkers, typically genes, need to be compared. When evaluating candidate biomarkers for the identification of different OTUs, one intuitively is looking for molecular markers that at the same time show the least amount of heterogeneity within OTUs and result in a maximal separation between the different OTUs. The first requirement must guarantee that members of the same OTU have the same (or at least similar) biomarkers, so that they can easily be grouped together based on those markers. The second requirement must guarantee that members of different OTUs have sufficiently different biomarkers, so that an identification based on those markers cannot erroneously suggest assignment of the members to the same OTU. TaxonGap was especially designed to produce a compact graphical representation of the resolution of individual biomarkers within and between taxonomic units, allowing easy and reliable inspection of the data for evaluation across different OTUs and different biomarkers.

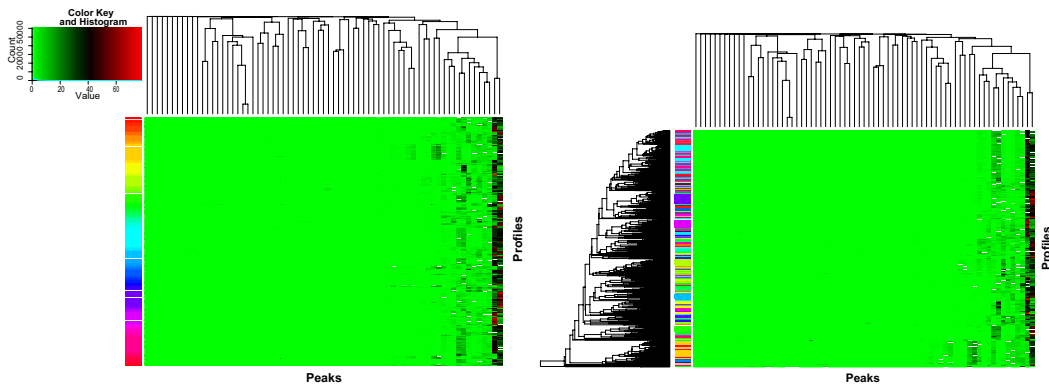


Figure 3.9: Heatmap of data with clustering of the genus *Bacillus* data set. The left figure illustrates only peak clustering, while the right plot illustrates peak and profile clustering. Clustering is based on the Canberra metric. Rows correspond to the different FAME profiles, which are alphabetically ordered and coloured by species name (presented at the left of each heatmap). Columns correspond to the different FAME peaks. Peak percentages are coloured from green to red with an increasing value, as represented by the colour key in the top-left corner.

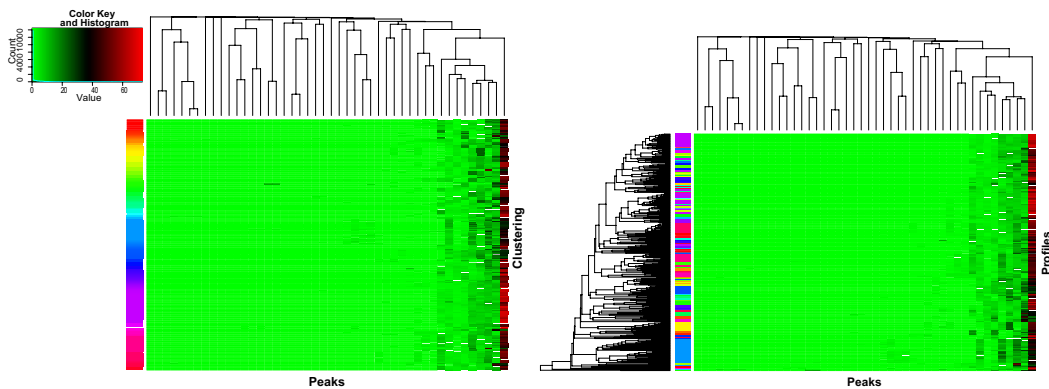


Figure 3.10: Heatmap of data with clustering of the genus *Paenibacillus* data set. The left figure illustrates only peak clustering, while the right plot illustrates peak and profile clustering. Clustering is based on the Canberra metric. Rows correspond to the different FAME profiles, which are alphabetically ordered and coloured by species name (represented in the colour bar at the left of each heatmap). Columns correspond to the different FAME peaks. Peak percentages are coloured from green to red with an increasing value, as represented by the colour key in the top-left corner.

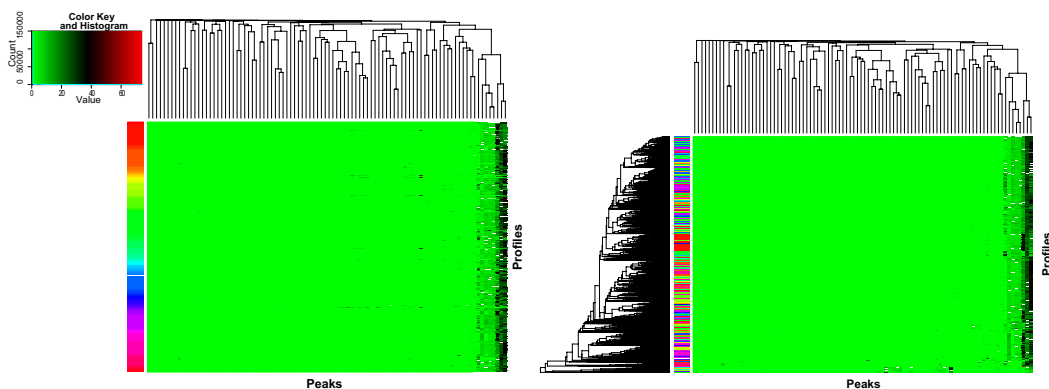


Figure 3.11: Heatmap of data with clustering of the genus *Pseudomonas* data set. The left figure illustrates only peak clustering, while the right plot illustrates peak and profile clustering. Clustering is based on the Canberra metric. Rows correspond to the different FAME profiles, which are alphabetically ordered and coloured by species name (represented in the colour bar at the left of each heatmap). Columns correspond to the different FAME peaks. Peak percentages are coloured from green to red with an increasing value, as represented by the colour key in the top-left corner.

3.3.3.1.2 Algorithm and Graphical Representation

For a given set of OTUs $\{O_1, O_2, \dots, O_n\}$, the s -heterogeneity within taxon O_i is defined by

$$\max_{x,y \in O_i, x \neq y} d_s(x, y). \quad (3.2)$$

Herein, $d_s(x, y)$ represents the distance between the (different) members x and y of the taxon O_i as measured from the biomarker s . These distances are presented in a separate distance matrix for each biomarker. Likewise, the s -separability of taxon O_i is defined by

$$\min_{x \in O_i, y \notin O_i} d_s(x, y). \quad (3.3)$$

The taxon containing that member y for which a minimum distance is reached in the computation of the s -separability, is called the closest neighbour of taxon O_i . Note, however, that the closest neighbour relationship is not necessarily symmetric: the fact that O_j is the closest neighbour of O_i does not imply that O_i is also the closest neighbour of O_j . TaxonGap calculates the matrix of s -heterogeneity and s -separability values with the different OTUs as matrix rows and the different biomarkers as matrix columns. Headers are placed to the left and on top of the matrix. To improve interpretability of the resulting graphical representation, the OTUs are presented according to their position in a phylogenetic tree, as an alternative to listing them in alphabetic order. With the aim to improve visual inspection and interpretation of the data and to support optimal comparability of the values across the biomarkers, TaxonGap presents the s -heterogeneity and the s -separability values respectively as light gray and dark gray horizontal bars for the individual biomarkers. The name of the closest neighbour is attached to the right side of the dark gray bar. Light gray bars are printed on top of and are made less thick than dark gray bars. Although not a strict requirement, it is advised that the same OTUs are used for evaluation of the different biomarkers. Missing biomarker data for a given OTU leads to holes in the TaxonGap output matrix. Note also that there is no necessity to use the same OTU members for evaluating different biomarkers. Importantly, the application of TaxonGap is not restricted to the comparison of genetic or molecular markers. In fact, TaxonGap accepts pairwise distance matrices generated from any kind of biomarkers. All biomarkers that enable the calculation of pairwise distance matrices may be compared using TaxonGap.

The graphical representation produced by TaxonGap offers a number of advantages over comparing individual phylogenetic trees for the evaluation of different biomarkers in identification studies. First of all, a separate row is reserved in the TaxonGap output for heterogeneity and separability values of different biomarkers for a single taxon, which is not the case when comparing phylogenetic trees. Even after a tedious process of swapping branches, it is not always possible to draw phylogenetic trees in a way that enables clear visual comparison. This is especially true when phylogenetic trees for multiple genes need to be compared. In addition, TaxonGap uses the same scaling for depicting distance values based on individual biomarkers. Few software tools for drawing phylogenetic trees allow precise control over scaling. Both placement and scaling improve comparability of the heterogeneity and separability for individual taxa. Secondly, it is important to point out the fact that phylogenetic trees present approxima-

tions of the underlying distance values instead of using minimum and maximum as aggregation operators. This is important when comparing s -heterogeneity and s -separability values for all species of a given biomarker s . To underscore the overall success rate of individual biomarkers to discriminate between the OTUs, TaxonGap depicts the overall separability (dark gray) per OTU as a vertical line for each biomarker. This line is omitted when the overall separability is too small. Finally, the graphical output of TaxonGap remains compact, even for data sets where the number of OTU members or biomarkers grows large. This is because the software has a built-in aggregation based on the individual OTUs and biomarkers. TaxonGap thus allows for a more straightforward evaluation of the discriminatory power of individual biomarkers in an OTU identification scheme, as opposed to the need of comparing separate trees drawn for each of the OTUs in the scheme.

3.3.3.1.3 Software and Publication

TaxonGap is implemented as an executable JAVA archive (JAR) with a graphical user interface. The graphical output is formatted as an enhanced postscript (EPS) document. Sequence alignments, pairwise distance or similarity matrices and phylogenetic trees can be generated using third-party software packages. To make TaxonGap even more user-friendly, the command line interface allows to use TaxonGap as a back-end plugin into this graphical software. TaxonGap 2.4.1, together with all previous builds, is currently freely available for download at the KERMIT website <http://www.kermit.ugent.be/TaxonGap>.

This work is published as an Applications Note in the international peer-reviewed journal *Bioinformatics* with reference B. Slabbinck, P. Dawyndt, M. Martens, P. De Vos and B. De Baets (2008). TaxonGap: a visualisation tool for intra- and inter-species variation among individual biomarkers. *Bioinformatics*, 24(6), 866-867.

3.3.3.2 FAME as a Taxonomic Marker

Hence, the intra- and inter-species variation of the three genera was also analyzed by the TaxonGap software, with FAME considered as biomarker. Using the software package BioNumerics (Applied Maths, Sint-Martens-Latem), a clustering was performed of the different FAME data sets with the Canberra distance metric and the resulting distance matrix was exported for analysis with the TaxonGap tool. For graphical representation, we did not opt for a FAME tree but rather chose for integrating the 16S rRNA gene sequence tree, covering all included species of the respective genus. The main reason is that the 16S rRNA gene allows to discriminate between most bacterial species and because it is interesting to investigate how the resolution of FAME analysis for species discrimination relates to the results of 16S rRNA gene sequence analysis. For sequence analysis, one high-quality 16S rRNA sequence was manually selected from the SILVA database (Pruesse et al., 2007). From the exported aligned sequences, a maximum likelihood tree was calculated using 1000 bootstraps by the RAxML software (Stamatakis, 2006). As this software package allows distributed computing with the message passing interface (MPI), tree inference was performed in parallel on an Intel Blade cluster (Intel Corporation, Santa Clara, CA, USA). The resulting tree, formatted in the Newick Tree Format,

was adjusted for analysis with the TaxonGap software tool. For input-output scaling and formatting between the different software programs, several Perl scripts and JAVA programs were developed.

A TaxonGap visualization was generated for each data set of the three genera, see Figures 3.12–3.14, respectively. For the considered genera, it is immediately clear that FAME is not a good taxonomic biomarker when considering the species rank as OTU. For almost all species, the intra-species variability (light-gray bars, heterogeneity) was much larger than the inter-species variability (dark-gray bars, separability), while the opposite is needed for a good taxonomic biomarker. This result was a clear indication of highly similar species, given the considered data. The clustering experiment confirmed this result as high similarities were seen between the FAME profiles of the different species. For many species, it could be seen that the separability values were quite small. Low separability values were especially seen for closely related species such as the species of the *Bacillus cereus* and *Bacillus subtilis* group. In these cases, the closest neighbour was almost always another species of the corresponding group, except for *B. cereus* which closest neighbour was *B. lentus* (possibly due to a misannotation). For the *B. cereus* group this was expected as the different species are mainly described by phenotypic and pathogenic traits. In case of *Pseudomonas*, other examples are the species with plant-pathogenic strains such as *P. syringae*, *P. amygdali*, *P. genomospecies 3*, *P. genomospecies 7*, *P. avellanae*, etc. for which the closest neighbour was also one of the mentioned species. The same trend was seen in the *P. beteli* group of species for which the closest neighbours were also species from this group. Note here, that from 16S rRNA sequence analysis this group was clearly an outgroup in the tree, assuming that the according species do not relate to *Pseudomonas sensu stricto*. From the TaxonGap visualization of the genus *Pseudomonas*, it could also be concluded that a lot of *Pseudomonas* species were very related in their FAME profiles, as shown by the many low separability values. This could be assigned to the small number of major fatty acid peaks present in the average FAME profiles and to the majority of low peak values (see also figures above). This was again confirmed by the clustering analysis.

3.3.4 Principal Component Analysis

When dealing with datasets with an excessive dimensionality, 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 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). Subsequent linear combinations are composed that are orthogonal to the previous PCs, repeatedly based on the combination representing the highest variance. A valuable visualization of the PCA analysis is achieved by plotting the variance and accumulated variance of the top- x PCs in a so-called scree plot. 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 or dimensions, resulting in a less complex model. This is not



Figure 3.12: Intra- and inter-species discrimination by FAME in the genus *Bacillus*. Visualization generated by TaxonGap 2.4.1. The axis denotes the percentage of heterogeneity within a species and the percentage of separability from other species. Dark grey bars denote the minimum separability from the other species, while the light grey bars denote the maximum species heterogeneity. The arrow indicates a line corresponding to the minimum separability over all species.

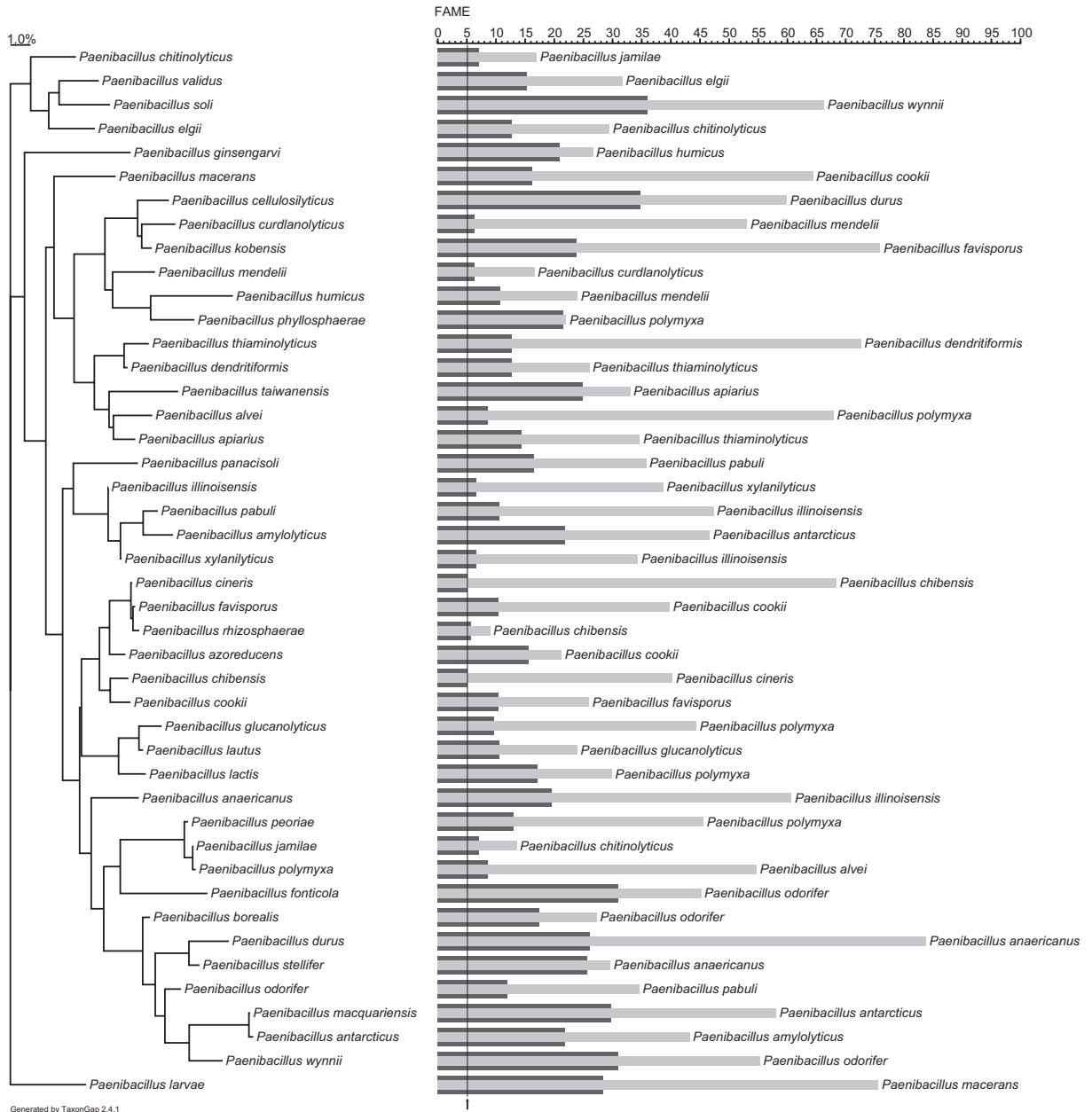


Figure 3.13: Intra- and inter-species discrimination by FAME in the genus *Paenibacillus*. Visualization generated by TaxonGap 2.4.1. The axis denotes the percentage of heterogeneity within a species and of the percentage separability from other species. Dark grey bars denote the minimum separability from the other species, while the light grey bars denote the maximum species heterogeneity. The arrow indicates a line corresponding to the minimum separability over all species.

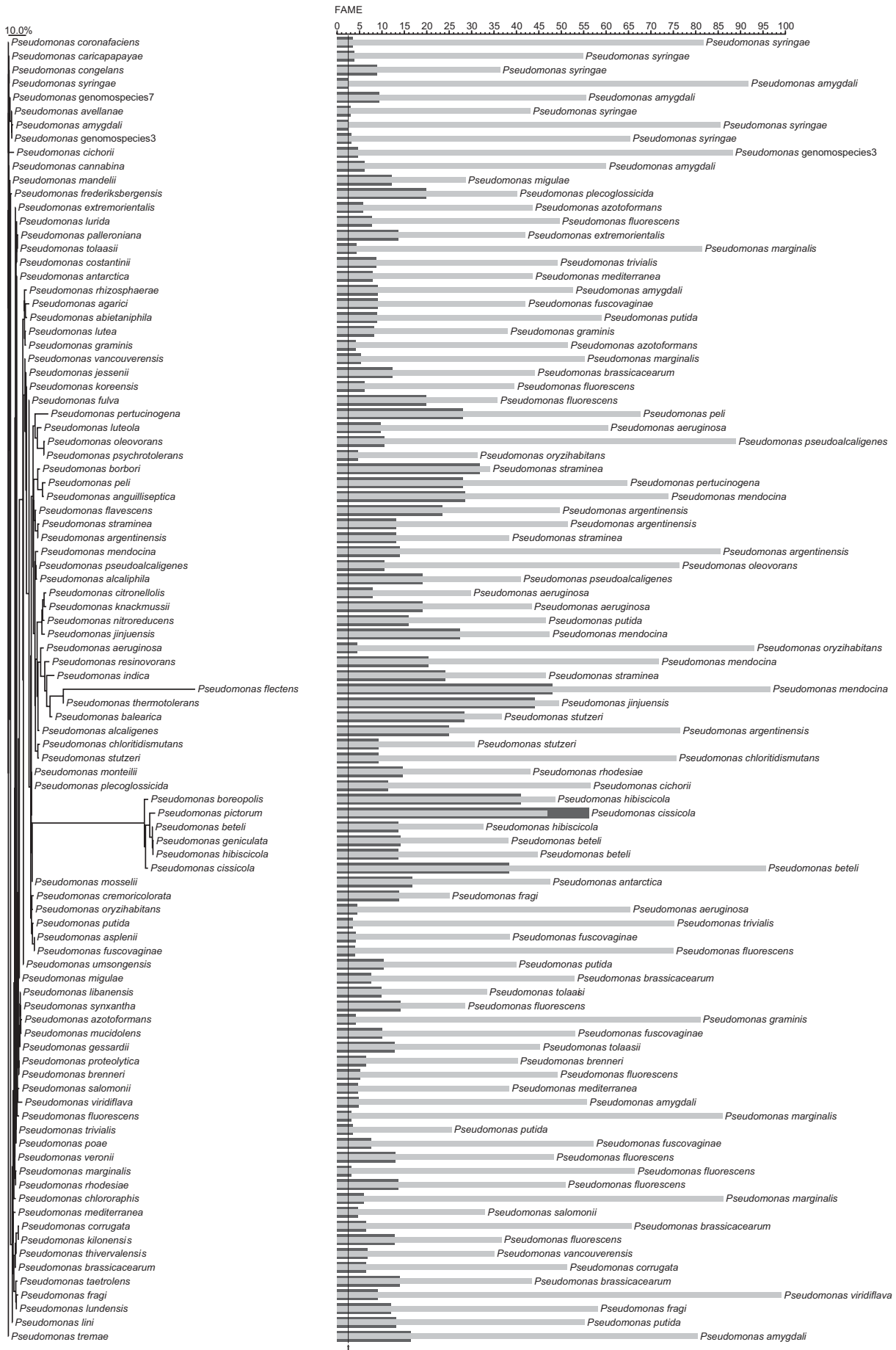


Figure 3.14: Intra- and inter-species discrimination by FAME in the genus *Pseudomonas*. Visualization generated by TaxonGap 2.4.1. The axis denotes the percentage of heterogeneity within a species and of the percentage separability from other species. Dark grey bars denote the minimum separability from the other species, while the light grey bars denote the maximum species heterogeneity. The arrow indicates a line corresponding to the minimum separability over all species.

only advantageous for computational reasons but may also be very effective for increasing the performance of an identification model.

A PCA analysis was performed for the three genus data sets and visualizations were subsequently done by skree plots (see Figure 3.15). It was immediately clear that for the three genera with only about five PCs approximately 95% of the variability in the FAME data could be represented. This implied that the different features are highly correlated, which can mainly be attributed to the fatty acid biosynthesis pathway where fatty acids are typically converted into other fatty acid molecules and to the activity of certain enzymes (e.g. desaturase enzyme) (Madigan et al., 2009). Knowledge of highly correlated features is important in view of constructing identification models with machine learning techniques as this implies that certain patterns will be present in the data set and that certain features will become redundant and non-informative.

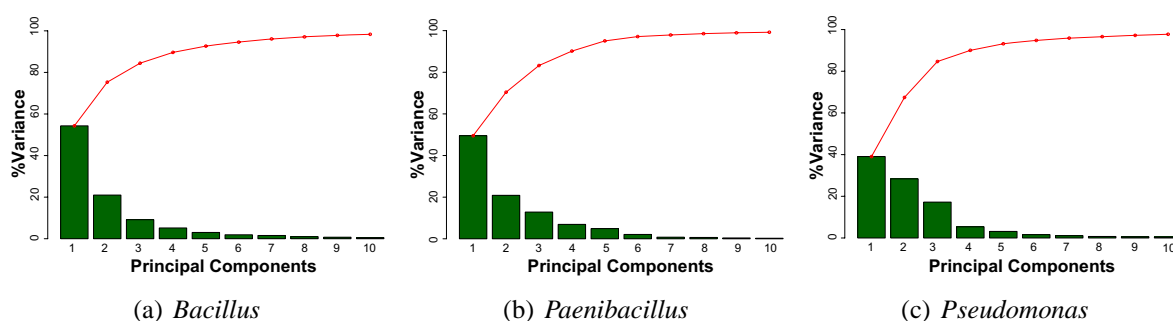


Figure 3.15: Principal component analysis of the three genera *Bacillus*, *Paenibacillus* and *Pseudomonas*. Skree plots are given for the data set of each genus, with only the ten first principal components.

Following PCA analysis, it is also possible to draw biplots of the different PCs. Biplots were generated only for the first two components of each data set and are visualized in Figures A.1–A.3, respectively. The different species are annotated by different colours. To improve the visibility of the large number of species, the most important species and species distinct in the plot are additionally annotated by a particular shape or mark.

For the genus *Bacillus*, the *cereus* group species are annotated by an ‘×’ mark, while the *subtilis* group species are annotated by a circle. The *cereus* group species could clearly be separated from the other species, though species discrimination in the group will be very hard. This latter remark also holds for the *subtilis* group species. In the biplot, other distinct species could be found such as *B. galactosidilyticus*, *B. decolorationis*, *B. ruris*, *B. niacini* and *B. pycnus*. In the biplot of the second and third PC (figures not shown), the *subtilis* group species became more separated. And, in the biplot of the first and third PC (figures not shown), the *B. cereus* moved into the global data cloud. Interestingly, in this biplot, temperature-related species were more separated such as *B. fumarioli*, *B. gelatini* and *B. thermantarcticus* (all grown at 52°C). Also, *B. coagulans* and *smithii* became distant from the data cloud and from these species it is known that their optimal growth temperature lies in the interval 40–57°C and 25–60°C, respectively (Logan and De Vos, 2009). Temperature seems to play a certain role in this biplot.

For the genus *Paenibacillus* only one data cloud was visible with some distinct species. Examples are *Pa. anaericanus*, *Pa. stelifer*, *Pa. ginsengarvi* and *Pa. humicus*. Species with a

large amount of FAME profiles are annotated by a specific shape, such as *Pa. polymyxa* ('×') and *Pa. larvae* (circle). In the biplot of the second and third PC as well as in the biplot of the first and third PC, the data cloud became more compact (figures not shown).

For the genus *Pseudomonas* three distinct data clusters could be seen: a cluster with *P. aeruginosa* (red '×'), a cluster with the plant-pathogenic species (circles) and a cluster with *P. beteli* ('+'). In this biplot, the data was much more compact than in the corresponding biplot of the other genera. In the biplot of the first and the third PC (figures not shown), the *P. beteli* group species became more separated from the other species, and the same three clusters were observed. Also in the biplot of the second and the third PC (figures not shown), the *P. beteli* group was more separated, though the plant-pathogenic cluster and the *P. aeruginosa* cluster overlapped.

It was also interesting to see how the three genera were located with respect to each other. A biplot of the first two PCs is visualized in Figure 3.16. These first two PCs represented about 82% of the variance in the data (figure not shown). From the biplot, it could be concluded that the three genera showed to be clearly separated from each other. Logically, the genera *Bacillus* and *Paenibacillus* were more closely related with each other than with the genus *Pseudomonas*. Nonetheless, some *Pseudomonas* appear in the *Bacillus* cluster. A closer look into the data showed that this could be attributed to the *P. beteli* group species and *P. flectens*. This again supports research, such as the work of Anzai et al. (2000), stating that the corresponding species should actually not be integrated to the genus *Pseudomonas*.

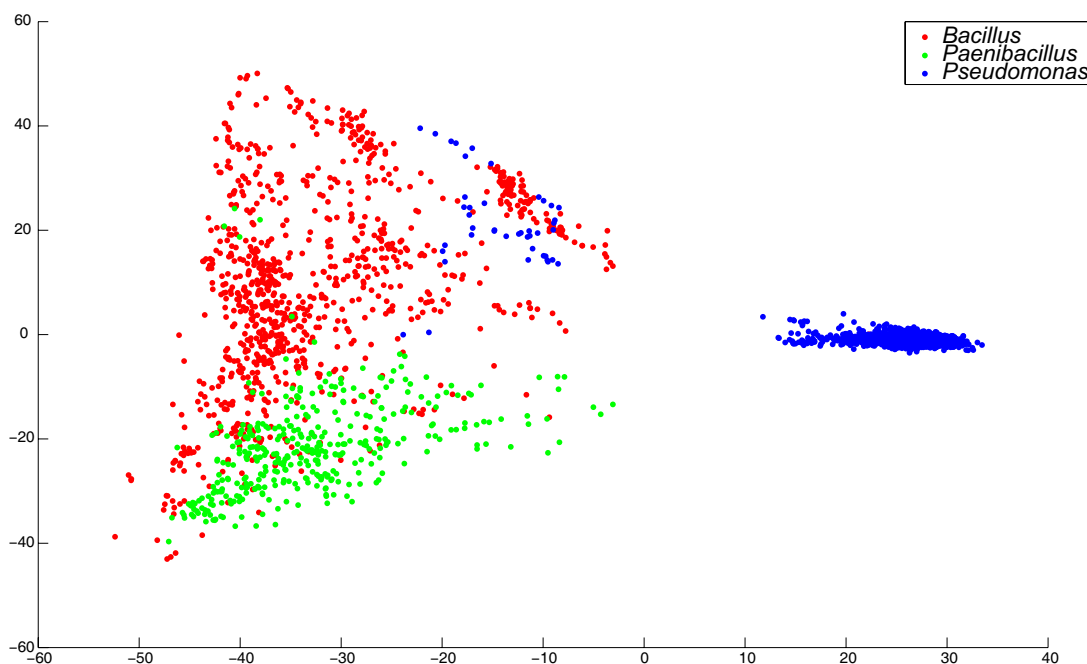


Figure 3.16: Principal component analysis of the genus data set. A biplot is visualized of the first two principal components.

Finally, we also investigated how FAME data could be used to discriminate between plant-pathogenic *Pseudomonas* species, and between this group and non-plant-pathogenic *Pseudomonas* species. A biplot of the first two principal components is given for both data sets, see Figures 3.17 and 3.18. From the first biplot, corresponding to the plant-pathogenic data set, it is

clear that plant-pathogenic species were hard to distinguish from each other. In this biplot, the species belonging to the *P. beteli* group and the species *P. flectens* are not included. A better scaling of the other species was obtained without these species. When compared to non-plant-pathogenic species, the second biplot showed that the plant-pathogenic FAME data clustered in one data cloud, and the non-plant-pathogenic data clustered into two FAME clouds, with one cloud clearly overlapping with the plant-pathogenic FAME cloud.

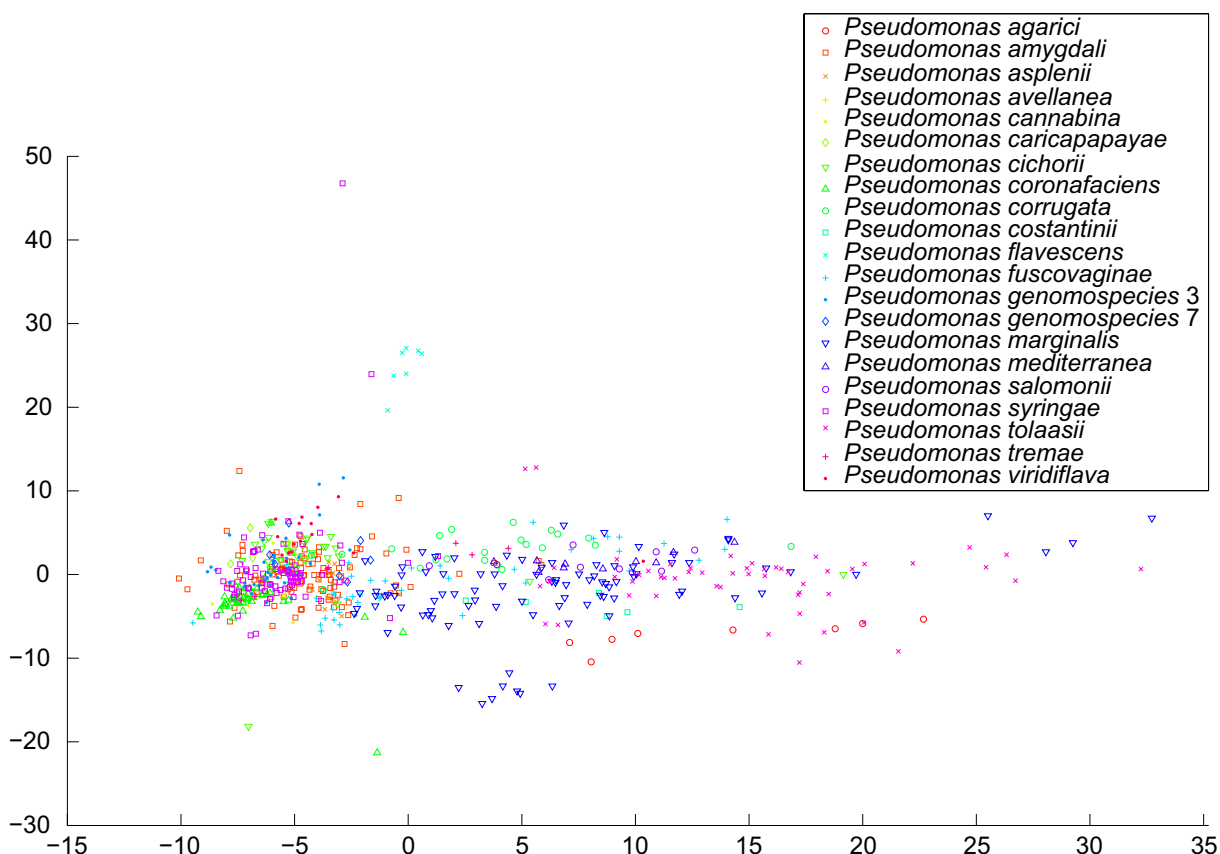


Figure 3.17: Principal components analysis of the plant-pathogenic *Pseudomonas* data set. A biplot is visualized of the first two principal components. The species belonging to the *P. beteli* group and *P. flectens* are not included in the plot.

3.4 Conclusion

In this section, we have discussed a FAME data analysis of the three genera *Bacillus*, *Paenibacillus* and *Pseudomonas*. Notice that in this study, we deal with a very specific problem setting. In view of classification by machine learning techniques, a lot of classes are present and these are represented by a small number of data instances or FAME profiles. On average, a small number of peaks was seen per species, even though a lot more peaks were present in the complete data set. Core-genus peaks, that occurred in all species of the genus, were present together with species- and strain-specific peaks. When averaged over all species of the genus, peak values and their standard deviations showed that species discrimination can be done in a quantitative and/or qualitative manner. Clustering and TaxonGap analyses showed that, when

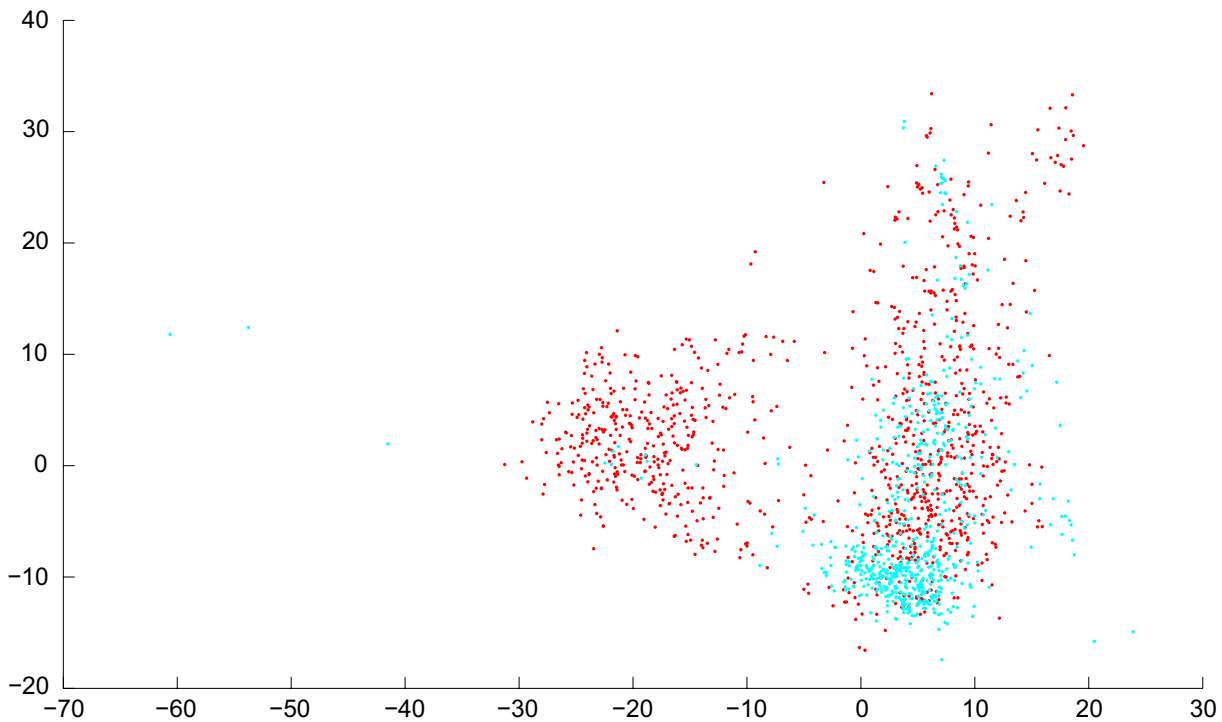


Figure 3.18: Principal components analysis of the 2008 *Pseudomonas* data set with species labeled as being plant-pathogenic or not. A biplot is visualized of the first two principal components. Red points correspond to non-plant-pathogenic species, while green points correspond to plant-pathogenic species.

calculating similarities or distances from the FAME data, a bad discrimination and identification will be attained as FAME profiles of several species are closely related. Also, it was clearly visible that species show different FAME subgroups and that FAME data will mostly allow to only discriminate between species groups. Moreover, from PCA analysis, it was shown that FAME peaks are indeed correlated. Therefore, redundant information was present in the data sets. Furthermore, based on the PCA analysis of the genus data set, the three genera could be well separated from each other. From the different genus-specific PC biplots, it could be concluded that the *Pseudomonas* species were more related in FAME data than the *Bacillus* and *Paenibacillus* species. In the latter case, also more distinct species and species groups were present. The conclusions regarding the *Pseudomonas* species were supported by the study of the plant-pathogenic *Pseudomonas* species, that showed very related FAME patterns between the different species. With all this knowledge, it could be underscored that identifying bacterial species by similarity scores will probably not have a large power. Nevertheless, it is worth investigating how machine learning techniques will be able to generalize over the different species and if optimal margins/boundaries can be calculated between these species. The high variability within the species will be an obstacle in calculating reliable margins or boundaries, but the peak correlations will allow a learning process and model construction with only a subset of informative features.

CHAPTER 4

FAME-based Bacterial Species Classification

Learning without thought is labor lost.

Thought without learning is perilous

CONFUCIUS

Die vier wil èn, moe de rook kun' verdrag'n

A BRUGES SAYING

4.1 Introduction

The growing list of validly published bacterial species clearly indicates that the bacterial landscape is continuously evolving. On 03/11/2009, 7,995 bacterial species were validly described (Euzéby, 1997). Given this rapid change in taxonomy, back-end identification libraries of first-line identification methods need constant updates. As gas chromatographic whole-cell FAME analysis is cheap, easy to handle and automated, it is used by many laboratories as a first-line identification method for bacterial species. Routine use at the Laboratory of Microbiology (Ghent University, Belgium) and the BCCMTM/LMG Bacteria Collection (Ghent, Belgium) has led to a joint FAME database, currently containing more than 71,000 bacterial FAME profiles. This database lends itself to keep track of the most recent changes in taxonomy of those taxa that are vastly represented in the database and for machine learning purposes. FAME analysis for bacterial identification relies, however, on the commercial Sherlock Microbial Identification System (MIS, MIDI Inc., Newark, Delaware, USA) for which the back-end identification libraries are only updated every few years and only cover part of all known species. The accuracy of bacterial species identification is therefore highly compromised, making this update latency a major drawback of the Sherlock MIS system. Even though microbial taxonomy is changing rapidly, the development of up-to-date identification libraries can be realized by computer systems and back-end databases.

Before 2006, the use of machine learning techniques for the classification of bacterial species based on FAME profiles was restricted to a very small taxonomic scope. The first FAME-based classification of bacteria was oriented towards seven genera of several marine bacteria by means of artificial neural networks. The scope of this research was rather small as the genera were represented by only 36 strains (Ruggiero et al., 1993). Extension of this

research was done by classifying 4, 5 and 14 genera of marine and environmental bacteria respectively, covered by 35, 26 and 39 species, and 71, 50 and 45 strains (Bertone et al., 1996; Giacomini et al., 2000, 2004). In all cases, research was also directed to a restrictive number of parameters concerning architecture and training of neural networks. Interestingly, however, the researchers concluded that, as FAME data yields information at the species level, it would be worthwhile to build a FAME-based bacterial species identification system. Until 2006, no large-scale FAME-based and genus-wide bacterial species classification and identification was established on the basis of machine learning techniques.

In the following two sections, we demonstrate the potential of classification and identification of species within the present genera *Bacillus*, *Paenibacillus* and *Pseudomonas* using different machine learning techniques. In the first section, only the genus *Bacillus* is considered together with artificial neural networks (ANNs). Several data sets were built according to different experimental setups considering various validation strategies and parameter settings. Comparison of the identification results ultimately led to some promising setups towards genus-wide *Bacillus* species identification. In the second section, we explore the realization of an extended and up-to-date FAME-based bacterial species identification system powered by machine learning. The genera *Bacillus*, *Paenibacillus* and *Pseudomonas* were considered together with three machine learning techniques: ANNs, support vector machines (SVMs) and random forests (RFs). Based on a laboratory information management system and the FAME database, we analyzed the identification at genus and species level. Analyses were evaluated both from a computational and a microbiological perspective. Furthermore, the identification results were subjected to an in-depth comparison with those obtained by the commercial Sherlock MIS.

4.2 *Bacillus* species classification: an ANN Approach

4.2.1 Methodologies

4.2.1.1 Artificial Neural Networks

In this section, we demonstrate the potential of classification and identification of species within the genus *Bacillus* using supervised ANNs. The *Bacillus* species as validly described in October 2006 were considered. More information concerning the data set can be found in Table 3.1. Feed-forward ANNs with backpropagation were trained with the resilient propagation learning algorithm. The ANN architecture consisted of one hidden layer and two activation functions were considered: the sigmoid and bipolar sigmoid activation function, further abbreviated as ‘s’ and ‘b’. An ANN with activation function f_1 on the hidden output neurons and activation function f_2 on the output neurons is further denoted as ‘ANN f_1/f_2 ’. Stratified simple validation (abbreviated as ‘val’) and cross-validation (abbreviated as ‘cv’) were used for parameter optimization by early stopping. An example is given in Figure 4.1 for the optimization of the number of hidden neurons, using the mean squared error as error function. Following trial-and-error experiments, different optimization intervals were finally chosen. The interval [10,200] was used for optimizing ANN b/b and ANN s/b models, the interval [10,250]

for optimizing ANN b/s models and the interval [10,350] for optimizing ANN s/s models. The number of hidden neurons in the last two intervals is increased due to the sigmoid activation function, as set on the output neurons, which corresponds to a smaller scope of output values (see also Subsection 1.2.1.3.2). As we dealt with a multi-class setting, identification of test data was done with the winner-take-all rule in which each data instance was labelled with the label of the output neuron corresponding with the highest output value.

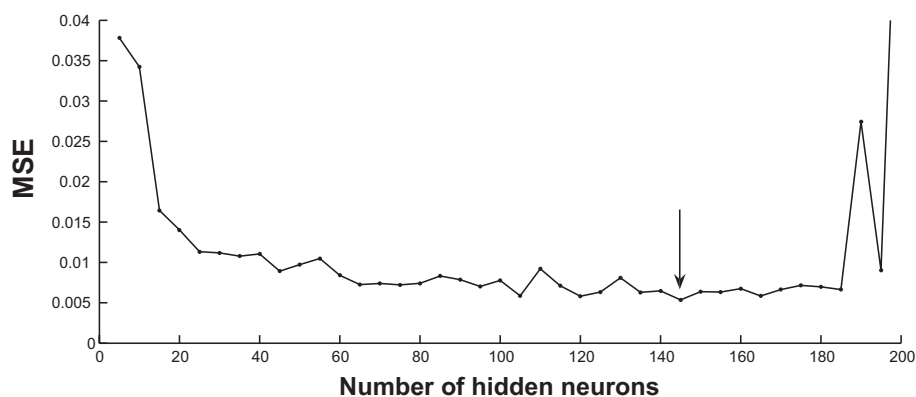


Figure 4.1: An example of optimization of the number of hidden neurons. Optimization was done by stratified cross-validation. Mean squared error (MSE) values are plotted for different numbers of hidden neurons (step size of 5). The final number of hidden neurons is pointed by the arrow.

4.2.1.2 Balanced and Imbalanced Data Sets

The sampled *Bacillus* data set contained FAME profiles with a different number of profiles for each species. Two possible directions could be considered for data set creation. Hence, the main research question in this section elucidates whether all data in the data set can be used or whether an equal number of profiles per species should be sampled. The former case is better known as data sets with an imbalanced class distribution. When data is highly costly, as in our case, this type of data set should be preferred. The latter case is better known as under-sampling the data set to deal with class imbalances. These two types of data sets are further denoted as imbalanced and balanced data sets. Different types of imbalanced data sets have already been analyzed, such as highly imbalanced two-class data sets (Japkowicz and Stephen, 2002) and imbalanced multi-class data sets (Weiss and Provost, 2003). As in a genus-wide identification scheme bacterial species are regarded equally important, our identification setup differed from those described.

Balanced data sets were created by randomly sampling three FAME profiles per species. From each balanced and imbalanced data set, a test set was created by randomly sampling one-third of the profiles of each species while using the remaining data for training. Cross-validation was performed with two folds. In the case of simple validation, a validation set was created by randomly sampling 50% of the training profiles of each species. Ultimately, these setups resulted in four experiment types. Each experimental setup was repeated ten times, each time with randomly sampled profiles. According to the four possible activation function combinations, four ANNs were trained for each experimental setup. This finally resulted in 160 ANN experiments. For high-performance computing these experiments were performed on a

Blade cluster (Intel Corporation, Santa Clara, CA, USA). Identification itself could easily be done on any modern PC.

4.2.1.3 Statistical Significance

For each experiment type, which was repeated ten-fold, the mean AUC and the standard deviation were calculated. Significant difference in performance was tested by a Wilcoxon rank-sum test and defined by a p -value below the significance level 0.05. This test was chosen because the normality assumption of the underlying distributions could not be guaranteed. Consequently, the Wilcoxon rank-sum test was a good alternative to the t -test (see also Subsection 1.3.3). All assumptions were met, except for the shapes of the underlying distributions which were assumed identical.

4.2.2 Results and Discussion

4.2.2.1 ANN Performance

To evaluate ANN-based *Bacillus* species classification, an initial data set was composed of 1,077 whole-cell FAME profiles, originating from standard growth conditions and covering 82 species represented by a heterogeneity of 477 strains. Different training, validation and test sets were randomly created, with varying balance type and validation type. An overview of ANN performance for each experiment type is given in Table 4.1 for the activation combination leading to the best results. The mean overall AUC of each experiment type is plotted in Figure 4.2 for each activation combination. Note that the AUC was calculated for each class (species) separately and an average was calculated over these classes (overall AUC), and subsequently, an average was taken over the different repeats (mean overall AUC). Also, a slightly different approach was used for the calculation of the ROC curve, as described in Subsection 1.3.2. Instead of using the output score of each profile as a threshold, here 10 thresholds were set in the interval [0,1] by steps of 0.1. The four highest mean overall AUC values were obtained for the imbalanced and cross-validated experiment type. The best result was obtained for the experiment type with the bipolar sigmoid activation function on both neuron types. The corresponding mean overall AUC value was 0.914 and the mean TP% over all test profiles was 75.2%.

Experiment type	Activation	# Training profiles	# Validation profiles	# Test profiles	Mean AUC (stdev)
bal3val	s/b	82	82	82	0.829 (0.041)
bal3cv	b/b	164	0	82	0.881 (0.023)
imbal3val	s/b	384	362	331	0.882 (0.031)
imbal3cv	b/b	746	0	331	0.914 (0.010)

Table 4.1: Overview of the identification results of each experiment type with the activation combination leading to the highest mean AUC. Number of training, validation and test profiles, and the mean and standard deviation of the area under the ROC curve (AUC) are reported.

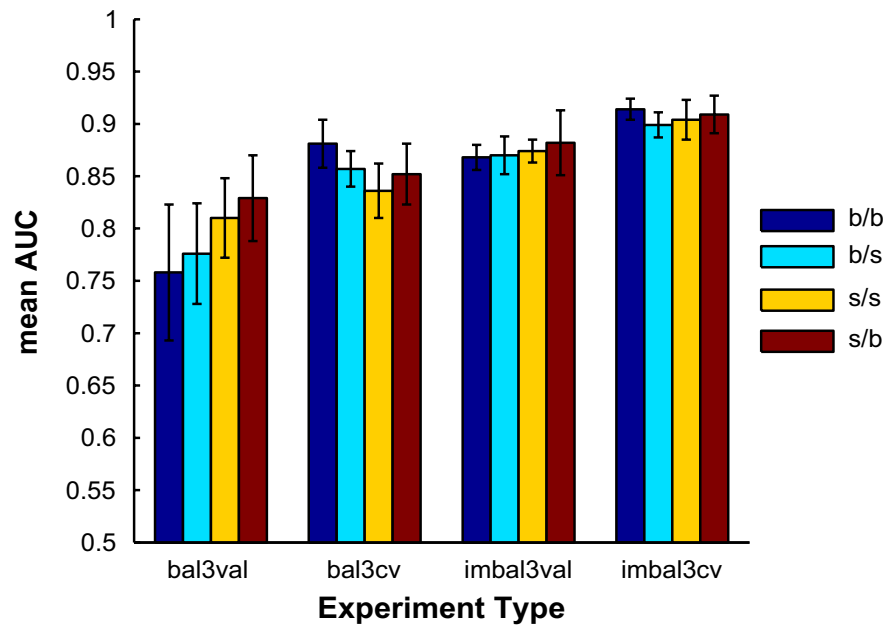


Figure 4.2: Mean overall area under the ROC curve (AUC) and standard deviation of each experiment type for each activation combination.

When looking at the experiments corresponding to the different repeats, the two experiments with the highest overall AUC values were obtained by the imbalanced and cross-validated experiments with the sigmoid and the bipolar sigmoid function on both neuron types. The overall AUC values obtained were 0.933 and 0.932, respectively. These experiments resulted in a TP% over all test profiles of 74.3% and 79.2%, respectively. Important to note is that even though stratified test sets were used, due to the imbalance effect this metric is somewhat biased towards the major classes. This bias problem was tackled by an alternative evaluation approach described in the following section. Together with the results in Table 4.1, it can be concluded that, given the possibility that some species are hard to distinguish from others based on FAME profiles alone, a quite good classification can be achieved using ANNs.

To analyze significant difference in performance, a Wilcoxon rank-sum test was performed for each relevant pair of experiment types. Only comparisons were done to analyze the effect of the balance type and the validation type. This implies that only those experiment types were compared differing in the respective parameter. An overview of the p -values is given in Table 4.2.

Several general conclusions can be drawn from Figure 4.2 and Table 4.2:

1. Classification of the species of the imbalanced data sets led to higher AUC values
2. Cross-validation led to better results
3. No winning activation combination was detected

These conclusions are discussed into more detail in the following sections.

4.2.2.2 Effect of Balance Type

To analyze the effect of class balance type, experiment types with equal validation type were compared. Figure 4.2 shows that imbalanced experiment types led to higher mean AUC values.

<i>p</i> -value	bal3val	bal3cv	imbal3val	imbal3cv
bal3val		0.0007*	0.0002*	
bal3cv	0.0002*			0.0003*
imbal3val	0.0002*			0.0013*
imbal3cv		0.0008*	0.0002*	
bal3val		0.3256	0.0032*	
bal3cv	0.1403			0.001*
imbal3val	0.0002*			0.0172*
imbal3cv		0.0002*	0.0013*	

Table 4.2: *p*-values of the Wilcoxon rank-sum test based on the mean overall AUC of each experiment type for each activation combination. Each activation combination corresponds to a triangle: b/b (top-left), b/s (top-right), s/s (bottom-left) and s/b (bottom-right). Significantly better identification performance is indicated by an asterisk ($p < 0.05$).

Table 4.2 indicates that all eight comparisons of balanced versus imbalanced experiments types showed a significantly better performance when considering imbalanced data sets. For our experimental setup, this is extremely important as our data were highly costly. These results can be explained due to the incorporation of a larger number of profiles and thus integration of a larger intra-species heterogeneity. Beside this, errors in imbalanced experiments were averaged over more data points, resulting in a delayed early stopping. This implied longer training which, together with the larger heterogeneity, led to a better generalization of the data and a better identification of unknown patterns. From Table 4.1 and Figure 4.2 it can be seen that the standard deviations in the balanced experiment types were larger than those of the imbalanced experiment types. This implies that the balanced data sets were inherently linked to the loss of critical information and that the corresponding classification was highly dependent on data sampling. Altogether, we can conclude that it is more difficult to distinguish between FAME profiles of the different species when considering balanced data sets. This led to the final choice of training on imbalanced data sets.

4.2.2.3 Effect of Validation Type

To analyze the effect of the validation type, experiments with equal balance type were compared. Figure 4.1 and Table 4.2 indicate that stratified cross-validation led to higher mean AUC values than simple validation. From Table 4.2 it can be concluded that in six out of eight comparisons of the experiment types, the experiments validated by stratified cross-validation led to a significantly better performance. Cross-validation can generally be expected to lead to better results, in contrast to simple validation, as the errors during cross-validation are averaged over the different folds and, thus, have less impact on the stopping of the ANN training. As a consequence, the final error function during cross-validation increases less quickly. In the experimental setup, stratified cross-validation should be preferred over simple validation.

4.2.2.4 Effect of Activation Type

As mentioned above, no clear winning activation combination can be deduced from Figure 4.2. A Wilcoxon rank-sum test was performed to test significantly different activation combinations. Tests were performed for each pair of activation combinations by comparing equal experiment types. Significant test results were obtained with a p -value below 0.05. The numbers of significantly different mean AUC values are reported in Table 4.3. No significantly better activation combination over all experiment types was found. Consequently, this parameter should be determined empirically through optimization.

p -value	b/b	b/s	s/s	s/b
b/b				
b/s	$\frac{1}{4}$			
s/s	$\frac{2}{4}$	$\frac{0}{4}$		
s/b	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{0}{4}$	

Table 4.3: Effect of the activation type. Number of activation combinations for the four experiment types leading to significantly different AUC values based on the Wilcoxon rank-sum test ($p < 0.05$)

4.2.2.5 Winner-Take-All

The goal of first-line identification tools, such as gas chromatographic FAME analysis, is not to achieve an exact identification but rather to narrow down the bacterial spectrum. Hence, an additional test was done by analyzing, for each test profile, the species labels corresponding to the output neurons with the five highest output values. In this experimental setup, a TP was seen as a hit when the correct species label corresponded with one of the five highest scores (annotated as 'five-best' in Fig 4.3). The mean TP% was calculated by dividing the number of TPs by the number of test profiles and averaging it over the ten repeats. Note that the aforementioned remark about the imbalance effect also holds here. Similarly, this metric was calculated for all experiment types in the original experimental setup (annotated as 'first' in Figure 4.3). Accuracies of each experiment type with each activation combination are plotted in Figure 4.3. Similar results were obtained for different activation combinations. It is obvious that for all experiment types the 'five-best' approach led to better identification results and that this approach met the goal of rapidly narrowing down the bacterial spectrum.

4.2.2.6 Closely Related Species

A closer look at the identification results leads to the conclusion that some species are better identified than others. An important issue related to this problem is the distinctness between species. The genus *Bacillus* contains two groups of species that are closely related: the *Bacillus cereus* group, which contains the species *B. anthracis*, *B. cereus*, *B. mycoides*, *B. pseudomycoides*, *B. thuringiensis* and *B. weihenstephanensis* (Euzéby 2007), and the *Bacillus subtilis*

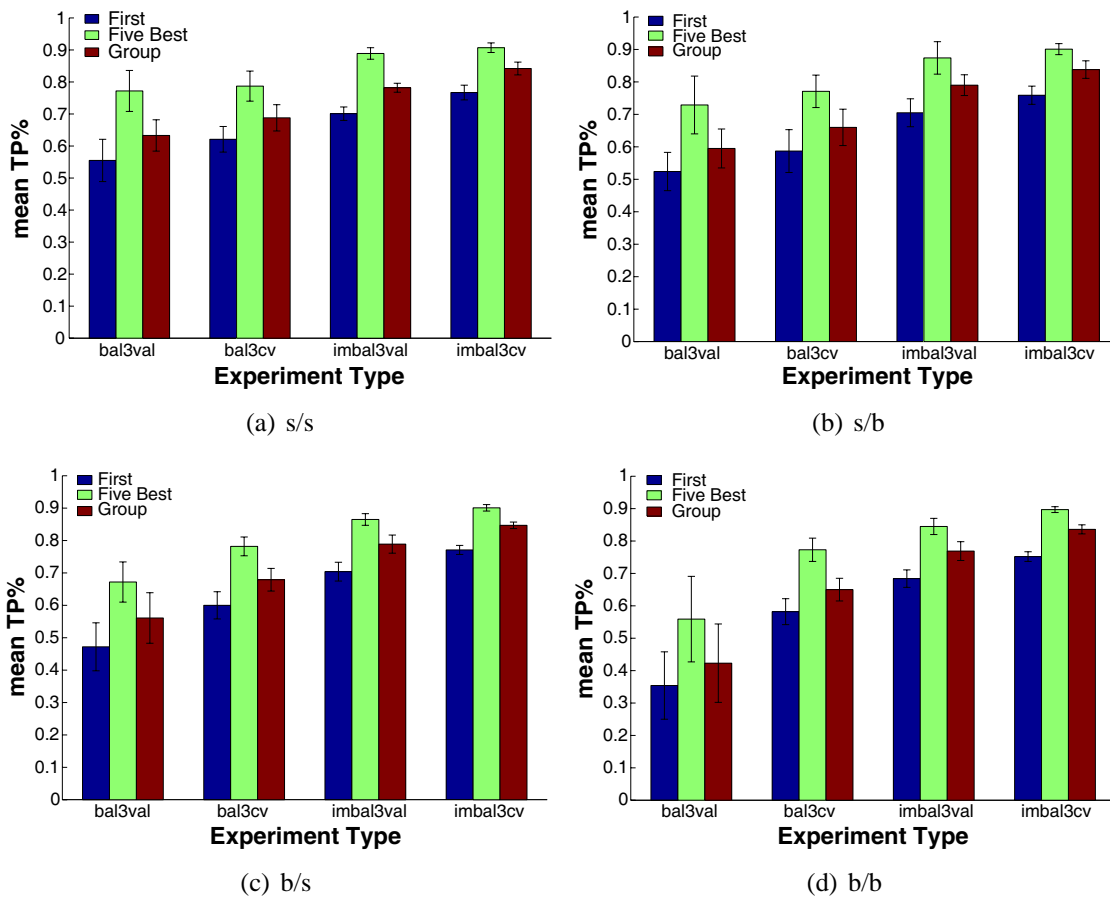


Figure 4.3: Mean true positive (TP) percentages for each experiment type and for each activation combination. Percentages are given for the identification of the correct species name as highest output score (First), as present in the five highest output scores (Five Best) and as highest output score when considering the *B. cereus* and *B. subtilis* species groups (Group). An experiment type consists of a balance type (bal/imbal) and a validation type (val/cv).

group, which contained in 2006 the species *B. amyloliquefaciens*, *B. atropheus*, *B. axarquienensis*, *B. licheniformis*, *B. malacitensis*, *B. mojavensis*, *B. pumilus*, *B. sonorensis*, *B. subtilis*, *B. tequilensis*, *B. vallismortis* and *B. velezensis* (Gatson et al., 2006; Hutsebaut et al., 2006). Note that some rearrangements have taken place in the last years. Updated information can be found in Subsection 2.2.3.1. Identification tools will hardly be able to distinguish between the *B. cereus*-related species. This can be deduced from the highly similar fatty acid profiles of *B. cereus* and *B. thuringiensis* (see also data analysis experiments); from genetic rRNA, gene and plasmid sequence analysis; from population genetic studies and comparative genomic analysis; and because these species can only be differentiated based on their morphology, phenotype and pathogenicity (Drobniewski, 1993; Kämpfer, 1994; Bavykin et al., 2004; Tourasse et al., 2006). For the *B. subtilis* group a similar conclusion can be drawn. Fatty acid profiles of *B. amyloliquefaciens*, *B. licheniformis* and most strains of *B. subtilis* are highly similar and strains of *B. amyloliquefaciens* show fatty acid patterns almost indistinguishable from those of *B. subtilis* (Kämpfer, 1994). Nonetheless, two studies showed that *B. amyloliquefaciens*, *B. licheniformis* and *B. pumilus* species could be discriminated (Vaerewijck et al., 2001; Coorevits et al., 2008). Based on 16S rRNA analysis, Ash et al. (1991) showed that *B. subtilis*-related species form a distinct clade in the phylogenetic *Bacillus* tree. As a result, one cannot expect that ANNs

will be able to classify group-related species perfectly. Based on this prior knowledge, a new evaluation of the experiments was done. When the profile of a species belonging to a species group was identified by the winner-take-all rule as a member of that group, then the species was annotated as a TP (annotated as 'Group' in Figure 4.3). Also in this case, a mean TP% was calculated and the results are plotted in this figure. Also, this figure clearly shows that considering species groups improves the mean TP%. This also confirms that FAME profiles of group-related species are highly similar and that it is difficult to differentiate between these species based on FAME data. Beside the existence of species groups, it is also possible that some species that are generally not regarded as belonging to species groups, are closely related to other species. Therefore, identification methods should take into account that these species cannot be separated based on their whole-cell fatty acid content. In both cases, analysis of the resolution of FAME analysis for species discrimination and the integration of this resolution information into the machine learning models will further enhance the FAME-based species identification and contribute to the goal of a first-line identification tool. This integration is discussed into more detail in Chapter 5.

4.2.3 Publication

This section is published in the international peer-reviewed journal *Antonie van Leeuwenhoek International Journal of General and Molecular Microbiology* with reference: B. Slabbinck, B. De Baets, P. Dawyndt and P. De Vos (2008). Genus-wide *Bacillus* species identification through proper artificial neural network experiments on fatty acid profiles. *Antonie van Leeuwenhoek International Journal of General and Molecular Microbiology*, 94, 187–198.

4.3 Three Genera - Three Techniques

4.3.1 Methodologies

In this section, we study the performance of the three machine learning techniques ANNs, RFs and SVMs for species prediction within the three genera *Bacillus*, *Paenibacillus* and *Pseudomonas*. Only the data sets of March 2008 were considered, together with a genera data set composed by merging the different genus data sets (see also Table 3.1). Thus, also genus identification was considered. For ANNs, the same setup was used as defined in Section 4.2.1. Parameter optimization was done by three-fold cross-validation. For RFs, the number of trees was first optimized in the intervals [1000, 4000] in steps of 250 and by setting the number of split variables on its default value, i.e. the root of the number of features. Subsequently, the number of split variables was optimized by evaluating the default value, twice the default value and half of the default value. The optimal number of trees was used during this optimization step. The RF technique has the advantage of not overfitting given a large number of trees and, therefore, optimization of the parameters was performed with the test set and not by cross-validation within the training set. Thus, the test set error was used as error criterion. For SVMs, the multiple one-versus-one optimization as implemented in the LibSVM software

(further denoted as SVM) and the single optimization as implemented in the BSVM software (further denoted as BSVM) were used with the linear (abbreviated as ‘lin’) and RBF kernels. The cost parameter C and the RBF kernel parameter γ were optimized using three-fold stratified cross-validation grid search. For the RBF kernel, the LibSVM and BSVM package provides a Python script for optimizing both parameters. Default value ranges of both scripts were used and are set to $[2^{-5}, 2^{15}]$ in steps of 2^2 and $[2^{-15}, 2^3]$ in steps of 2^2 , respectively. For the linear kernel, the C parameter was optimized over the value range $[2^{-5}, 2^{15}]$ in steps of 2^2 in LibSVM and $[2^{-8}, 2^8]$ in steps of 2^2 in case of BSVM. A smaller range was chosen for BSVM as computation time of this method was multiple times higher than that of LibSVM. If the maximum value was reached, the interval range was extended with two steps. The probability outputs as given by the software programs were used for further statistical analysis. The performance of each technique was statistically analyzed as described in Section 1.3.

As mentioned above, also a genera data set was created. With this data set, genera identification was aimed at and a complete species identification over the three genera (see also following subsection). For genera classification, ten-fold cross-validation was performed, due to the larger amount of data available per class.

For all techniques, custom JAVA programs were developed for processing of the data, parameter optimization, learning and testing. Hereby, the pipeline from data set to statistical analysis of the prediction results was completely automated. Overall, each experiment was performed ten-fold, each time with randomly sampled training and test sets. Statistical measures resulting from the analysis were subsequently averaged over the experiments. These experiments were also performed on a high-performance Blade cluster (Intel Corporation, Santa Clara, CA, USA). Identification was done on any a single PC.

4.3.2 Experimental Design

Two strategies for classification and identification were evaluated which are schematically represented in Figure 4.4. In the stratified identification strategy, genus and species identification were performed by separate identification models. Identification was first performed at genus level, followed by identification at species level. For genus identification, FAME profiles were only annotated by genus name. At species level, a species identification model was generated based on a genus-specific data set. This data set comprised only FAME profiles of the species of a particular genus, which were annotated by genus and species name. At species level, profiles were only identified by the genus identification model corresponding with the identification label of the profiles retrieved by genus identification. By considering only the highest score of each profile following genus identification, this restricts the flexibility in identification. As a more flexible approach, alternative solutions could be considered here such as an interval of highest output scores or a weighted approach. The second approach was the straight identification strategy. Herein, one single species identification model was generated based on the complete data set, in which the FAME profiles were annotated by both genus and species name. This approach made the classification task quite hard with respect to the closely related species of the genera *Bacillus* and *Paenibacillus*. The genus *Pseudomonas* is quite distant from

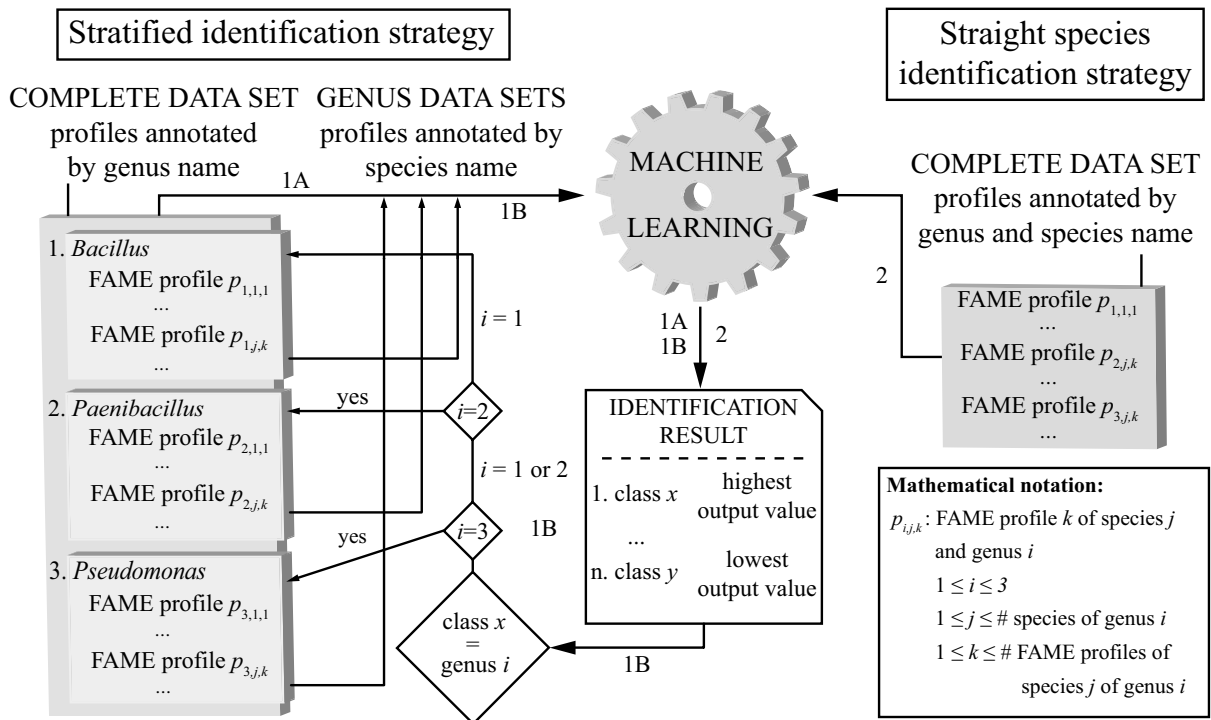


Figure 4.4: Schematic presentation of the experimental design. 1A/B. Stratified identification strategy. 1A. Genus identification is performed by the genus identification model. This model relies on the complete FAME data set in which the profiles are annotated by genus name (dark grey box). 1B. For each genus, a species identification model is built based on the FAME profiles corresponding to that specific genus. The respective FAME profiles are annotated by species name (light grey boxes). In both cases, each profile is labelled with the genus or species name associated with the highest output value. However, species identification is only performed for the genus associated with the highest output value following genus identification. 2. **Straight species identification strategy.** The complete data set of FAME profiles is annotated by genus and species name (dark grey box). Identification is performed by a single identification model. Each profile is labelled with the genus and species name associated with the highest output value.

these two other genera, as clearly visualized in the biplot of the first two principal components (see Figure 3.16), and will not hamper the calculation of margins between *Pseudomonas* species and species of the two other genera. Importantly, for evaluation of the global performance of each individual identification model, all data were considered, even though profiles could be misidentified by the genus identification model in a stratified identification strategy!

For comparison with Sherlock MIS, the stratified identification strategy was considered. For each species identification model, the training and test set combination resulting in the highest AUC value was considered. Ultimately, identification at genus level was achieved by merging the respective training and test sets and by training and testing a genus identification model based on the merged data sets. As such, it was possible to identify the same profiles for genus and species identification and to rule out those profiles incorrectly identified at genus level. Importantly, the same machine learning technique was considered for the species identification models as well as for the genus identification model. Even though it is possible that different machine learning techniques result in better performance on different data sets, we focused on the same technique for ease of implementation.

Metric	ID strategy	ANN s/s	ANN b/s	ANN b/b	ANN s/b	RF	SVM RBF	SVM lin	BSVM RBF	BSVM lin
AUC	Stratified ID									
	Genus	0.992 (0.004)	0.992 (0.006)	0.991 (0.006)	0.993 (0.003)	0.998 (0.001)	0.997 (0.002)	0.996 (0.002)	0.993 (0.004)	0.971 (0.039)
	Bacillus	0.966 (0.011)	0.972 (0.008)	0.966 (0.008)	0.964 (0.011)	0.988 (0.007)	0.981 (0.005)	0.977 (0.008)	0.962 (0.108)	0.950 (0.012)
	Paenibacillus	0.970 (0.008)	0.971 (0.014)	0.971 (0.007)	0.965 (0.013)	0.990 (0.015)	0.976 (0.013)	0.983 (0.005)	0.970 (0.087)	0.947 (0.007)
	Pseudomonas	0.944 (0.008)	0.951 (0.005)	0.950 (0.006)	0.937 (0.014)	0.987 (0.003)	0.979 (0.003)	0.979 (0.003)	0.938 (0.013)	0.929 (0.007)
	Straight ID	0.971 (0.003)	0.973 (0.004)	0.974 (0.004)	0.964 (0.005)	0.991 (0.002)	0.988 (0.002)	0.988 (0.002)	0.968 (0.013)	0.970 (0.004)
Se	Stratified ID									
	Genus	0.979 (0.006)	0.978 (0.007)	0.979 (0.009)	0.978 (0.005)	0.977 (0.006)	0.985 (0.005)	0.979 (0.006)	0.985 (0.006)	0.969 (0.011)
	Bacillus	0.753 (0.028)	0.740 (0.024)	0.731 (0.024)	0.748 (0.036)	0.847 (0.021)	0.544 (0.053)	0.457 (0.036)	0.824 (0.291)	0.764 (0.030)
	Paenibacillus	0.753 (0.047)	0.734 (0.037)	0.724 (0.045)	0.749 (0.046)	0.901 (0.040)	0.610 (0.068)	0.551 (0.044)	0.824 (0.323)	0.823 (0.043)
	Pseudomonas	0.551 (0.039)	0.537 (0.037)	0.501 (0.035)	0.523 (0.047)	0.673 (0.014)	0.281 (0.028)	0.272 (0.026)	0.702 (0.037)	0.700 (0.025)
	Straight ID	0.669 (0.021)	0.669 (0.016)	0.633 (0.032)	0.634 (0.021)	0.732 (0.015)	0.239 (0.011)	0.232 (0.010)	0.740 (0.019)	0.751 (0.015)
Pr	Stratified ID									
	Genus	0.984 (0.004)	0.981 (0.007)	0.984 (0.004)	0.984 (0.003)	0.982 (0.004)	0.989 (0.003)	0.983 (0.006)	0.988 (0.003)	0.974 (0.009)
	Bacillus	0.812 (0.030)	0.798 (0.026)	0.803 (0.031)	0.748 (0.036)	0.908 (0.013)	0.829 (0.043)	0.751 (0.055)	0.853 (0.231)	0.834 (0.023)
	Paenibacillus	0.815 (0.049)	0.796 (0.036)	0.775 (0.043)	0.803 (0.062)	0.947 (0.018)	0.775 (0.264)	0.800 (0.247)	0.854 (0.250)	0.846 (0.042)
	Pseudomonas	0.671 (0.023)	0.669 (0.041)	0.645 (0.031)	0.643 (0.038)	0.851 (0.023)	0.708 (0.035)	0.688 (0.021)	0.743 (0.026)	0.722 (0.021)
	Straight ID	0.745 (0.026)	0.757 (0.022)	0.728 (0.031)	0.718 (0.021)	0.882 (0.009)	0.661 (0.019)	0.634 (0.031)	0.787 (0.012)	0.787 (0.013)
F	Stratified ID									
	Genus	0.980 (0.005)	0.980 (0.004)	0.981 (0.005)	0.981 (0.006)	0.979 (0.004)	0.987 (0.004)	0.982 (0.006)	0.986 (0.004)	0.971 (0.010)
	Bacillus	0.821 (0.025)	0.826 (0.026)	0.808 (0.019)	0.815 (0.024)	0.893 (0.014)	0.810 (0.020)	0.782 (0.024)	0.854 (0.182)	0.850 (0.018)
	Paenibacillus	0.854 (0.024)	0.839 (0.026)	0.840 (0.023)	0.843 (0.022)	0.940 (0.018)	0.850 (0.023)	0.818 (0.021)	0.876 (0.179)	0.870 (0.020)
	Pseudomonas	0.699 (0.024)	0.718 (0.018)	0.702 (0.012)	0.697 (0.034)	0.805 (0.016)	0.692 (0.022)	0.661 (0.025)	0.757 (0.014)	0.758 (0.015)
	Straight ID	0.783 (0.020)	0.786 (0.011)	0.771 (0.016)	0.764 (0.016)	0.843 (0.016)	0.677 (0.018)	0.666 (0.204)	0.810 (0.008)	0.807 (0.007)

Table 4.4: Overview of the results for genus and species identification. Classification performance is indicated by the Area Under the ROC Curve (AUC). Identification performance is indicated by sensitivity (Se), precision (Pr) and F-score (F). Results are reported for two identification (ID) strategies: stratified and straight identification. The stratified identification strategy performs identification at genus level and at species level for the three genera *Bacillus*, *Paenibacillus* and *Pseudomonas*. Three machine learning techniques are used: artificial neural networks (ANNs) with a sigmoid (s) and/or bipolar sigmoid (b) activation function on the hidden and output neurons ($f_{\text{hidden}}/f_{\text{output}}$), random forests (RFs) and support vector machines (SVMs) with RBF and linear (lin) kernel. Performance values and standard deviations are reported. Highest performance values are indicated in bold face.

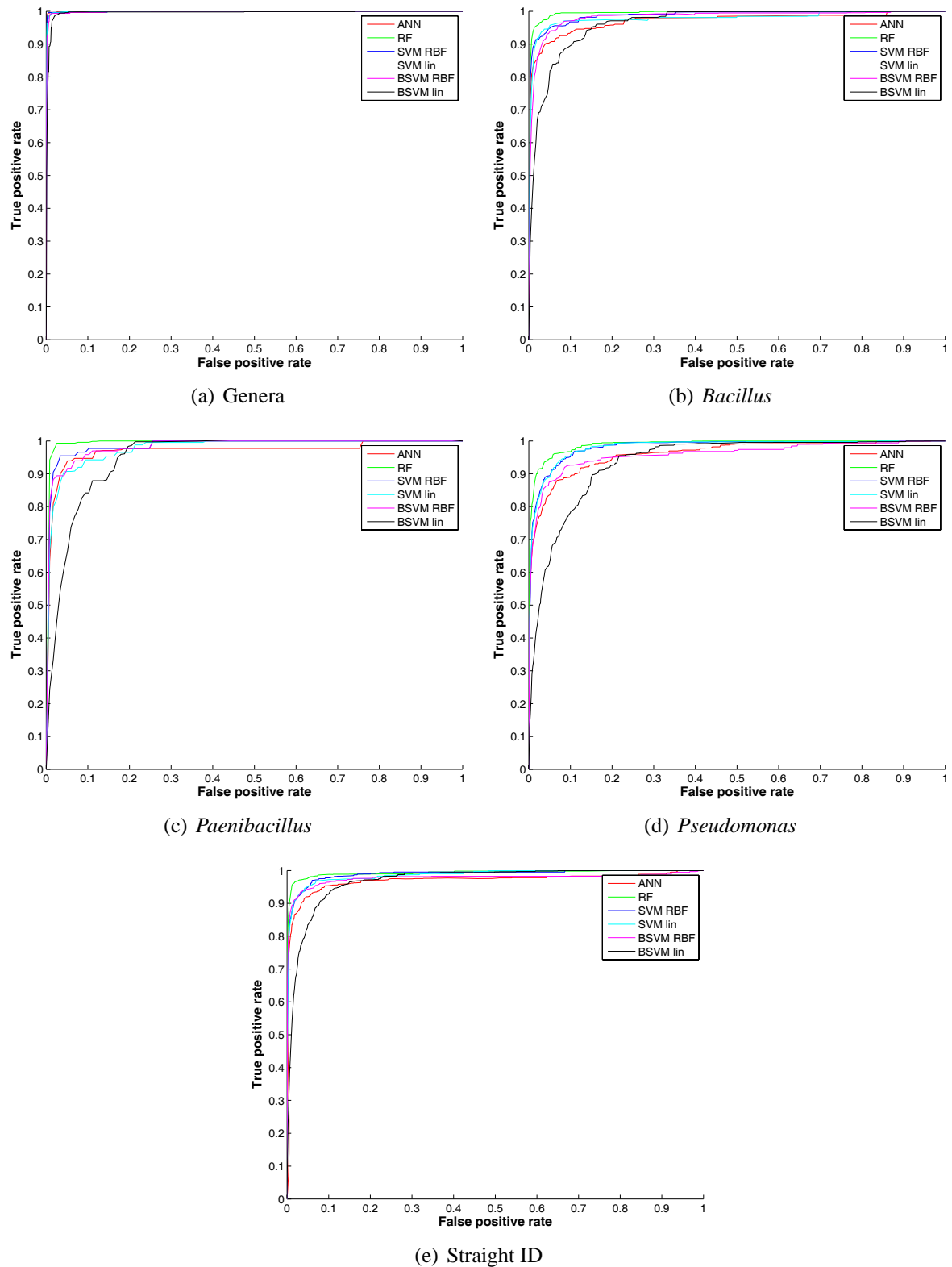


Figure 4.5: Average ROC. Average ROC plots of the ANN, RF, SVM with RBF and linear (lin) kernel and BSVM with RBF and linear kernel classification experiments for each genus, the genera classification experiment and the straight identification experiment. Each time, the experiment with the highest AUC was selected. ROC values were calculated for each class separately and vertically averaged over the classes.

4.3.3 Results

In the stratified identification setting, genus identification was performed preliminary to species identification. A detailed report of the accuracy values is given in Table 4.4. A simple visualization of the individual AUC values calculated for each class (species or genus) is given in Figure 4.5, which shows the average ROC curves as obtained for each model and each technique. An average ROC curve is an approximation of the average of the individual ROC curves of each class. More information can be found in Subsection 1.3.2. As could be expected, and could be deduced from PCA (see Figure 3.16), the three machine learning techniques resulted in very high FAME-based genus classification performances. At genus level, among all experiments, the highest Se value of 0.997 and Pr value of 0.994 was attained with the same BSVM RBF model, with respective standard deviations 0.005 and 0.009. The multi-class confusion matrix of the best RF model (based on the AUC value) is shown in Table 4.5. The three *Bacillus* profiles corresponding with an identification as *Paenibacillus* corresponded to the species *B. coagulans* (1 profile) and *B. simplex* (2 profiles). As both genera are closely related, it is not surprising that some profiles of either genus were identified as the other. Thus, this is also the case for the four *Paenibacillus* profiles that were identified as the genus *Bacillus*, i.e. *Pa. cineris*, *Pa. kobensis*, *Pa. larvae* and *Pa. macerans*. The *Pseudomonas* profile identified as *Bacillus* corresponded to *P. flectens*, which according to PCA analysis is more related to *Bacillus* than to the different *Pseudomonas sensu stricto* species. Note that *P. flectens* is not considered as a *Pseudomonas sensu stricto* species but, based on 16S rRNA sequence analysis, clusters in the *Enterobacteriaceae* group (Anzai et al., 2000). Generally, it can be concluded that almost perfect genus identification is achieved. Note the remark in the section about the experimental design, where we stated that misidentified profiles at genus level were still considered for the evaluation at species level. It can be considered that, due to the very small number of misidentifications at the genus level, the inclusion of these misidentified profiles only have a very small effect on the global performance.

		Predicted genus		
		<i>Bacillus</i>	<i>Paenibacillus</i>	<i>Pseudomonas</i>
True genus	<i>Bacillus</i>	317	3	0
	<i>Paenibacillus</i>	4	122	0
	<i>Pseudomonas</i>	1	0	556

Table 4.5: Multi-class confusion matrix resulting from genus identification by the best RF experiment. The number of correct predictions are presented on the main diagonal, the other cell values show the number of incorrect predictions. Row labels correspond to the true genus names, column labels correspond to the predicted genus names.

At species level and for each genus considered, the identification accuracies were different for the three machine learning techniques. AUC values for each genus and each classifier were quite high. As these were calculated in a one-versus-others setting, this means that the obtained identification scores were quite high but not the highest scores, while the latter scores were used for the calculation of the metrics sensitivity, precision, F-score and specificity. The specificity

metric is not shown as, due to the large number of classes, the imbalanced nature of the data set and the one-versus-others evaluation approach, the number of true negatives outnumbered the number of false positives, making discrimination between specificity values irrelevant (all values approximated 100%). Note also, that precision and F-score can result in a value equal to ∞ . The respective species are not considered in the calculation of an average metric value (see also Section 1.3). The number of metric values per experiment equal to ∞ is not reported in Table 4.4. This number was not very high, except for the SVM experiments. Among all *Bacillus* species identification experiments, the highest Se value of 0.885 and Pr value of 0.926 were achieved by a RF model, with respective standard deviations of 0.216 and 0.143. The overall *Bacillus* species identification could be considered very high, given the presence of species groups of closely related species and the many overlapping species data patterns as seen in the clustering and PCA analysis. As compared to the *Bacillus* ANN experiments of the previous section, a small increase was seen in the AUC values. This is due to the different approach of calculating the AUC. By setting the output score of each profile as a threshold, a better approximation was obtained in this section. For *Paenibacillus* species identification, among all experiments, the highest Se value and Pr value was also achieved by RFs (standard deviations in brackets): 0.974 (0.092) and 0.981 (0.081), respectively. *Paenibacillus* species identification could, thus, be considered very high. Among all *Pseudomonas* species identification experiments, one BSVM RBF model led to the highest Se value of 0.753 (0.310) and the highest Pr value of 0.887 (0.201) was obtained by a RF model. This result was quite surprising (positively) given the many overlapping species as concluded from clustering and PCA analysis. Overall, the three machine learning techniques resulted in a very high FAME-based genus identification performance, given the constraints as shown by the preceding data analysis. In the case of species identification, the RF and BSVM techniques outperformed the ANN and SVM technique, where the ANN technique also outperformed the SVM technique. Based on identification performance and computation time, RFs was clearly the best technique for this type of classification. From these statistics, it could be concluded that it is possible to achieve a moderate to good accuracy for FAME-based *Bacillus*, *Paenibacillus* and *Pseudomonas* species identification by machine learning.

It is also interesting to look at the straight species identification strategy. Table 4.4 shows classification and identification accuracies of the different methods for the data set covering the species of all three genera. This strategy concerned a straight species identification in which genus identification was not considered. Generally, a moderate to good FAME-based species classification performance was achieved by the three techniques. Among all species identification experiments, the highest Se value and Pr value was 0.778 (0.322), obtained by a BSVM lin experiment, and 0.898 (0.186), obtained by a RF model. Generally, the RF, BSVM and ANN methods outperformed the SVM technique. From these statistics, it could be concluded that FAME-based species identification in this identification strategy resulted in a moderate accuracy.

4.3.4 Discussion

4.3.4.1 Stratified Identification Strategy

The results obtained indicate that, when considering FAME data, the three machine learning techniques RFs, ANNs and SVMs resulted in a nearly perfect genus identification. A first approach towards FAME-based identification of bacterial genera by machine learning was taken by Bertone et al. (1996) and Giacomini et al. (2000, 2004) who successfully identified a limited number of marine and environmental bacteria at genus level by ANNs. The researchers concluded that FAMEs are good biomarkers for bacterial genus identification and that it would be worthwhile to build a FAME-based bacteria identification system at species level. In a taxonomic context, a first in-depth study on FAME-based species identification by machine learning techniques was performed for the genus *Bacillus* as described in Section 4.2. From this study, we concluded that species identification by FAME data and machine learning techniques is very promising, taking into account the limited resolution of FAME analysis for species discrimination. The research as described in this section extended the scope of the previous section by also evaluating species identification in the genera *Paenibacillus* and *Pseudomonas* by three machine learning techniques: RFs, ANNs and SVMs. These genera were selected because two genera, *Bacillus* and *Paenibacillus*, belong to the same phylum and are closely related. The third genus, *Pseudomonas*, was selected to also include a distantly related genus belonging to a different phylum. From a genus-wide identification perspective, all genera were represented in the LMG FAME database by a sufficient number of species to cover at least half of the validly published species (see Table A.2). When considering genus classes only, the three genera *Bacillus*, *Paenibacillus* and *Pseudomonas* could easily be distinguished from each other based on FAME data. Furthermore, analysis of the multi-class confusion matrices showed that misclassifications of the FAME profiles were mainly due to misclassifications of *Bacillus* profiles as *Paenibacillus*, and conversely (example given in Table 4.5). This result was expected as both genera are evolutionary more related to each other than to the genus *Pseudomonas*. Nonetheless, genus identification was surprisingly good when taking into account that the genera *Bacillus* and *Paenibacillus* were reported are closely related. Even though different bacterial genera can possibly be hard to distinguish based on FAME data, the strategy of selecting the highest output value for final identification will fail when extending the taxonomic scope towards dozens of bacterial genera. Therefore, the development of an alternative scoring and weighing mechanism will become indispensable for reliable genus and species identification.

Kämpfer (1994) concluded that fatty acid analysis has a potential for species differentiation within the genus *Bacillus*. The application of machine learning techniques for FAME-based *Bacillus* species identification supports this hypothesis. The identification results clearly indicated that species of the genus *Bacillus* could be distinguished and that the application of RFs resulted in the best identification accuracy. From a taxonomic perspective, some species are closely related and are consequently assigned to a species group such as the *B. subtilis* and *B. cereus* groups. Integration of this prior knowledge into computational classification models confirmed that wrong identifications are mostly due to identifications as species of the

same group (see Section 4.2). Moreover, species that are more distantly related through evolution might also show highly similar FAME patterns. When considering this information about species groups and species distinctness, and the presence of 74 *Bacillus* classes, the identification accuracy achieved by RFs could be considered as very good. By looking more into detail to the identification of the different *Bacillus* species, some interesting facts could be seen. A focus was given to those species with an F-score smaller than 0.667. As expected, species residing in a species group were identified as another member of the group. Examples are the four *B. thuringiensis* profiles that were identified as *B. cereus* (*B. cereus* group), one profile of *B. sonorensis* (two in total) that was identified as *B. licheniformis* and four of the five profiles of *B. vallismortis* that were identified as *B. subtilis*. The latter two are *B. subtilis* group examples. These misidentification were also found to be the closest neighbours in the TaxonGap experiment (see Figure 3.12). Other examples were described in literature as closely related: the two *B. muralis* profiles were identified as *B. simplex* (Heyrman et al., 2005), one *B. vireti* profile that was identified as *B. novalis* (two in total, same isolation source and confirmed by TaxonGap) (Heyrman et al., 2004) and one profile of *B. clausii* identified as *B. lentus* (Nielsen et al., 1995). Except for the latter example, the mentioned relationships of all previous examples were confirmed in the maximum likelihood 16S rRNA gene phylogenetic tree constructed in this study (see Figure 2.4). No literature description was found for the misidentification of *B. jeotgali* as *B. subterraneus* (1 profile), though both species grouped together in the same cluster of the maximum likelihood tree and *B. subterraneus* was also its closest neighbour in the TaxonGap experiment. Of course, low F-score values may be found due to a large number of false positives (resulting in a high precision), as found for the species *B. decolorationis* and *B. subterraneus*. No clear relation could be put on some misidentifications, not in literature nor in the constructed maximum likelihood tree. Examples of these misidentifications were the identification of *B. azotoformans* as *B. megaterium** (1/1 profile), *B. firmus* as *B. aquimaris* (2/3 profiles), *B. fortis* as *B. aquimaris* (1/1 profile), *B. gibsonii* as *B. lentus** (1/2 profiles), *B. halmapalus* as *B. subterraneus* (1/2 profiles), *B. lentus* as *B. pumilus* and as *B. cereus** (1/4 and 1/4 profiles), and *B. pseudocaliphilus* as *B. pumilus* (1/1 profile). Some relationships were confirmed by the TaxonGap experiment and are denoted by an asterisk. Two main possible reasons for these results are similarities in the FAME profiles or misidentification by a species representing a large number of FAME profiles. The latter case is a consequence of the class imbalance problem. Herein, the margins for these species are more confident and minor classes (species with a small to very small number of FAME profiles) are dominated by these classes. Or, profiles of these classes are more likely identified as the major class. Almost all relationships were also confirmed by the biplots of the PCA analysis described in Subsection 3.3.4.

Similar to the FAME analysis of Kämpfer (1994), Heyndrickx et al. (1996) concluded that FAME analysis allows genus identification and identification of *Paenibacillus* species into several species groups. As about one-fourth of the species in the genus *Paenibacillus* has been validly published since January 2006, no in-depth study of species discrimination by FAME analysis has previously been performed. Our identification results show that species in the genus *Paenibacillus* could be distinguished from each other based on their FAME profiles and machine learning techniques. When looking at the species with an F-score ≤ 0.667 , only one

species was not correctly identified. The two profiles of *Pa. xylanilyticus* were identified as *Pa. illinoisensis*. This relates to 16S rRNA analysis which showed a close relatedness between these species (Rivas et al., 2005), which is also nicely shown in the maximum likelihood 16S rRNA gene phylogenetic tree constructed in this study (see Figure 2.6). This relationship could also be seen in the data analysis experiments where in the PCA biplots the data of both species were overlapping and TaxonGap (see Figure 3.13) also confirmed that *P. illinoisensis* was the closest neighbour of *P. xylanilyticus*. Three other species had a lower F-score but these were mainly due to a large precision, or thus a large number of false positives.

Identification results showed that *Pseudomonas* species were harder to distinguish than those of *Bacillus* and *Paenibacillus*. This could be expected from preliminary data analysis and clustering. Fatty acid analysis of pseudomonads has been a matter of discussion for several decades (Ikemoto et al., 1978; Moss et al., 1972; Moss and Dees, 1976; Moss, 1981; Oyaizu and Komagata, 1983; Welch, 1991). Two broad studies on this issue were reported by Stead (1992) and Vancanneyt et al. (1996) showing that analysis of whole-cell fatty acid fingerprints of pseudomonad strains revealed major groups and subgroups corresponding well to the groupings based on DNA-DNA and DNA-rRNA hybridization techniques. The strains of rRNA group I in the study of Vancanneyt et al. (1996) represented 29 different and validly described *Pseudomonas* species which could be grouped into four major FAME subgroups. Subgrouping of various phytopathogenic species was also found. The authors demonstrated that whole-cell fatty acid data show some qualitative and quantitative differences among the various subgroups and concluded that some species can only be distinguished based on smaller quantitative differences (Stead, 1992; Vancanneyt et al., 1996). Nonetheless, machine learning techniques clearly take advantage of these quantitative differences as the RF identification results showed that on average 67.3% of the species were assigned a correct species label. The above mentioned issues are, however, not the only reason for a lower identification percentage. As mentioned in the introduction, the taxonomy of the genus *Pseudomonas* has been under revision for several decades. In particular, the taxonomic position of various pathovars is still under discussion. In the present study, we chose to follow the *Pseudomonas syringae* taxonomy as proposed by Gardan et al. (1999). Both the limitations of whole-cell FAME analysis for species discrimination and the uncertainties in the taxonomic position of various *Pseudomonas* species were most likely the two main reasons for the lower identification percentage. Nonetheless, RFs maximally exploited the FAME analysis resolution to distinguish *Pseudomonas* species on a genus-wide scale. Also for this genus, species identification was analyzed by setting a threshold on the F-score of 0.667. 16 species were not identified and most misidentifications were species of the *P. aeruginosa* group, *P. fluorescens* group, *P. putida* group and *P. syringae* group (Moore et al., 1996; Anzai et al., 2000). Evidence was found in the tree and/or in literature for the following species (each time the number of misidentified and total number of profiles are given in brackets): *P. abieta-niphila* (*P. putida* group (1/3) and one as *P. vancouverensis*; Mohn et al., 1999), *P. asplenii* as *P. fuscovaginae* (2/4) (Vancanneyt et al., 1996; Tvrozová et al., 2006), *P. avellanae* (*P. syringae* group (1/1); Gardan et al., 1999), *P. brenneri* (*P. fluorescens* group (3/3); Baïda et al., 2001), *P. citronellosis* (*P. aeruginosa* group (2/2); Lang et al., 2007), *P. congelans* (*P. syringae* group (2/2); Behrendt et al., 2003), *P. fluorescens* (*P. fluorescens* group (13/49); Vancanneyt et al.,

1996; Gardan et al., 1999), *P. hibiscicola* as *P. beteli* (1/2) (Van Den Mooter and Swings, 1990), *P. knackmussi* (*P. aeruginosa* group (1/1); Stolz et al., 2007), *P. lurida* (*P. syringae* group (1/2); Behrendt et al., 2007), *P. marginalis* (*P. fluorescens* group (8/31); Vancanneyt et al., 1996; Anzai et al., 2000), *P. plecoglossicida* (*P. putida* group (2/2); Nishimori et al., 2000), *P. poae* (*P. fluorescens* group (1/2); Behrendt et al., 2003), *P. putida* (*P. putida* group (9/30); Moore et al., 1996; Anzai et al., 2000), *P. resinivorans* (*P. aeruginosa* group (2/2); Moore et al., 1996; Anzai et al., 2000), *P. rhodesiae* (*P. fluorescens* group (1/2); Coroler et al., 1996), *P. straminae* as *P. argentinensis* (1/3, other was *P. mendocina*) (Peix et al., 2005), *P. tolaasii* (*P. fluorescens* group (8/17); Vancanneyt et al., 1996; Moore et al., 1996), *P. trivialis* (*P. fluorescens* group (2/2); Behrendt et al., 2003), *P. vancouverensis* as *P. abietaniphila* (1/2, other was *P. putida*) (Mohn et al., 1999) and *P. veronii* (*P. fluorescens* group (2/2); (Elomari et al., 1996)). Only evidence in literature was found for *P. constantinii* which was identified as *P. putida* (1/2 profiles) (Munsch et al., 2002), *P. extremorientalis* also identified as *P. putida* (1/2 profiles) (Ivanova et al., 2002), *P. genomospecies 3 en 7* identified as *P. syringae* group (6/12 and 2/2, not considered in maximum likelihood tree) (Gardan et al., 1999), *P. luteola* identified as *P. aeruginosa* (2/4 profiles) (Anzai et al., 1997) and *P. mediterranea* as *P. antarcticus* (2/3 profiles, the other was *P. fluorescens*) (Catara et al., 2002; Reddy et al., 2004). No clear support for misidentifications was found in the maximum likelihood tree and in literature for the species: *P. argentinensis* as *P. mendocina* (1/2 profiles), *P. azotoformans* as *P. marginalis* (2/2 profiles), *P. brassicacearum* as *P. amygdali* (1/3 profiles) and *P. syringae* (1/3 profiles), *P. caricapapayae* as *P. amygdali* (1/1 profile), *P. jessenii* as *P. fluorescens* (1/2 profiles) and *P. marginalis* (1/2 profiles), *P. korensis* as *P. fluorescens* (2/2 profiles), *P. lini* as *P. fluorescens* (1/3 profiles), *P. monteillii* as *P. fluorescens* (1/2 profiles), *P. psychrotolerans* as *P. oryzihabitans* (1/2 profiles), *P. salomonii* as *P. mediterranea* (2/3 profiles) and *P. umsongensis* as *P. putida* (1/2 profiles). Low F-scores due to a large precision were found in the species: *P. abietaniphila*, *P. argentinensis*, *P. fluorescens*, *P. lini*, *P. marginalis*, *P. mendocina*, *P. putida* and *P. salomonii*. For this latter case, the same arguments hold as mentioned above in the paragraph with the *Bacillus* results. It is clear that the large number of misidentifications is related to two effects: the many overlapping data patterns as found in the data analysis experiments and the large number of profiles with only a very small number of FAME profiles (see also Figure 3.2).

4.3.4.2 Straight Species Identification Strategy

Instead of considering a layered identification system, it is also possible to build a single model including all species classes. As the FAME profiles of the three genera could clearly be distinguished by machine learning techniques, the performance of a straight species identification model could be considered as a superposition of the three individual genus models. This is nicely shown by the ultimate identification performance which is close to the average of the identification performance of the separate models. Results indicate that the identification accuracy of this approach was confined by the genus *Pseudomonas* which had the lowest identification performance (the largest number of species and profiles) but, had also a larger weight in the calculation of the statistical metrics (averaged over all classes). Besides this, additional

issues should be taken into account for not choosing this type of identification approach. The complete data set comprised 213 classes of which some had few data points per class. This led to a harder classification task with respect to the stratified strategy where genus classification was performed with three genus classes that comprised many data points per class and, subsequently, species classification with 74, 44 and 95 classes per genus model. Moreover, considering the rapidly evolving taxonomy, retraining of the complete model will become necessary in order to achieve up-to-date identification. More classes also result in longer training times. In contrast, in a layered system only those species identification models which correspond to updated data sets need to be retrained. As such, dropping the genus identification model should not be an option as a layered system clearly resulted in better identification performance and better scalability.

4.3.4.3 Comparison of Machine Learning Techniques

Regarding genus identification, making a final choice between the different machine learning techniques is not possible as the values of all statistical measures are (very) similar. When looking at computation times, a preference could be given to the BSVM approach with the RBF kernel which resolves the classification task in less than ten minutes. From all the machine learning techniques evaluated in this study, the best species identification accuracy was achieved by RFs, while the one-versus-one SVM approach as implemented in LibSVM resulted in the worst accuracy. The application of this SVM approach resulted in a very poor identification performance, which could possibly be due to two main reasons. Besides the choice of an inappropriate kernel, the many species classes with only few FAME profiles per class played a crucial role in this classification approach. The main disadvantage of this method is related to this latter fact as, most possibly, a bad performance was obtained when handling two classes which contained only few data points per class. It is clear that a reliable and high-performant classification model cannot be constructed based on a data set with very limited class sizes, which is the case in our setup as the species classes contained only few FAME profiles. Nevertheless, a nice advantage of this technique is its rapid computation time. For example, the *Bacillus* data set was handled in less than ten minutes. Results comparable to RFs were achieved by the BSVM approach in which the class boundaries are calculated in one single optimization (Crammer and Singer, 2001). The optimization as aimed at by this approach is, however, quite hard and, when considering the parameter optimization, the computation time was consequently very long (from several hours for the *Bacillus* data set to more than one month for straight species ID). ANNs showed a performance somewhere in between that of the RF and SVM technique. Training times were, however, longer than those of RFs. Because of these properties together with the large number of parameters to optimize (activation functions, number hidden neurons, training algorithms, etc.), this technique was not quite attractive for classification in the presented setting. Overall, we can conclude that for future expansion of this research, the random forests technique is most preferable for FAME-based bacterial species classification and identification.

4.3.5 Comparison with Sherlock MIS

Routine identification of bacterial species based on FAME analysis is traditionally performed by the commercial system Sherlock MIS (MIDI Inc., USA). Even though Sherlock MIS is the standard technology for routine FAME-based bacterial identification, the commercially exploited identification system has one main disadvantage when aiming at genus-wide bacterial species identification. As an example, in March 2008, the TSBA50 library entries covered only 30 of the 142 validly published *Bacillus* species (21%), 18 of the 86 validly published *Paenibacillus* species (21%) and 31 of the 112 validly published *Pseudomonas* species (26%). An overview is given in Table B.2. By making use of the LMG FAME database and machine learning techniques, we were able to partially fill this gap and to respond to the dynamic character of taxonomy by rapidly creating new data sets and training new up-to-date identification models. Given the fact that Sherlock MIS is a pioneer in FAME analysis, this system was a good benchmark to compare the power of both identification systems. However, reliable benchmarking is only possible by taking only those *Bacillus*, *Paenibacillus* and *Pseudomonas* species into consideration that were present in both the FAME data sets and the TSBA50 identification library. In the present study, a FAME profile was identified in a stratified setting. In other words, identification was performed first at genus level and, subsequently, at species level by one of the species identification models following successful genus identification. For example, a *Bacillus* profile was not further taken into account when identified as *Paenibacillus* in the genera identification model. For each identification model and from the ten random RF experiment repeats, the RF experiment resulting in the highest AUC value was chosen and the corresponding test set was evaluated. Remark that the test set of the RF experiment for genus identification was not evaluated but that the corresponding RF model was used for genus identification of the *Bacillus*, *Paenibacillus* and *Pseudomonas* test profiles. The indices of the experiment repeat with the corresponding parameter settings (number trees and number of split variables) were for the genera classification model experiment 1 with 1000 trees and 5 variables, for the *Bacillus* species classification experiment 2 with 1750 trees and 8 variables, for the *Paenibacillus* species classification experiment 2 with 1000 trees and 6 features, and for the *Pseudomonas* species classification experiment 5 with 1250 trees and 4 features. Corresponding to the identification type (genus or species), identification was evaluated by considering only the genus name or species name associated with the highest identification output value. For Sherlock MIS, this corresponded to the species with the highest SI (similarity index) value. Two approaches were used for comparing both identifications systems. In the first approach, for each genus, the number of correct identifications was averaged for each considered species, as these were possibly represented by a different number of FAME profiles. Next, a global average and standard deviation was calculated over all species. In a second approach, one test instance for each considered species was randomly sampled and the percentage of correct identifications was calculated. This procedure was repeated ten-fold and a final average and standard deviation was calculated. This second approach was alternatively chosen to also prevent the calculation of an estimate that could be biased because of an imbalanced test set. In the first approach, this problem was dealt with by averaging the identification results of the test samples of each

species.

Identification at genus level was nearly perfect. Three *Bacillus* profiles were rejected due to identification as *Paenibacillus* and one *Paenibacillus* profile was rejected due to identification as *Bacillus*. Figure 4.6 shows a comparison between the *Bacillus*, *Paenibacillus* and *Pseudomonas* species identification accuracy obtained with the RF technique in the stratified strategy setup and with the TSBA50 identification library of Sherlock MIS. It is important to underscore again that model construction was only based on those species present in both the sampled data sets and in the Sherlock MIS identification library. Both evaluation approaches resulted in a similar performance for the three genera and a distinct gap was observed between the RF identification performance and the Sherlock MIS performance for each of the genera. By the first evaluation approach, the RF method correctly identified on average, with standard deviations in brackets, 78.28% (31.37) of the *Bacillus* species, 94.49% (10.70) of the *Paenibacillus* species and 75.65% (28.19) of the *Pseudomonas* species. This in contrast to the Sherlock MIS which correctly identified only 55.77% (41.20), 51.22% (43.40) and 27.00% (33.69) of the species of the respective genera. These averages are visualized in Figure 4.6. Similar values were found by the second approach, in which the RF technique correctly identified 81.78% (6.17) of the *Bacillus* species, 96.43% (3.76) of the *Paenibacillus* species and 74.84% (5.65) of the *Pseudomonas* species. This in contrast to Sherlock MIS that correctly identifies 55.67% (6.52), 54.29% (7.68) and 27.42% (5.10) of the species of the respective genera. The standard deviations distinctly differ between both approaches, due to a different approach of averaging (over species versus over test sets, respectively). The standard deviations in the first approach were quite high due to the fact that species are either very well identified or just not identified and because values are bounded in the percentage interval [0,100]. In other words, we were confronted with a very skewed distribution. The averages showed that the number of well identified species was in favor when compared to the bad identifications. Importantly, these standard deviations should not be interpreted under the assumption of a normal distribution as the distribution of the results was quite skewed. One possibility to interpret this skewedness is to calculate the 25% and 75% percentiles of the results. These percentiles are also visualized in Figure 4.6. It is clear that in the first approach, in which an average identification is calculated over all species, the RF approach was more consistent in its identification. This in contrast to the Sherlock MIS system, which for the species of the genera *Bacillus* and *Paenibacillus* either identified a species or failed to identify the species. For the species of the genus *Pseudomonas*, Sherlock MIS clearly fails to attain high identification percentages. In the case of identification by a RF model, the differences of the 25% percentile with the average are smaller than those between the 75% percentile and the average. Though the 25% percentile was high, this implied that an important number of identifications resulted in a score near zero. A final important remark is that in the second approach, profiles of minor classes had a higher probability of being reselected than profiles of major classes.

It is immediately clear that the machine learning approach outperformed the commercial identification system for species identification for the three bacterial genera considered. The main reason for the resulting gap can be found in the different approach of identification. Sherlock MIS calculated correlation values between unknown FAME profiles and the TSBA50 iden-

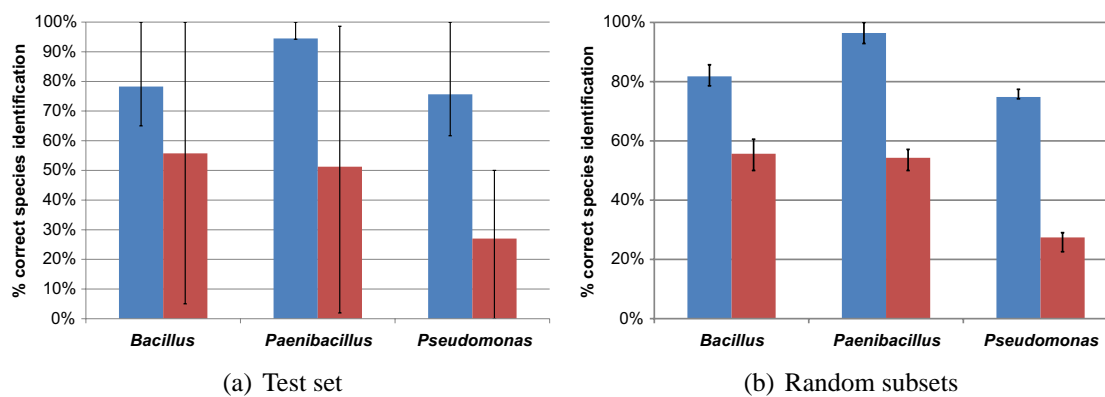


Figure 4.6: Comparison of the identification performance of random forests and Sherlock MIS in a stratified setting. Blue bars correspond to random forests and red bars with Sherlock MIS. In the left diagram, evaluation is performed by the test set with averaging of the identification results over each species and globally over all species. In the right diagram, evaluation is performed by randomly sampling ten subsets consisting of one profile for each species from the test set and by calculating an average performance over each subset. In this setting, the performance of each subset is calculated by the percentage of correct identifications. The bottom and upper bar respectively denote the 25% and 75% percentile of the identification results.

tification library entries based on the Mahalanobis distance where, in contrast to MIDI, machine learning techniques took advantage of learning from the data. Based on the knowledge inside the different data classes, machine learning techniques learnt to distinguish the different classes from one another. Next, probability values were given to an unknown FAME profile of belonging to each class. It is clear that machine learning really took advantage of learning from the data in contrast to the naïve Sherlock MIS approach of comparing each FAME profile with each library entry. Besides this, it is important to remark that Sherlock MIS included significantly more genera in its identification libraries, making the identification potentially more prone to wrong identification results. Consequently, comparison with Sherlock MIS will become more reliable when more genera are implemented in the proposed identification scheme.

4.3.6 Independent Test Sets

In the construction of identification models, the most challenging part of research is the identification of independent test sets. As in this study the number of FAME profiles and the number of strains was restricted for each species, other researchers contributed to this study by sharing private data sets. Important herein, the growth and culture conditions of the considered bacterial strains needed to be identical to those of the data sets for model construction.

4.3.6.1 *Bacillus simplex*

A first independent test set was obtained by dr. Johannes Sikorski (DSMZ, Germany) containing 131 bacterial FAME profiles from bacteria isolated from soil samples of the Evolution Canyons I and II in Israel. 16S rRNA analysis of all profiles showed very high similarities to the strain *Bacillus simplex* LMG 21002. For more information we refer to the papers of Sikorski and Nevo (2005, 2007). FAME profiles resulting from standard growth and culture conditions

(28°C, 24h, TSA medium) were identified by the best random forests model in a stratified setting.

The shared FAME profiles were constructed by the TSBA40 peak naming table, which contains one major difference with respect to the TSBA50 peak naming table: the disposal of fatty acid C_{15:0}. MIDI considered to treat this straight chain fatty acid as a zero feature in the TSBA50 peak naming table, in order to avoid artificial variance in the fatty acid profiles that results in poor similarity index calculations. Note that the chromatographic peak 15:0 is still identified by the peak naming method TSBA50, but it is no longer taken into account during calculation of the relative amounts of the fatty acid compounds. This decision is supported by MIDI following work with coryneforms and related organisms, which often produce unknown peaks located in the 15:0 naming window and the fact that acid-fast organisms often produce fragments that also fall within the 15:0 naming window, although they are not related to this fatty acid compound (Dawyndt, 2004). As the Sikorski FAME profiles are resulting from the TSBA40 peak naming table and our identification models rely on the peaknaming of the TSBA50 method, the C_{15:0} fatty acid was removed from the profiles. For the respective experiment, also the peak C_{14:0} 2OH was not present in the training set. The C_{15:0} peak corresponded with only 1% relative peak area and the C_{14:0} 2OH was only present in two profiles with values of approximately 0.1%. A recalculation of the relative peak areas was performed by equally distributing the totally removed peak area percentage over the different other peaks. Finally, the rescaled profiles were identified by the RF model. Results are reported in Table 4.6. A distinct difference was found between both systems, with a clear advantage for the machine learning approach. Note that, as compared to just deleting the respective peaks, rescaling of the data had no effect when only *Bacillus* species identification was considered, as the resulting performance with both approaches resulted in a 100% correct identification. Also important to remark is that the scores for the identification of the genus *Paenibacillus* were very close to those of the genus *Bacillus*. Implementing a more flexible system instead of only considering the highest output scores would place the identification results in a better context. For example, considering all identifications within a certain interval from the highest score.

Random forests		Sherlock MIS	
107/131(81.68%)	<i>Bacillus simplex</i>	90/131(68.70%)	<i>Bacillus-megaterium</i> -GC subgroup B
16/131 (12.21%)	<i>Paenibacillus validus</i>	19/131(14.50%)	<i>Paenibacillus-gordoniae</i> *
3/131 (2.29%)	<i>Paenibacillus pabuli</i>	12/131 (9.16%)	<i>Brevibacillus-brevis</i> *
	<i>Paenibacillus taiwanensis</i>	5/131 (3.82%)	<i>Bacillus-simplex</i>**
2/131 (1.53%)	<i>Paenibacillus chitinolyticus</i>	2/131 (1.53%)	<i>Bacillus-megaterium</i> -GC subgroup A
		1/131 (0.76%)	<i>Brevibacillus-centrosporus</i> **
			<i>Bacillus-chitinosporus</i>
			<i>Sporosarcina-ureae</i>

Table 4.6: Identification of the J. Sikorski independent FAME data set. The identification results of the best random forests experiment and of Sherlock MIS are given for 131 *Bacillus simplex* FAME profiles. The percentage of *B. simplex* identification is marked in bold.

4.3.6.2 Milk Data Set

Stratified identification was also performed for 38 FAME profiles originating from different *Pseudomonas* strains isolated from milk samples (personal communication with An Coorevits, Laboratory of Microbiology, Ghent University). Sequence analysis was based on the *rpoB* gene, as far as this single gene shows a good resolution for species delineation and resolves into the correct species name. All profiles were identified as the genus *Pseudomonas*, while only one profile was correctly identified at species level (as *P. aeruginosa*). This again underscores the difficult identification of *Pseudomonas* species. However, when considering the *Pseudomonas* groups as described by Anzai et al. (2000) and the maximum likelihood 16S rRNA gene tree (see Figure 2.8), 22 of the 38 profiles (57.90%) corresponded to the correct species group or were a close relative of the species group. This data set was mainly concerned with *P. fluorescens* group species. Sherlock MIS only placed two profiles into the correct group and mainly identified the profiles as the species *P. putida*.

4.3.6.3 Double-blind Study

Since several years, a double-blind study has been performed regarding the taxonomic position of several *Pseudomonas* species (personal communication with Paul De Vos, Laboratory of Microbiology, Ghent University). For identification purposes in this study, 40 species were considered, covered by 75 (synonymous) type strains and deposited as a *Pseudomonas* species. One FAME profile per strain was obtained from two independent laboratories. Recalculation of the peak areas of most profiles was also necessary due to the use of the Sherlock MIS TSBA40 peak naming table (or, presence of fatty acid C_{15:0}) and to the presence of several peaks that were not covered during the construction of the machine learning model. Identifications were compared to the species name as deposited in a culture collection. Stratified identification by the best RF experiment was performed. At genus level, all profiles of both laboratories were correctly identified as belonging to the genus *Pseudomonas*. At species level, one laboratory obtained a correct *Pseudomonas* species identification percentage of 61%, 69% and 75%, when considering the highest, the two highest and the three highest output scores (without averaging of results within and over species). Even though *Pseudomonas* species identification showed to be a hard job when applying only FAME data, these results indicated that whole-cell FAME analysis has a potential for species identification. The results of the second laboratory were lower: 41%, 60% and 65%. When comparing the RF *Pseudomonas* species identifications of both laboratories, a consistent identification was found of 53%, 84% and 95% when focusing again on the highest, the two highest, and three highest output scores. It can be concluded that a certain consistency in the independently generated FAME profiles was found. Using Sherlock MIS, the first laboratory correctly identified approximately 53% of the species, while the second laboratory identified 43% of the species correctly. In this latter comparison, only the highest SI value was considered.

4.3.7 The Plant-pathogenic *Pseudomonas* Species

In the description of the 2008 data set, we stated that the genus *Pseudomonas* comprises different plant-pathogenic strains. Several FAME studies on this topic have already been performed, such as the work of Stead (1992); Stead et al. (1992). Computational analysis of this plant-pathogenic group, however, remained untouched. Therefore, we investigated how well machine learning techniques could distinguish between plant-pathogenic species based on FAME data. In a second experiment, we evaluated how well the FAME data of the group of plant-pathogenic species could be distinguished from the FAME data of non-plant-pathogenic species. Herein, a *Pseudomonas* species was considered plant-pathogenic when at least one of its strains is known as a plant or mushroom pathogen. All respective species (25) are reported in Table 3.2 and are denoted in Table A.2 by the superscript ‘p’.

4.3.7.1 The Group of Plant-pathogenic Species

In this experiment, only the machine learning techniques RFs and SVMs were considered. For SVMs, both the linear and RBF kernel were considered. Optimization of the variables was done by a gridsearch. For RFs, this corresponded with a forest size interval of [1000,4000] in steps of 1000 and a split variable interval [1,#features] in steps of 5. For SVMs, the intervals described in Subsection 4.3.1 were used. Nested cross-validation was performed with pooling of the test results (Witten and Frank, 2005; Varma and Simon, 2006; Parker et al., 2007) (see also Subsection 1.1.1). As the minimum number of FAME profiles per class was four, a 4-fold outer cross-validation was performed. For SVMs, 3-fold inner cross-validation was performed and, as RFs have a low tendency to overfit, no inner cross-validation was performed for this technique and optimization was done by the test folds. The results of the *Pseudomonas* species identification are reported in Table 4.7. Analysis of the test results in this multi-class setting was also performed in a one-versus others approach as described in Section 1.3. As mentioned in this section, the metrics precision and F-score could become undefined due to a zero denominator, though the number of these undefined metrics was very small.

Technique	AUC	Se	Pr	F
RF	0.975 (0.026)	0.631 (0.297)	0.811 (0.152)	0.711 (0.181)
SVM lin	0.962 (0.042)	0.498 (0.384)	0.758 (0.215)	0.663 (0.244)
SVM RBF	0.964 (0.039)	0.475 (0.387)	0.675 (0.268)	0.625 (0.271)

Table 4.7: Results of *Pseudomonas* species identification by the 2008 *Pseudomonas* plant-pathogenic species data set. Two machine learning techniques were evaluated: random forests (RFs) and support vector machines (SVMs). For support vector machines, the linear kernel (lin) and the RBF kernel were considered. Nested cross-validation was performed, with no inner cross-validation in the case of RFs. The AUC, sensitivity (Se), precision (Pr) and F-score (F) were calculated as an average over all classes in a one-versus-others settings. Standard deviations are also reported.

These results were comparable to those resulting from the machine learning analysis of the complete 2008 *Pseudomonas* data set, in the sense that similar metric values were found. Even though a smaller number of species was investigated, it was not straightforward to distinguish

the plant-pathogenic *Pseudomonas* species from one another. These findings are supported by the principal components analysis described in Subsection 3.3.4 and Figure 3.17. From the corresponding biplot of the first two principal components, it can clearly be seen that the FAME patterns of the different species overlap, making it not easy for machine learning techniques to find good flexible margins. Some species could, however, clearly be distinguished by the machine learning experiments. 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*. Note, however, that it remains to be seen whether this conclusion also holds for plant-pathogenic species of other genera.

4.3.7.2 Plant-pathogenic Species versus Non-plant-pathogenic Species

In this experiment, only RFs and SVMs were considered in the same setting as described in the previous subsection. However, because a larger number of data points per class was available, we chose to perform a 10-fold nested cross-validation. In the case of RFs, no inner cross-validation was performed for this technique and parameter optimization was done on the test folds (see also section above). The results of this two-class identification experiment are reported in Table 4.8. The ROC curve for the RF experiment is visualized in Figure 1.10.

Technique	AUC	Se	Pr	F
RF	0.965	0.961	0.907	0.933
SVM lin	0.897	0.935	0.794	0.859
SVM RBF	0.939	0.920	0.902	0.911

Table 4.8: Results of *Pseudomonas* identification of plant-pathogenic species in the 2008 *Pseudomonas* data set. Two machine learning techniques were evaluated: random forests (RFs) and support vector machines (SVMs). For support vector machines, the linear (lin) and RBF kernel was considered. 10-fold nested cross-validation was performed without an inner cross-validation in the case of RFs. The AUC, sensitivity (Se), precision (Pr) and F-score (F) were calculated from the two-class identification results.

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 were somewhat surprising given the overlapping data clouds visualized by principal component analysis (see Figure 3.18). It is clear that some relation exists between plant-pathogenicity and the fatty acid content. When statistically analyzing the probability estimates output from the RF experiment by a Wilcoxon rank-sum test, a p -value of approximately zero was obtained, implying statistically different probability estimates at the significance level of 0.05. A possible reason for this good performance could be found in the paper of Stead (1992) who describes a clustering of a multitude of plant-pathogenic *Pseudomonas* species by hydroxy fatty acids. Though, in this study *Pseudomonas sensu lato* is considered and major discriminations are discussed with the first major group corresponding 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-OH and C_{12:0} 3-OH. Subgrouping could also be achieved by these hydroxy fatty acids, together with the hydroxy fatty acid C_{12:0} 2-OH. 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 discriminations can also be assumed to be very valuable for discrimination between plant-pathogenic and non-plant-pathogenic species. Future study of this relation should reveal all determining FAME constituents and the corresponding qualitative and quantitative differences.

4.3.8 Publication

The main part of this section is published in the international peer-reviewed journal of Systematic and Applied Microbiology with reference: B. Slabbinck, B. De Baets, P. Dawyndt and P. De Vos (2009). Towards large-scale FAME-based bacterial species identification using machine learning techniques. *Systematic and Applied Microbiology*, 32, 163–176.

4.4 Conclusions

Initial research applying ANNs for FAME-based classification and identification of bacterial genera showed promising results. Bertone et al. (1996) and Giacomini et al. (2004) concluded that it is worthwhile to build a system for FAME-based identification that discriminates bacteria at the genus level. It remained, however, to be seen how classification of bacterial FAME profiles could be used to discriminate between species of a single genus. Kämpfer (1994), Heyndrickx et al. (1996) and Vancanneyt et al. (1996) proved that fatty acid analysis has a potential for species differentiation within the genera *Bacillus*, *Paenibacillus* and *Pseudomonas*.

In a first research part, multiple experimental setups were analyzed to verify the possibilities for genus-wide species identification by combining machine learning techniques and FAME data. Even though FAME analysis is routinely performed in many laboratories, this mass of data has never been subjected to a machine learning strategy for FAME-based species classification. Generally seen, good identification results for *Bacillus* species were obtained. We successfully proved that bacterial species of a single genus could be distinguished based on their FAME content using ANNs, which are a good option for the identification of species in the genus *Bacillus*. From the experimental setups, we concluded that ANN-based identification improves with imbalanced data sets validated with stratified cross-validation. It is expected that bacterial identification will generally improve when more profiles are available for each species due to a larger intra-species heterogeneity. Furthermore, the combination of activation functions used by ANNs should be determined empirically. By determining the best activation functions and optimizing the number of hidden neurons, a good genus-wide FAME-based species identification system could be achieved for the genus *Bacillus*. Proper experimental setups, however, imply that there are still other setup preconditions to be tested, which could possibly contribute to an improved classification performance. Some species are very closely related both genotypically and phenotypically, such as the species within the *B. cereus* group and the *B. subtilis* group. In these cases, species identification should be done as a group identification. Better results were achieved following this strategy. When considering the use of FAME-based identification as first-line identification tool, narrowing the bacterial spectrum could also be achieved more effi-

ciently by analyzing the highest scores given by the identification model. Both cases prove that integrating the resolution of FAME analysis in the identification system will enhance first-line species identification. For this part, it could be concluded that the presented results were highly promising for the classification of *Bacillus* species.

With the 'three genera - three techniques' research, the next step was taken towards a computational genus-wide species identification system based on whole-cell FAME data. FAME-based genus and species identification was evaluated using the machine learning methods ANNs, RFs and SVMs. The three machine learning techniques showed a similar and nearly perfect identification performance at genus level. At species level, experiments and subsequent evaluation demonstrated that RFs is the best technique for species identification for each of the three genera. Besides this, RFs has also several advantages as opposed to ANNs and SVMs such as robustness against overfitting, computation time and optimization of a small number of parameters. Consequently, it is advised to perform further work on various other genera and species by the RF technique in a stratified identification strategy. Considering the limited discriminative power of FAME analysis for species identifications and ongoing discussions regarding the taxonomic positions of several *Bacillus*, *Paenibacillus* and *Pseudomonas* species, a moderate to high identification performance was achieved. For the genus *Pseudomonas* specifically, the resolution of FAME analysis for species discrimination showed also to be very limited. In this case, the integration of this prior knowledge in the identification system should be considered. Comparisons with the identification reports of the commercial Sherlock MIS (MIDI, Inc., USA) showed that the performance of the presented machine learning approach for the identification of *Bacillus*, *Paenibacillus* and *Pseudomonas* species was clearly improved, even though Sherlock MIS has a lot more genera and species included in its identification libraries.

The plant-pathogenic *Pseudomonas* species were analyzed in a separate data set. Both RFs and SVMs resulted in a moderate discrimination between the different species, which could also be interpreted from principal component analysis. However, when considering a setting of plant-pathogenic and non-plant-pathogenic *Pseudomonas* species, a very high discrimination was achieved. And, a clear relation exists between certain fatty acids and plant pathogenesis.

As bacterial taxonomy is rapidly evolving, flexible solutions are required to achieve up-to-date first-line bacterial species identification. We presented a machine learning approach to tackle this problem. Up-to-date and accurate identification are two of the main advantages of this approach as opposed to the Sherlock MIS. Nonetheless, the current approach has some drawbacks. According to the List of Prokaryotic names with Standing in Nomenclature as published in October 2006 the genus *Bacillus* comprised 143 validly published species while in March 2008, *Bacillus*, *Paenibacillus* and *Pseudomonas* comprised 145, 86 and 117 validly published species, respectively (Euzéby, 1997). The 2006 data set extracted from the LMG FAME database contained 82 *Bacillus* species covering 1,071 FAME profiles, while the 2008 data set contained 961, 378 and 1,673 standard FAME profiles of 74 *Bacillus* species, 44 *Paenibacillus* species and 95 *Pseudomonas* species, respectively. Particularly for the *Bacillus* and the *Paenibacillus* data set, only half of the validly published species were included for training of the machine learning techniques. Based on the FAME database alone, we are thus still far away from a complete genus-wide bacterial species identification. No single computational FAME

analysis had been performed on this scale yet. Moreover, knowledge about the heterogeneity of each species is limited by the restricted number of strains and FAME profiles present in our FAME database. These drawbacks are, however, inherent to the rapidly evolving taxonomy as well as to research performed at a single institute, which is restricted to specific ecological, clinical and industrial niches. Thus, in general, one institute can be regarded as ‘data-restricted’. This problem may only be solved by future cooperation between different research institutes performing bacterial FAME analysis under the same standardized conditions. Even though cooperation is not straightforward, it should not be a huge obstacle as the proposed approach would benefit all cooperating parties and would improve bacterial species identification in many microbiology-related fields. The ultimate solution for this problem lies, however, in building a public FAME database. Where gene and genome sequence databases are hugely expanding in number and content, databases of phenotypic data are still far behind. This topic is further discussed in Chapter 6.

It is clear that identification fully relies on the species library and the resulting data sets. Training of the data can be compromised when species are renamed or strains are wrongly assigned. In a continuously changing field such as microbial taxonomy, these errors are inherent to the data set. However, with the power of our identification approach and in contrast to the commercial system, we can rapidly update the back-end library and retrain the classification model in order to obtain an up-to-date identification scheme. Because the advantages of machine learning techniques are fast training and the ability to handle large data sets, future work on FAME-based bacterial species identification by machine learning techniques focusing on the implementation of more genera and species will, therefore, easily be handled. Nevertheless, increasing the number of genera and species will make training of new identification models a harder but challenging computational task, and will lead to more error-prone results. The degree of reduced identification power will not only depend on the number of genera and species described/included in the new system, but also on the intra- and inter-genus/species variation of the additional and new taxa. This encompasses a need for more strains and more FAME profiles as, without a sufficient number of the latter, the calculation of reliable boundaries between different classes remains a challenging task. This again takes us back to the above-mentioned restrictions of one research institute or laboratory. Herein, it is also important to state that the ‘one strain–one taxon’ descriptions do not provide this natural variation and microbiologists should be discouraged to create such new taxa because of their weak phenotypic discrimination. Regarding the expansion of the presented identification system towards more genera and species, two more remarks should be emphasized. First, including more genera and species will make the comparison with the identifications by the Sherlock MIS more reliable and objective. Second, the expansion of the identification system is, however, limited as most bacteria do not grow under the same standardized growth conditions or are even unculturable. Finally, when further expanding the stratified identification system, future work will also need to integrate an alternative scoring and/or weighing mechanism to obtain reliable species identification as this fully relies on the power of genus identification. Summarized, by cooperation and extending this research in the future, the automated FAME-based identification tool for bacteria will become most valuable in microbiology and many related fields.

CHAPTER 5

Phylogenetic Learning

*As I say it's a bit dingy at present but it's
surprising what a lick of paint'll do isn't it?*

WALLACE AND GROMIT

5.1 Introduction

For the three genera considered in this work, different numerical FAME studies have already underscored that FAME profiling cannot be used to discriminate all valid species from each other (Heyndrickx et al., 1996; Kämpfer, 1994; Stead, 1992; Vancanneyt et al., 1996). With the data analysis performed in Chapter 3, we extended the scope of these studies and showed that the FAME profiles of the considered bacterial species are indeed highly similar, making them hard to distinguish. Where FAME-based bacterial species identification is typically performed by comparing FAME profiles against identification libraries with fixed peak percentages, we demonstrated in the previous chapter that machine learning techniques are able to maximally exploit the pattern information in the FAME data. By learning mathematical functions to delineate the different classes or species, an improved species identification was achieved.

In bacterial taxonomy, strains are classified at the taxonomic level of species according to their relatedness in genotypic data. At present, the 70% DNA-DNA hybridization (DDH) threshold is considered the gold standard for circumscribing the taxonomic rank of species. Interestingly, whole-genome sequence analysis revealed a correlation between a DDH value of 70% and a 95% average nucleotide identity (Konstantinidis et al., 2006a; Konstantinidis and Tiedje, 2007). But, even though DNA reassociation is the gold standard and genome studies flourish, 16S rRNA gene sequence analysis is still widely preferred for species delineation for two important reasons: 16S rRNA gene sequence identity greater than 97% may indicate a specific species and sequencing the 16S gene is much cheaper and faster due to the massive technological improvements. As a consequence of this explosive trend, the nucleotide sequence databases of the International Nucleotide Sequence Database Collaboration (INSDC) have known an exponential growth. Nonetheless, sequence analysis and phylogenetic reconstruction studies should rely on high quality sequences. With the exponential growth of the sequence databases, the number of poor quality sequences also grows extensively and sequence curation becomes indispensable. To circumvent manual curation, the SILVA database project allows users to retrieve quality controlled and aligned rRNA sequences as stored in the EMBL sequence database (Pruesse et al., 2007). Since the species resolution of 16S rRNA gene sequence analysis is moderately high to high and that of FAME profiling only moderate, this data

type and the resulting phylogenetic trees can be perfectly used for knowledge integration in species classification models based on FAME data.

As an alternative to flat multi-class classification as handled in the previous chapter, different tree-based approaches for multi-class classification were suggested in literature. Many studies handled multi-furcating trees, mostly for multi-label classification. More information on this topic is given in Subsection 1.1.2.2. Importantly, most of these studies did not involve hierarchical classification for single-label multi-class classification, meaning that each instance is classified at leaf level. From another perspective, hierarchical classification has also been proposed for standard multi-class classification tasks. In this setting, the idea consists of improving multi-class classification methods by constructing a tree of binary classifiers (Lee and Oh, 2003; Cheong et al., 2004; Fei and Liu, 2006). The tree architecture is inferred from the considered data based on a particular algorithm that calculates distance measures or similarities between the considered classes. We exploited and studied this approach for FAME-based species classification.

This chapter focuses on the integration of taxonomic and phylogenetic knowledge into species classification models, with the goal of evaluating the resolution of FAME data. In the first section, two different strategies for the integration of taxonomic and phylogenetic knowledge were investigated, using Random Forests (RFs) as base classifiers. In the first strategy, we considered the integration of relationships between species solely based on FAME data. Herein, a FAME tree was constructed and evaluated for hierarchical multi-class classification. In the second setting, we considered knowledge integration from the perspective of bacterial phylogeny. Using 16S rRNA gene sequence analysis, phylogenetic trees were constructed and, subsequently, used for hierarchical single-label multi-class classification (based on FAME data). This last strategy is further referred to as ‘phylogenetic learning’, an approach that utilizes two types of data: 16S data was considered to incorporate phylogenetic knowledge in the form of a hierarchy or tree and, on each node of this tree, a binary classifiers was constructed by means of FAME data. No hierarchical single-label multi-class classification on phylogenetic information in microbial taxonomy has been investigated so far. In the second section, the phylogenetic learning approach was evaluated as a first step towards a further post-processing. An initial approach was investigated to put the identification results in a proper context. More specifically, a highlighting system was evaluated for easy visualization of those species and species groups that were hard to distinguish from each other based on FAME data.

5.2 From Learning Taxonomies to Phylogenetic Learning

5.2.1 Methodologies

5.2.1.1 Machine Learning

In this chapter, we only focused on the machine learning technique Random Forests (RFs). In contrast to the study performed in the previous chapter, a different parameter approach was used. A grid search was performed to optimize the number of trees and the number of split

variables. All numbers of features were considered for split variable selection and 1000 to 4000 trees in steps of 250 trees were selected for tuning the number of trees. Optimization of the parameters was performed by the error on the test set.

With our FAME data set, two problems arised: classes were imbalanced, meaning that a different number of samples was present in each class, and many classes contained only a small number of samples. To tackle a possible imbalance effect on the classifier performance, the true error rate could be estimated by stratifying train and test sets (Kohavi, 1995). For the second case, classification could also become problematic when two-class classifiers were created based on small data sets. This could be solved by performing cross-validation for performance estimation (Witten and Frank, 2005). This in contrast to the experiments in the previous section (see 4.3.1), where only one separate test set was used for performance estimation. A three-fold stratified cross-validation was performed for both the hierarchical classification and flat multi-class classification. To prevent overfitting, the number of folds was set equal to the minimum number of profiles over all bacterial species, which was three in this study. In this perspective, the stratification proportion corresponded to one-third. Given the identical nature of the probability estimates resulting from each RF model, we chose to aggregate all test sets in a joint test set for performance evaluation. This method is also better known as pooling (Parker et al., 2007). Finally, for the pooled test set, the same approach was followed for evaluation of classifier performance, as described in Section 1.3. In view of this cross-validation for performance estimation, RF parameter optimization within each fold was done by the corresponding test fold, similar to the procedure as described in section 4.3.1.

Besides the calculation of global performance measures, the performance at class level between flat multi-class classification and phylogenetic learning was also compared. The comparison is visualized in a bar diagram. Initially, flat multi-class classification and the corresponding classification results of each class were considered. A threshold was set on a metric using steps of 0.01. As metric, sensitivity and F-score were further analyzed. The corresponding thresholds are plotted along the X axis. For each threshold, those classes were selected corresponding to sensitivity or F-score values smaller than or equal to the threshold. Secondly, for each threshold and, thus, for each selected set of classes, the corresponding metric values obtained by phylogenetic learning were evaluated. The number of phylogenetic learning metric values that were larger than the corresponding metric values resulting from flat multi-class classification are plotted against the Y axis on the left. Also, this number is expressed as a percentage of the corresponding class set size. The corresponding percentages are plotted against the Y axis on the right (an example is given in Figure 5.7).

5.2.1.2 Learning Taxonomies

As our goal is to integrate taxonomic knowledge into the bacterial species classifiers, the construction of a FAME tree by supervised divisive clustering was first considered. A divisive clustering algorithm builds a top-down cluster hierarchy, also called dendrogram, by each time splitting a cluster in two and starting from the entire data set. In unsupervised clustering, distances are calculated between all points in the respective data set. In contrast, supervised

divisive clustering considers the class labels of the respective data points and calculates only distances between those data points with differing class labels. Consequently, the final number of clusters in supervised clustering equals the number of classes present in the original data set (Duda et al., 2001; Mirkin, 2005). Thus, the construction of a FAME tree by supervised divisive clustering was considered. Popular hierarchical clustering strategies are single linkage, complete linkage, average linkage, Ward linkage, etc. In these strategies, multiple metrics can be applied as distance measure. In this study, the identification performance of the respective classifiers served as distance metric. This implies not to extract class distances in input space but rather from output space. Initially, the FAME data set was randomly split in a training and test set. We chose to use the area under the ROC curve (AUC) as calculated from the test set as a splitting criterion. In case of ties, the splitting was refined by accounting for the average linkage of the probability estimates of both classes, as calculated by the Euclidean distance. One should prefer the classifier corresponding to the largest average distance between the probability estimates.

To this end, for each level in the considered top-down setting, a RF classifier was built for all possible two-group combinations of all considered species or classes. For each classifier, the initial train and test set were used to select these profiles of the species considered in each combination. The profiles present in the training set were used for training the RF classifier, while the profiles present in the test set were used for performance estimation. Thus, all combinations correspond to a two-class classification task. For each node or level, this results in $2^{K-1} - 1$ combinations, with K the number of classes considered. Note that, when considering four classes, the combination of classes 1 and 2 automatically excludes the combination of classes 3 and 4. The divisive clustering stops when only two-class clusters are retained. To speed up the divisive clustering and classification process, no grid search and no cross-validation were considered. Or thus, parameter optimization and performance estimation were done on the sampled test subset. The forest size range was identical to the interval described above and was optimized using the default number of split variables ($z = \sqrt{D}$, with D the number of features). Based on the forest size corresponding to the lowest error rate, parameters were optimized by choosing the number of split variables equal to $\frac{D}{2}$, $2D$ and \sqrt{D} . Ultimately, a rooted tree was constructed with equal branch lengths and the different nodes were labeled with the corresponding AUC value. The resulting tree was visualized with the treeing method of the TaxonGap software (see Section 3.3.3).

As an initial proof-of-concept, 15 *Bacillus* species were selected from the original data set. Selection was based on classes with reasonable sample size and classes that are taxonomically closely related to each other, e.g. species of the *Bacillus cereus* and *Bacillus subtilis* groups. The first selection criterion was chosen to avoid heavily imbalanced data subsets. The following species with respective number of data points were selected: *Bacillus aquimaris* (12), *Bacillus atrophaeus*^s (21), *Bacillus cereus*^c (62), *Bacillus coagulans* (32), *Bacillus drementensis* (38), *Bacillus fumarioli* (28), *Bacillus galactosidilyticus* (12), *Bacillus licheniformis*^s (74), *Bacillus megaterium* (28), *Bacillus mycoides*^c (11), *Bacillus patagoniensis* (12), *Bacillus pumilus*^s (57), *Bacillus sporothermodurans* (17), *Bacillus subtilis*^s (64) and *Bacillus thuringiensis*^c (12). Species annotated with superscript ‘c’ belong to the *Bacillus cereus* group, while species an-

notated with superscript 's' belong to a species of the *Bacillus subtilis* group (Euzéby, 1997; Hansen et al., 2001; Hutsebaut et al., 2006). It was expected that the species of these two groups cluster together.

To further speed up the clustering process, computations were performed in parallel on an Intel Blade cluster (Intel Corporation, Santa Clara, CA, USA).

5.2.1.3 Phylogenetic Analysis

5.2.1.3.1 16S rRNA Gene Sequence Trees

Because the resolution of 16S rRNA gene sequence analysis provides moderate to high species delineation for most bacterial species, this type of data is an ideal source for knowledge integration. The SILVA database was used for 16S rRNA gene sequence selection. This database subjects EMBL 16S rRNA gene sequences to a multiple alignment, different control procedures and annotates the corresponding sequences with quality scores. In SILVA, quality is denoted three-fold: pintail quality for sequence anomaly detection, sequence quality and alignment quality (Pruesse et al., 2007). For each type strain of each species present in our data set, one 16S rRNA gene sequence was selected from version 95 of the SILVA database. If multiple 16S rRNA gene sequences for each type strain were available, selection of the final sequence was based on best quality and longest sequence length. Remember that the type strain of a bacterial species is the fixed name bearer of the species and its phylogenetic position is, hence, determinative in the taxonomic framework. A list of the selected accession numbers can be found in Appendix C.

Sequence distance calculation was performed by the PHYLogeny Inference Package version 3.68 (PHYLP), using the program Dnadist. The Jukes Cantor evolution model was used for correcting the nucleotide distances (Felsenstein, 1989, 2004; Stackebrandt and Swiderski, 2002). This DNA sequence evolution model assumes an equal and independent change rate for each nucleotide. So, substitution of one nucleotide by one of the three other nucleotides occurs with equal probability. All other parameters were used as default except for sequential input sequences. Based on the resulting distance matrix, an NJ and a UPGMA tree were created using the PHYLP program Neighbor (Sokal and Michener, 1958; Saitou and Nei, 1987; Dawyndt et al., 2006). Default parameter settings were used, except for a randomized input order of the species. Phylogenetic trees were created for the species present in both the 2008 data set and in the 15 species data set. All trees were visualized using the iTol webtool version 1.5 (Letunic and Bork, 2007).

5.2.1.3.2 Tree Inference

In Chapter 2, phylogeny is defined as the evolutionary relationships between organisms as deduced from the genetic information in nucleic acids and proteins. These relationships can mathematically be modelled and visualized in a graph, called a phylogenetic tree. A phylogenetic tree is composed of nodes and branches (or edges), where branches only connect two adjacent nodes and define the relationships among the nodes in terms of descent and ancestry.

The terminal or external nodes are called the leaves. In taxonomy, one calls these operational taxonomic units (OTUs). The most basic tree is the cladogram, which simply show relative measures of common ancestry. Additive trees associate with each branch a particular branch length, that represents an amount of evolutionary change. Hence, the distance between two OTUs equals the sum of the branch lengths connecting them. Ultrametric trees or dendrograms are additive trees where the tips of the trees are equidistant from the root. With this tree it is possible to depict evolutionary time. Cladograms and additive trees can either be rooted or unrooted. A root is the common ancestor of all integrated OTUs and defines the order of descent by evolutionary time. Hence, unrooted trees only specify evolutionary relationships (but no path). Unrooted trees can become rooted by different rooting procedures. A root can be defined on a particular branch but is commonly created by using an particular outgroup that is distant from all considered OTUs. In this chapter, only rooted trees with bifurcating nodes were considered. This implies that each interior node is connected to three others and every leaf is connected to only one other node. Or, from mathematical perspective, a rooted phylogenetic tree can be regarded as a directed acyclic graph with bifurcating nodes (Li, 1997; Page and Holmes, 1998; Felsenstein, 2004).

A tree obtained by a certain data set and a certain tree reconstruction method is called an inferred tree. Many tree inference methods exist which can be distinguished in different ways. A major example is the principle of tree construction. Herein, the two main types are cluster methods and search methods. Cluster methods follow a specific cluster algorithm to arrive at a particular tree. Advantages are easy implementation and fast computation. The result is typically a single tree. These type of methods, however, are limited in analytical evaluation and do not allow for evaluation of competing hypotheses. Examples are the unweighted pair-group method with arithmetic mean (UPGMA) and the neighbour joining (NJ) method. The second class of methods are search methods in which an optimality criteria is used for choosing among a set of all possible trees. This criterium assigns a score to each tree as function of the relationship between the tree and the data. As such, these methods require an explicit function that relates data and tree and allow to evaluate the quality of any tree. Or, they allow to compare how well competing hypotheses of evolutionary relationship fit the data. However, computationally these methods are very expensive and pose the problem of determining a good optimality criterium. Finding the optimal tree(s), mostly requires heuristic methods. Examples are the maximum parsimony (MP) and maximum likelihood (ML) methods. Another difference between the four methods can be made on the basis of how they handle the data. The UPGMA and NJ methods handle pairwise distance matrices, while MP and ML operate directly on the data or on functions derived from the data. Therefore, the former methods are also called distance matrix methods and the latter methods are discrete methods. A major drawback of distance methods is the loss of information when translating, for instance sequences, into a pairwise distance matrix (e.g. loss of information concerning individual sites) (Li, 1997; Page and Holmes, 1998; Felsenstein, 2004). In this chapter, we only focused on clustering methods for tree inference. The simplest method of tree inference is the UPGMA method, which was originally developed for the construction of trees that reflect the phenotypic similarities between OTUs. UPGMA can be used for the construction of phylogenetic trees but one has to bear in

mind that UPGMA assumes a constant evolution rate in all evolutionary lineages (so-called molecular clock). The UPGMA method employs a sequential clustering algorithm, in which relationships are inferred in order of decreasing similarity. Initially, all OTUs are organised in a star-like tree with one interior node. Repeatedly, the two most similar OTUs (shortest distance d) are searched, further considered as a composite OTU and the distance matrix is recalculated, until only two OTUs are left. The branch lengths of the two constituent OTUs of the composite equals half of the distance d . Distances are recalculated by the formula

$$d_{(ij),k} = \left(\frac{n_i}{n_i + n_j} \right) d_{ik} + \left(\frac{n_j}{n_i + n_j} \right) d_{jk}, \quad (5.1)$$

with $OTU_{(ij)}$ the composite OTU constituted of OTU_i and OTU_j , $d_{(ij),k}$ the distance between this composite OTU and OTU_k , and n the number of OTUs in a particular composite OTU (for a single OTU n equals 1). This corresponds to the computation of the arithmetic mean of the pairwise distances between the constituent OTUs of the two composite OTUs or between the constituent OTUs of a composite OTU and another OTU (Li, 1997; Page and Holmes, 1998; Felsenstein, 2004).

In an unrooted bifurcating tree, two OTUs are said to be neighbours if they are connected through a single internal node. In this perspective, the computationally fast NJ method sequentially searches for those neighbours that may minimize the total length of the tree. In fact, NJ is a heuristic method that does not assume a clock but approximates the minimum-evolution tree, the tree with the smallest sum of all branch lengths. As with UPGMA, this method also starts with a star-like tree. The first step is to separate that pair of OTUs from all others that gives the smallest sum of branch lengths. This procedure is continued until all interior branches are found (Li, 1997; Page and Holmes, 1998; Felsenstein, 2004). The algorithm is described in Algorithm 3.

5.2.1.4 Phylogenetic Learning

Based on the 16S rRNA gene phylogenetic trees, a classification scheme was developed following the hierarchical class structure. As such, a rooted phylogenetic tree can be regarded as a directed acyclic graph with bifurcating nodes. The main idea is similar to that of binary tree classifiers (Lee and Oh, 2003; Cheong et al., 2004; Fei and Liu, 2006). However, in contrast to our study, these authors inferred a tree from the data used for classification, while we considered phylogenetic information (the 16S rRNA gene) for tree inference and used FAME data solely for classification. We called this approach ‘phylogenetic learning’. As a simple, naïve approach, at each node of the 16S rRNA gene phylogenetic tree, a two-class RF classifier was trained, based on a subset of the FAME data set. Herein, only the subset of profiles belonging to that part of the tree was considered for training and testing. The two branches of the node defined the two groups of the binary classification task, and at each node, a positive and negative dummy class label was created. As binary tree classifiers rely on rooted trees, we only focused on tree inference by cluster methods as these methods typically result in a single rooted tree. The UPGMA and NJ method were chosen. Unrooted trees, as commonly resulting from the ML and

Algorithm 3 Neighbour-joining algorithm.**Require:** Distance matrix \mathbf{D} of N OTUs

```

1: while  $N > 2$  do
2:   for all OTU $_i$   $i : 1 \rightarrow N$  do
3:      $u_i = \sum_{j:j \neq i}^N \frac{d_{ij}}{(N-2)}$ 
4:   end for
5:   Declare MIN, OTU $_a$  and OTU $_b$ 
6:   for all OTU $_i$   $i : 1 \rightarrow N - 1$  do
7:     for all OTU $_j$   $j : 2 \rightarrow N$  do
8:       if  $i == 1$  AND  $j == 2$  then
9:         MIN =  $d_{ij} - u_i - u_j$ 
10:        OTU $_a$  = OTU $_i$ 
11:        OTU $_b$  = OTU $_j$ 
12:       else if  $d_{ij} - u_i - u_j < \text{MIN}$  then
13:         MIN =  $d_{ij} - u_i - u_j$ 
14:         OTU $_a$  = OTU $_i$ 
15:         OTU $_b$  = OTU $_j$ 
16:       end if
17:     end for
18:   end for
19:   Join OTU $_a$  and OTU $_b$  into composite OTU $_{(ab)}$ 
20:   Branch length from OTU $_a$  to OTU $_{ab}$ :  $v_a = \frac{1}{2}D_{ab} + \frac{1}{2}(u_a - u_b)$ 
21:   Branch length from OTU $_b$  to OTU $_{ab}$ :  $v_b = \frac{1}{2}D_{ab} + \frac{1}{2}(u_b - u_a)$ 
22:   for all other OTU $_k$  do
23:      $D_{(ab),k} = \frac{(D_{ak} + D_{bk} - D_{ab})}{2}$ 
24:   end for
25:   Remove OTU $_a$  and OTU $_b$  from the distance matrix and replace them by the composite OTU $_{(ab)}$ 
26:    $N = N - 1$ 
27: end while
28: Connect remaining OTU $_l$  and OTU $_k$  by branch length  $D_{kl}$ 

```

MP methods, could be used but ask for a preceding rooting procedure.

Given the tree hierarchy, classifiers constructed on terminal nodes and a certain number of parent nodes can become biased due to a small training set size. Herein, terminal nodes are regarded as nodes splitting two leaves. Consequently, splitting data sets in a training and test set was not a good option. Cross-validation overcame this issue by dealing with the whole data set. The classification performance could easily be evaluated as the path of each instance was fixed and each data instance was presented to the classification hierarchy. In the case of an incorrectly classified instance at a specific node in the tree, propagation along the true path stopped and the corresponding data instance was further identified along the predicted path. Therefore, the path and ultimate predicted class of each instance could be determined and a multi-class confusion matrix could be generated for statistical analysis. The same evaluation was done as in the previous chapter (see also Section 1.3).

As an interesting feature, this method offered the possibility to investigate where misclassification mostly occurs along the phylogenetic tree. Hence, the misclassification distance for each class could be estimated by averaging the correct path length of each incorrectly classified instance. This implied that, for each incorrectly classified instance, the correct path length was incremented each time the corresponding classifier resulted in identification of the true branch.

Incrementing continued until the considered instance was incorrectly classified. Note that the node resulting in misclassification also incremented the path length and that the path length was also incremented when ultimate identification occurred in the correct leaf. In this latter case, the correct path length equals the maximal path length. For each class, the average correct path length was plotted against the maximal path length.

5.2.2 Results and Discussion

Within the framework of bacterial taxonomy, an interesting topic for subsequent machine learning research is that of knowledge integration. As FAME data does not allow for a global species discrimination, classification by FAME data should make clear which species are hard to distinguish from each other. This is easily achieved by learning in a hierarchical scheme. Two approaches are straightforward: tree inference by FAME data or by data that do allow for an identification at species level. In this section, we discuss the integration of these two particular types of knowledge into FAME-based bacterial species classification models.

5.2.2.1 Learning Taxonomies

In the first stage of this research topic, we investigated the possibility of reconstructing a small part of the hierarchical phylogenetic structure of the genus *Bacillus* by FAME data and RFs. This genus was chosen because of the profound expertise on this genus present at the Laboratory of Microbiology of Ghent University (Belgium). In this experiment, we focused only on FAME data to integrate taxonomic knowledge. A tree was constructed with a divisive clustering algorithm, in which the classifier performance of RFs, trained on FAME data, was used as splitting criterion. In this top-down approach, all possible splits between classes are initially considered in the root node. Since the data set consisted of a small number of data instances for the majority of species, we preferred divisive clustering over agglomerative clustering. The latter approach builds a tree in a bottom-up clustering procedure, so that, in our setting, clustering at leaf level would be based on the results of unreliable classifiers (due to a small number of class instances for many species). We wanted to avoid this type of instability in the tree construction phase.

An initial proof-of-concept experiment was performed based on a small data set of 15 species or classes, as selected from the original 2008 data set. Only species corresponding to a large number of FAME profiles were selected, with a minimum of 11 profiles. About half of the selected species belong to the two known *Bacillus* species groups, the *Bacillus cereus* and *Bacillus subtilis* group. Hierarchical divisive clustering was started in the root node, for which 16383 RF classifiers were trained and evaluated. In subsequent steps of the clustering algorithm, classifier training became less time-consuming, because the number of trained classifiers decreases exponentially for the remaining subtrees. When the algorithm was finished, in total 18589 classifiers had been trained. The computing time to build and evaluate the complete species hierarchy was 65h 10m 22s.

In this initial experiment, we aimed to evaluate whether a FAME tree constructed with divisive clustering indeed revealed the relations between the species of the different species

groups. Figure 5.1 shows the resulting tree, in which no branch lengths are specified and AUC values of the RF classifiers are given at each internal node. The species representing the *Bacillus cereus* group or the *Bacillus subtilis* group were clearly clustered together under the same parent nodes. Consequently, one can conclude that FAME data allows to discriminate between groups of species. Such a result was not expected because of the large number of combinations and related FAME profiles. However, this experiment clearly showed that RFs took advantage of the relatedness between species and/or groups of species. Selecting both *Bacillus* species groups out of 16383 classifiers showed that machine learning techniques can be employed to distinguish between different species and species groups based on FAME data. Consequently, building a FAME tree using classification techniques as treeing method could be a good base for further knowledge integration. Therefore, we evaluated the tree constructed from the different machine learning models also as a hierarchical classification scheme. Note that this classification task followed the main strategy as reported by (Lee and Oh, 2003; Cheong et al., 2004; Fei and Liu, 2006). Subsequent to the construction of the tree, a RF classifier was retrained at each node of the tree, so that the different classifier parameters were optimized by the approach described in Subsection 5.2.1.2. To this end, we considered both 3-fold and 11-fold stratified cross-validation.

The corresponding results are reported in the upper part of Table 5.1. These results showed that hierarchical single-label multi-class classification with 3-fold stratified cross-validation performed slightly worse than flat multi-class classification (see bottom part of Table 5.1). Performing 11-fold stratified cross-validation, however, resulted in a slightly better performance than flat multi-class classification. In summary, for the 15 species data set, it could be concluded that hierarchical single-label multi-class classification resulted in a performance comparable to that obtained with flat multi-class classification. Nonetheless, we were mainly interested in the classification of the 74 validly published *Bacillus* species present in our data set. Upscaling this experiment from 15 classes to 74 was, however, computationally infeasible, because the number of classifiers to be trained increases exponentially with the number of classes. When considering these 74 classes in our FAME data set, $2^{73} - 1$ classifiers must be trained in the root node. This could not be realized in a reasonable computing time, even when multiple processors are used in parallel. Furthermore, to obtain a good classification performance in the presented experiment, only species were selected for which a reasonable amount of data was available. Nonetheless, in the full data set, a lot of classes were present with a small number of FAME profiles (e.g. three or four profiles) which may result in an unreliable FAME tree. Even though this experiment with 15 species gave promising results, for the reasons above, knowledge integration by divisive clustering of FAME profiles was not further considered in this study.

5.2.2.2 Phylogenetic Learning

An alternative to the construction of a FAME tree is to infer a tree based on data resulting from a technique with a good resolution for species discrimination. In this perspective, the best possibilities are DDH, whole-genome sequence analysis and multi-locus sequence analysis

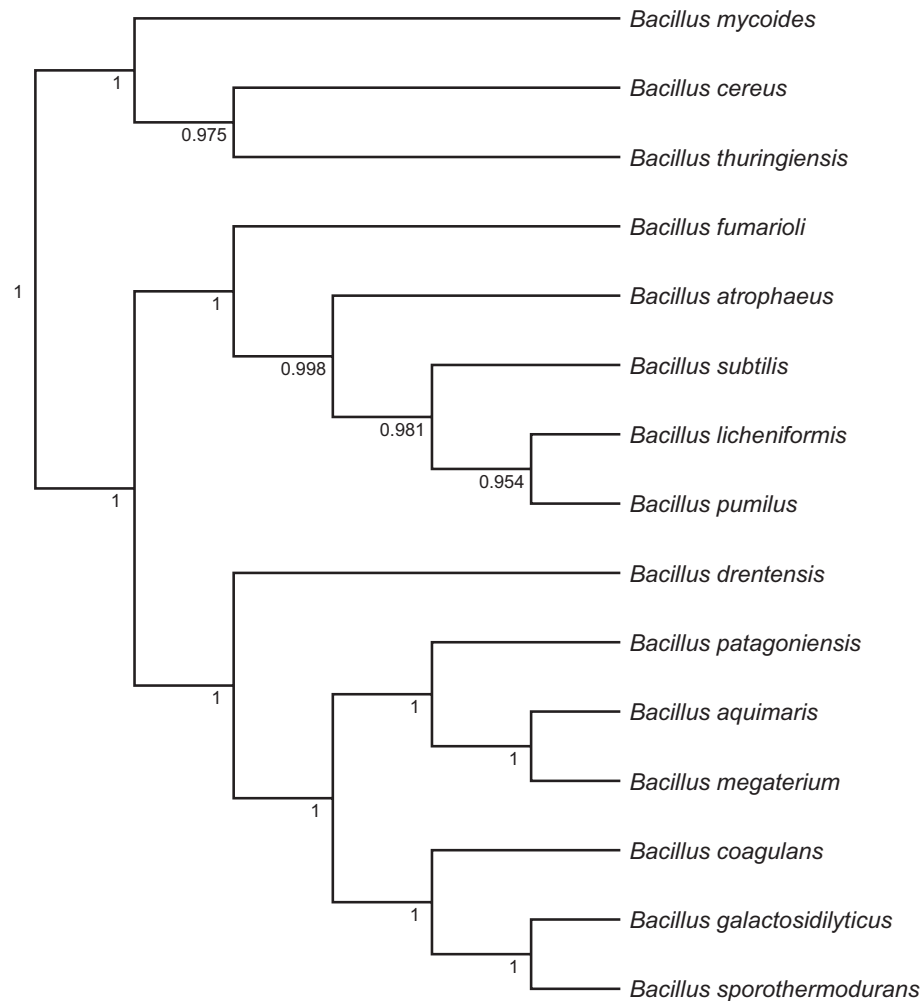


Figure 5.1: FAME tree. Phylogenetic tree resulting from the divisive clustering of the FAME data of 15 *Bacillus* species based on classification by Random Forests. Clustering is based on AUC and average linkage of the probability estimates calculated from identification by Random Forests. At the different nodes the corresponding AUC value is reported.

(MLSA). However, some problems occur with the corresponding data. DDH values are not publicly available and only a restricted number of whole-genome sequences is available for genus-wide studies. Moreover, the construction of multi-gene trees following MLSA leads to the problem of finding discriminating genes on a genus-wide scale. Moreover, finding such gene sequences of high quality for every species is not always straightforward. Therefore, one good other alternative is to focus on 16S rRNA gene sequence analysis. This technique is widely preferred for species delineation because of improved sequencing technology and public sequence databases. Nonetheless, the 16S rRNA gene may not allow for a delineation of every species (Wayne et al., 1987; Stackebrandt and Goebel, 1994; Konstantinidis et al., 2006a; Konstantinidis and Tiedje, 2007). Currently, 16S rRNA gene analysis is one of the techniques generally used in microbiology for phylogenetic analysis.

When using this technique as a starting point for knowledge integration, high quality 16S rRNA gene sequences can be exported from the SILVA database. This database subjects EMBL 16S rRNA gene sequences to different control procedures and annotates the corresponding sequences with quality scores (Pruesse et al., 2007). In this way, one 16S rRNA gene sequence

Classification Results					
	Sensitivity	Precision	NaN	F-score	NaN
HSMC - 15 species					
HSMC	0.887 (0.214)	0.945 (0.059)	0	0.895 (0.179)	0
HSMC (11-fold CV)	0.916 (0.130)	0.956 (0.037)	0	0.930 (0.083)	0
PhyLearn - 15 species					
PhyLearn - NJ	0.992 (0.007)	0.954 (0.041)	0	0.924 (0.099)	0
PhyLearn - UPGMA	0.860 (0.211)	0.931 (0.064)	0	0.873 (0.153)	0
PhyLearn - 74 species					
PhyLearn - NJ	0.741 (0.237)	0.846 (0.181)	1	0.768 (0.181)	1
PhyLearn - UPGMA	0.684 (0.256)	0.860 (0.174)	2	0.741 (0.180)	2
Multi-class					
15 species	0.902 (0.170)	0.944 (0.054)	0	0.911 (0.124)	0
74 species	0.851 (0.189)	0.901 (0.121)	0	0.863 (0.145)	0

Table 5.1: Results from the hierarchical single-label multi-class classification, phylogenetic learning and flat multi-class classification experiments. In this table, the three classification strategies are abbreviated as ‘HSMC’, ‘PhyLearn’ and ‘Multi-class’, respectively. The results of these three strategies are reported in the upper, middle and bottom part of the table, respectively. The results of hierarchical single-label multi-class classification were based on the FAME tree resulting from the divisive clustering experiment. Only the 15 species data set was considered and 3-fold and 11-fold stratified cross-validation (CV) was performed. In the case of phylogenetic learning, two 16S rRNA gene trees were used as template: neighbour-joining (NJ) and unweighted pair group method with arithmetic mean (UPGMA). For PhyLearn, both the 15 and the 74 species data set were considered and all PhyLearn experiments were performed using 3-fold stratified CV. Also the flat multi-class experiments were validated by this CV strategy. In the three strategies, classification performance was evaluated based on the pooled test set. Metrics reported are sensitivity, precision and F-score. Based on a multi-class confusion matrix, statistics were calculated in a one-versus-other setting with averaging of the corresponding statistic over the different classes. Standard deviations are reported between brackets. NaN denotes the number of classes that have resulted in a value ∞ (only in case of precision and F-score).

was selected for each type strain of each *Bacillus* species present in the original FAME data set. Note that the type strain of a bacterial species is the fixed name bearer of the species (according to the bacterial code (Lapage et al., 1992)) and its phylogenetic position is, hence, determinative in the taxonomic framework.

Following sequence selection, distance matrices were calculated using the Jukes-Cantor nucleotide evolution model and two phylogenetic trees were accordingly constructed, respectively with the NJ and the UPGMA method (Sokal and Michener, 1958; Saitou and Nei, 1987; Felsenstein, 1989, 2004; Dawyndt et al., 2006). The respective trees are shown in Figures 5.2 and 5.3. They were used as templates for hierarchical FAME-based species classification by the binary tree approach. As hierarchical classification relies on a phylogenetic tree, we called this approach ‘phylogenetic learning’. As the binary tree classifier is based on a rooted tree structure, we initially chose to select the NJ and UPGMA methods as these basically infer rooted trees. Two methods were selected, to allow for a comparison of binary tree classifiers based on different trees. In view of tree inference, several other methods exist (e.g. MP and ML). The MP and ML methods, however, infer unrooted trees and need consequently to be rooted for use in

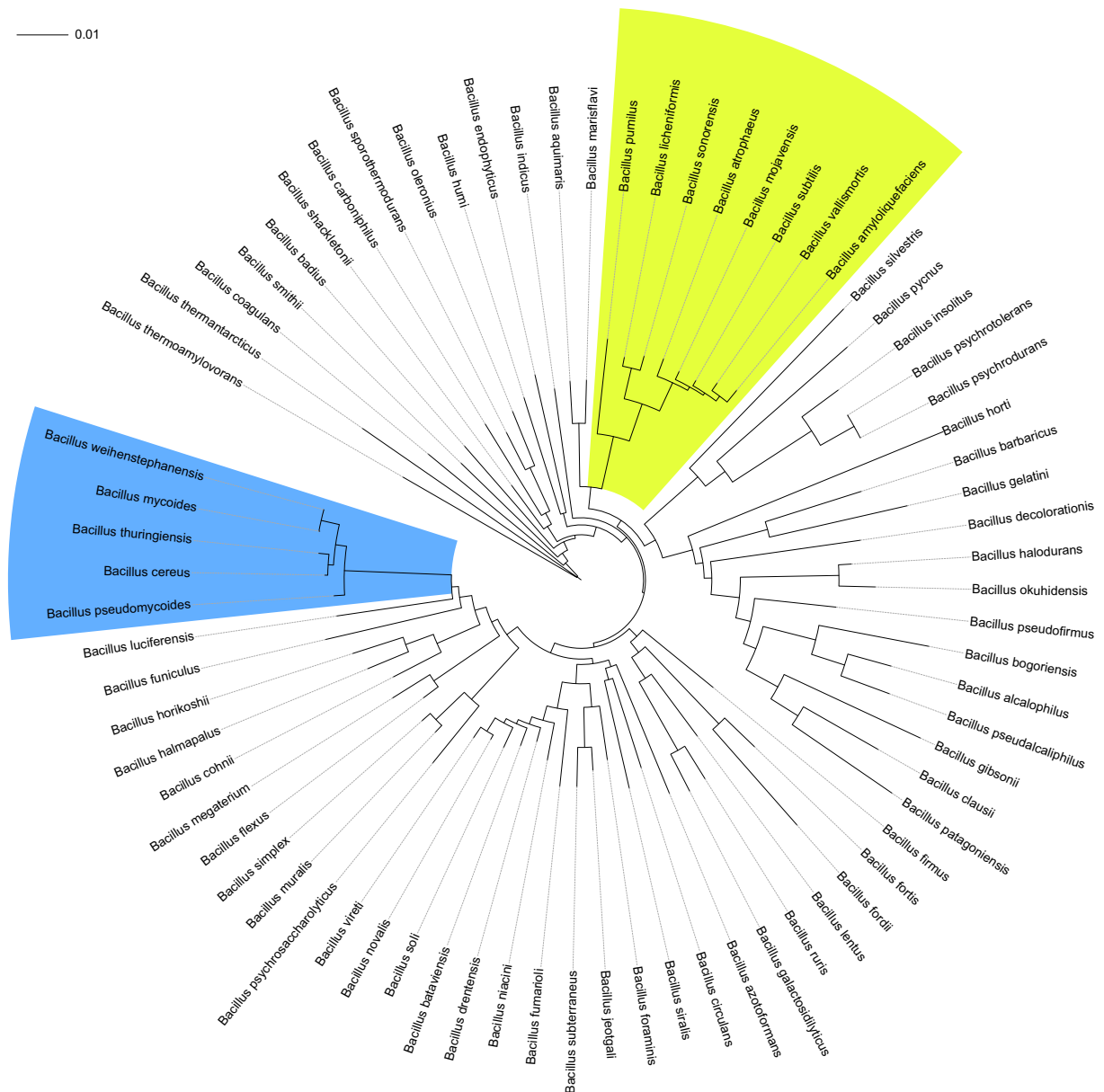


Figure 5.2: *Bacillus* 16S rRNA gene neighbour-joining tree as constructed by PHYLIP 3.68 and based on sequences selected from the SILVA database. Only the species present in the original 2008 data set are visualized. The tree was visualized using the iTol webtool (Letunic and Bork, 2007). The *Bacillus cereus* and *Bacillus subtilis* groups are coloured in blue and green, respectively.

a binary tree classification. Typically, this rooting is achieved by the inclusion of an outgroup (Felsenstein, 2004). Now, the constructed RF classifiers were evaluated for distinguishing between the FAME patterns of the two underlying groups of classes in every node of the tree. The collection of binary classifiers should be regarded as one classifier wrapping the multiple hierarchically structured classifiers. Three-fold stratified cross-validation for error estimation was performed during the training process of each classifier with pooling of the test results of all folds (Kohavi, 1995; Witten and Frank, 2005), i.e. the predictions on test data were pooled together in one big set, and the performance measures were calculated on this set. The results of phylogenetic learning based on the NJ and UPGMA trees are reported in the middle part of Table 5.1.

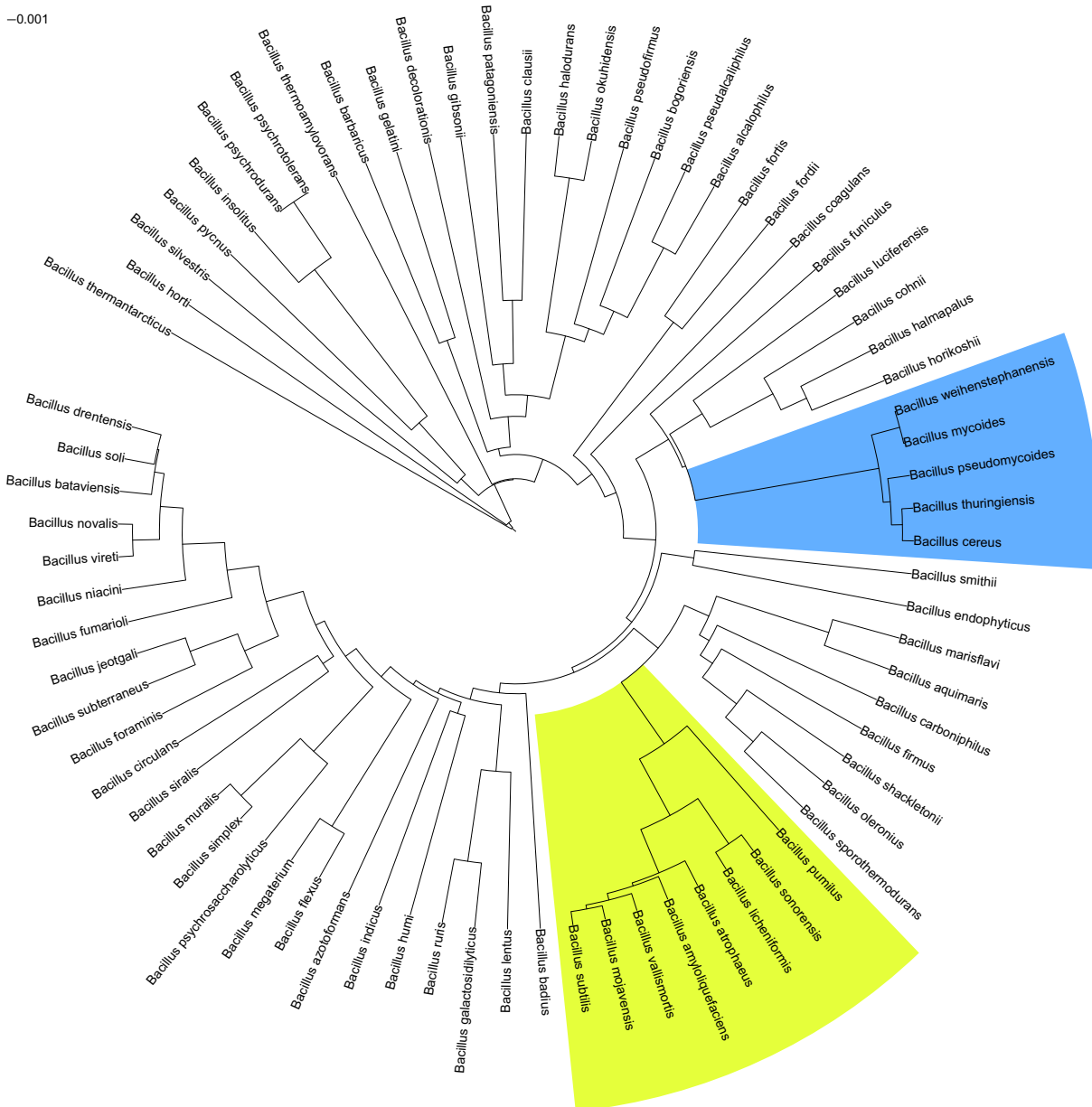


Figure 5.3: *Bacillus* 16S rRNA gene UPGMA tree as constructed by PHYLIP 3.68 and based on sequences selected from the SILVA database. Only the species present in the original 2008 data set are visualized. The tree was visualized using the iTol webtool (Letunic and Bork, 2007). The *Bacillus cereus* and *Bacillus subtilis* groups are coloured in blue and green, respectively.

These results were compared with those obtained from a FAME-based flat multi-class classification (see bottom part of Table 5.1), where only one multi-class classifier was trained by the same cross-validation strategy. First of all, phylogenetic learning and flat multi-class classification were also evaluated for the 15 species data set (as selected in the previous subsection). For the flat multi-class classification, an AUC value was achieved of 0.992 (0.010). The corresponding results are also reported in Table 5.1. Note that the flat multi-class classification with 3-fold stratified cross-validation in this study differed from the flat multi-class classification strategy performed in the previous chapter. In the latter case, 10 repeated experiments were carried out with averaging of the classifier performance on a randomly sampled test set. Average AUC, sensitivity and precision were then given by 0.988, 0.847 and 0.908, respectively

(see Table 4.4). These metric values were approximately identical to the values obtained by phylogenetic learning. As a result, the cross-validation with pooled metric calculation in a flat multi-class setting did not improve classification performance, when compared to the random test set selection.

It is also interesting to see that, even though flat multi-class classification of the 15 species data set resulted in a very high AUC value of 0.992, higher sensitivity and F-score values were obtained by phylogenetic learning on this data set (based on the NJ tree). Conversely, phylogenetic learning based on a UPGMA tree performed slightly worse than flat multi-class classification. When the study was scaled up to 74 species, flat multi-class classification performed better than phylogenetic learning on both trees. For flat multi-class classification, an according AUC value of 0.982 (0.042) was achieved. For the NJ and UPGMA trees, the difference in sensitivity between both techniques and flat multi-class classification was 11% and 16.7%, respectively, while the difference in F-score was, respectively, 9.5% and 12.2%. The contrast between the two data sets was, logically, based on the larger number of relations between the different species and the more complex hierarchical structure of the data. The main reason for the lower prediction performance of phylogenetic learning could be found in the 16S rRNA gene phylogenetic trees that defined the multiple learning tasks. These could become quite hard to solve when classifying the species based on FAME data. Flat multi-class classification is not confronted with these restrictions at all and allows for more flexible solutions. Moreover, in a 74 species hierarchical learning system, the probability of a misclassification probability along the identification path in the tree was much larger than the misclassification probability in a 15 species hierarchy. Also, in the 74 species data set, some species were known to be very closely related to each other, increasing the probability of misclassification in the hierarchy. Despite a lower classification performance compared to flat multi-class classification, phylogenetic learning allowed to evaluate the classification scheme at node level. In this way, it was possible to analyze the resolution of FAME data at different tree levels. Ultimately, the goal of this approach will be to investigate how a particular pruning strategy could be applied by which those species will be grouped that are hard to classify by the machine learning method of interest. As a consequence, it will also become possible to report identification scores for groups of species that are very related in their FAME content. A first attempt towards this goal is discussed in the second section of this chapter.

Further investigation could also be done on the improvement of classification performance. For instance, a variable misclassification cost could be defined along the classification path. As an example, nodes splitting groups of classes could be evaluated differently than nodes splitting one species from a group of classes and splitting two leaves. In this latter case, a more severe misclassification cost could be defined. Another approach could account for the different branch lengths of the phylogenetic tree.

As the multi-class classification problem was tackled by hierarchically structured binary classifiers, it was also interesting to look at the individual class statistics. As mentioned in the Section 1.3, a multi-class confusion matrix was generated by classification of each test data point and counting the different types of errors that are made. Using the iTol webtool (Letunic and Bork, 2007), we plotted a bar diagram of sensitivity and F-score values along the tree

and aligned the corresponding bars with the corresponding leaf or class of the tree. F-scores corresponding to a value of ∞ (i.e. sensitivity and precision equal to zero) were, however, not visualized. In this way, rapid inspection was possible to detect those classes that were hard to identify by the phylogenetic learning model and the flat multi-class classifier.

The results of phylogenetic learning with NJ and UPGMA trees and those of flat multi-class classification are displayed in Figures 5.4-5.6, respectively. In case of flat multi-class classification, the metric values are displayed along the 16S rRNA gene NJ tree. Following decomposition of the multi-class confusion matrix into a two-class confusion matrix for each class in a one-versus-others strategy, it was possible to compare the prediction performance of each technique at class level. When comparing the sensitivity values of each species obtained by phylogenetic learning based on the two considered 16S rRNA gene trees to those obtained by multi-class classification, only 15% of the species had a higher sensitivity value. 57% and 61% of the species had a lower sensitivity value, for the NJ and the UPGMA tree respectively. In case of the F-score, 22% and 19% of the species had a higher F-score value, for both trees respectively, while 69% and 70% of the species had a lower sensitivity value. Nonetheless, when looking more deeply into the results, those classes that were hard to distinguish from the other classes in a multi-class classification setting were better identified in the hierarchical classification setting. This is clearly illustrated by the cumulative plot in Figure 5.7. In this figure, identification by phylogenetic learning was compared to flat multi-class identification at class level. Even though phylogenetic learning performed globally worse than flat multi-class classification, it was clear that, when considering a threshold of 0.5-0.6, phylogenetic learning had an added value due to better identification of classes that were not well identified by multi-class classification. For example, all species corresponding with a sensitivity value below 0.5 (in flat multi-class classification) were better identified by phylogenetic learning based on a NJ tree. These species were *B. azotoformans* (0.333 \rightarrow 1; 3 profiles), *B. funiculus* (0.4 \rightarrow 1; 5 profiles), *B. halmapalus* (0.5 \rightarrow 1; 6 profiles), *B. jeotgali* (0.5 \rightarrow 1; 6 profiles), *B. thuringiensis* (0.333 \rightarrow 0.5; 12 profiles) and *B. vallismortis* (0.267 \rightarrow 0.667; 15 profiles). The increase in sensitivity is given between brackets, together with the number of FAME profiles. *B. funiculus* (0.4 \rightarrow 0.8; 5 profiles), *B. thuringiensis* (0.333 \rightarrow 0.417; 12 profiles) and *B. vallismortis* (0.267 \rightarrow 0.533; 15 profiles) corresponded with an increase in sensitivity when phylogenetic learning was based on an UPGMA tree. Here, also only species with a sensitivity below 0.5 were considered. Note that *B. thuringiensis* belongs to the *B. cereus* group and *B. vallismortis* to the *B. subtilis* group. In phylogenetic learning based on the NJ tree, for five other species a higher sensitivity value was obtained: *B. aquimaris*, *B. firmus*, *B. lentus*, *B. megaterium* and *B. subterraneus*. In phylogenetic learning based on the UPGMA tree, eight species attained a higher sensitivity value: *B. drementensis*, *B. firmus*, *B. lentus*, *B. mycoides*, *B. novalis*, *B. pseudocaliphilus*, *B. sporothermodurans* and *B. vireti*. In most cases, only small increases were seen.

As mentioned above, a hierarchical classification structure easily allows to analyze where misclassifications occurred along the tree. This could be regarded as an evaluation approach to further analyze the resolution of FAME data for species discrimination. Furthermore, it was also interesting to calculate an average misclassification path length. Results for phylogenetic learning based on an NJ tree are visualized in Figure 5.8. Results for phylogenetic learning

based on a UPGMA tree were similar and are visualized in Figure 5.9. Herein, importantly, only misclassified test points were considered. It becomes clear from both figures that misclassification mostly occurred at nodes near the correct leaf. This is not very surprising as, based on FAME data, a lot of species cannot be distinguished from each other. This again shows that the resolution of FAME analysis is restricted to distantly related species and species groups.

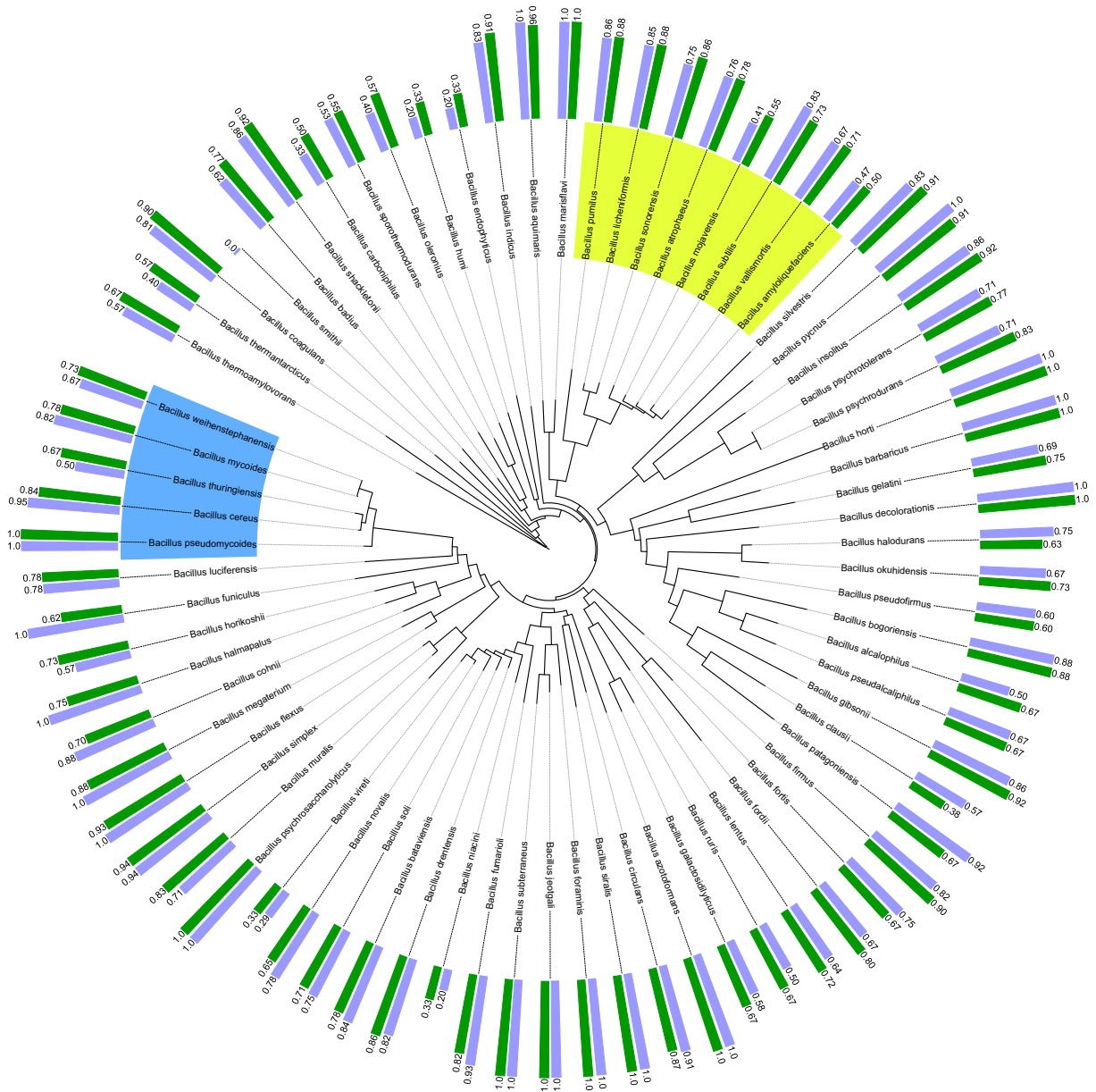


Figure 5.4: Sensitivity and F-score values by phylogenetic learning based on a 16S rRNA gene NJ tree. For each *Bacillus* species, the corresponding sensitivity and F-score value of phylogenetic learning based on a 16S rRNA gene NJ tree is displayed. Sensitivity is denoted by the light blue bars, F-score by the green bars. The tree is visualized using the iTol webtool (Letunic and Bork, 2007). The *Bacillus cereus* and *Bacillus subtilis* groups are coloured in blue and green, respectively.

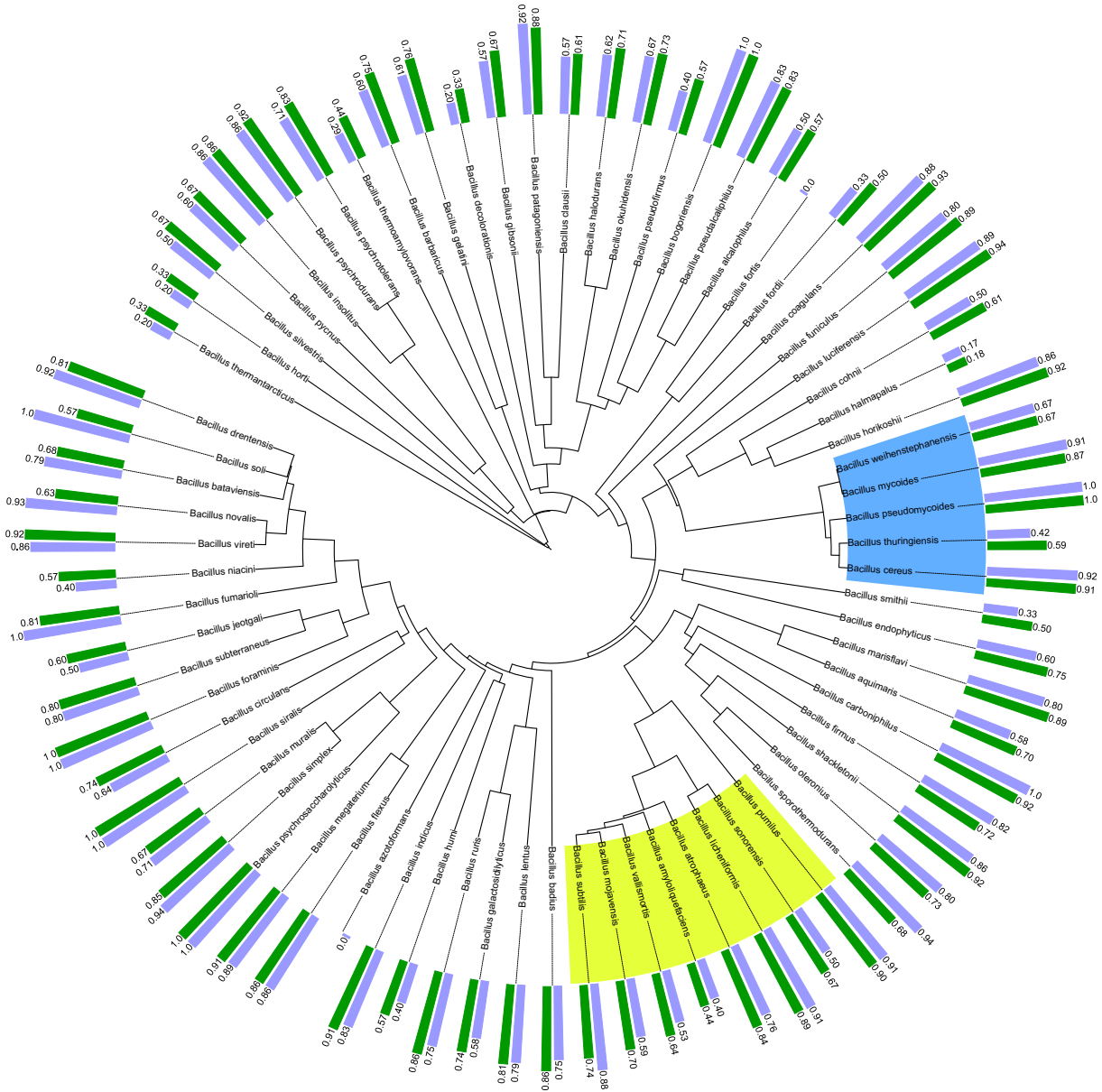


Figure 5.5: Sensitivity and F-score values by phylogenetic learning based on a 16S rRNA gene UPGMA tree. For each *Bacillus* species, the corresponding sensitivity and F-score value of phylogenetic learning based on a 16S rRNA gene UPGMA tree is displayed. Sensitivity is denoted by the light blue bars, F-score by the green bars. The tree is visualized using the iTol webtool (Letunic and Bork, 2007). The *Bacillus cereus* and *Bacillus subtilis* groups are coloured in blue and green, respectively.

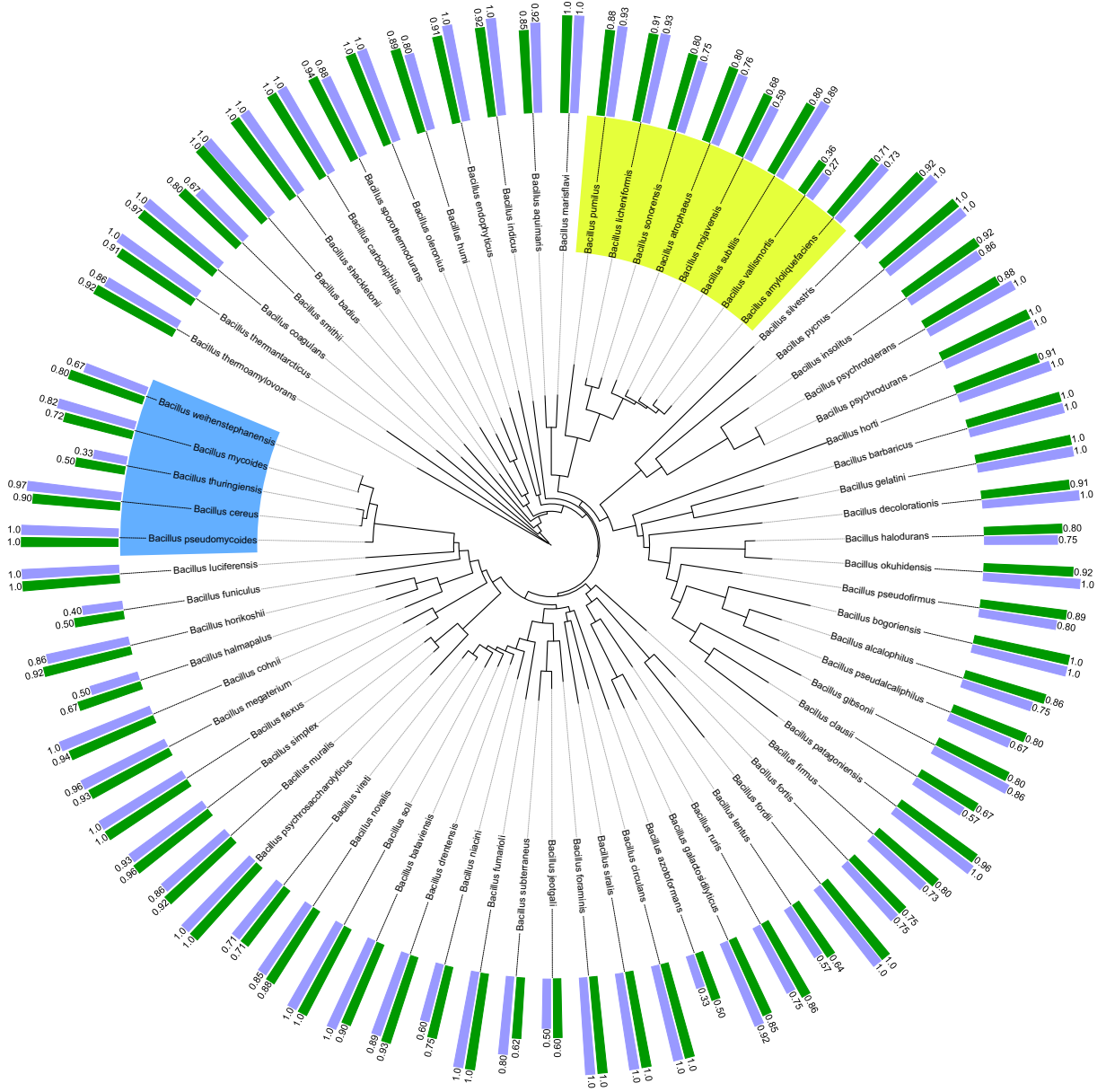


Figure 5.6: Sensitivity and F-score values for flat multi-class classification. For each *Bacillus* species, the corresponding sensitivity and F-score value for flat multi-class classification is displayed along the 16S rRNA gene NJ tree. Sensitivity is denoted by the light blue bars, F-score by the green bars. The tree is visualized using the iTol webtool (Letunic and Bork, 2007). The *Bacillus cereus* and *Bacillus subtilis* groups are coloured in blue and green, respectively.

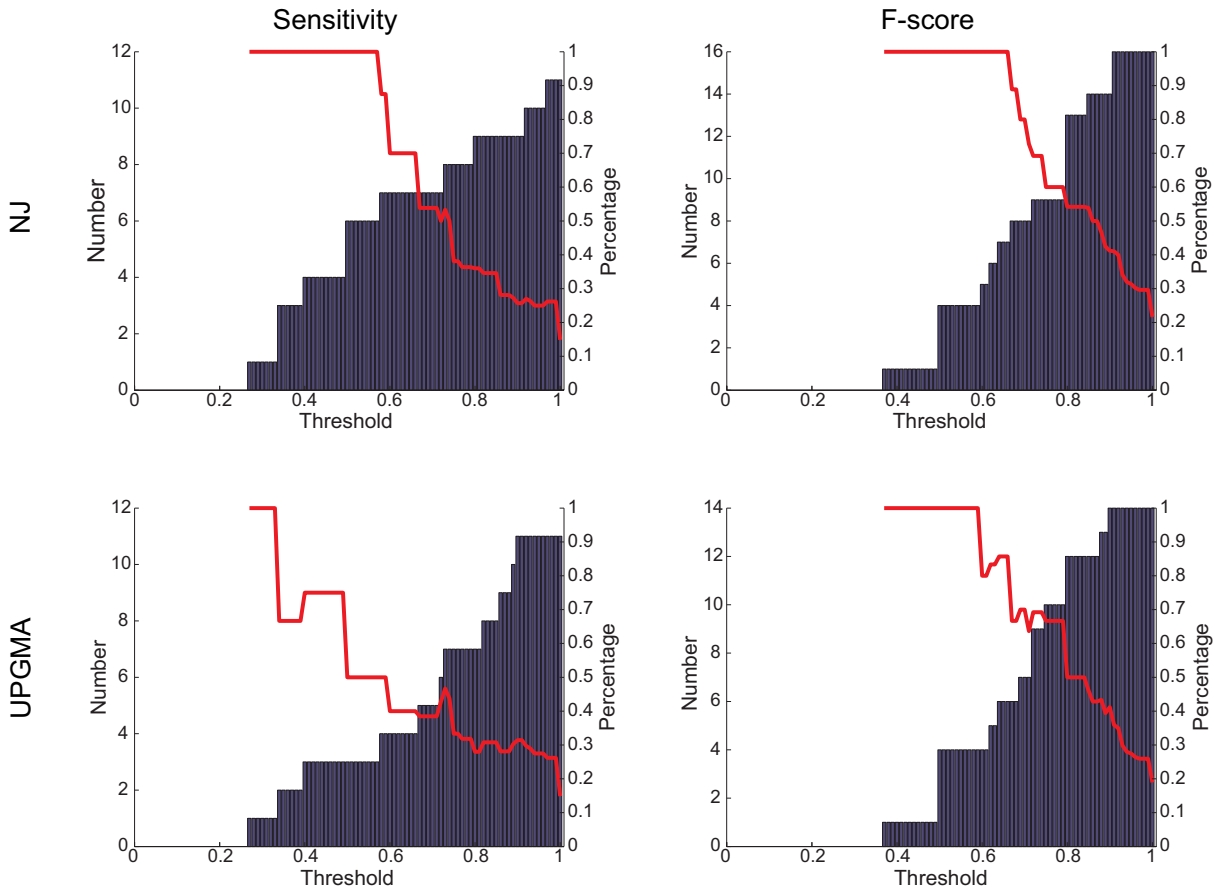


Figure 5.7: Performance comparison at class level. For each class, sensitivity and F-score values resulting from phylogenetic learning based on a 16S rRNA gene NJ or UPGMA tree were compared to those obtained by flat multi-class classification. Four plots are given. The X-axis corresponds to thresholds set on the corresponding metric values. Threshold steps of 0.01 were chosen. For each threshold, flat multi-class classification was evaluated at class level and those classes with metric values smaller than or equal to the threshold were selected. Classification performance by phylogenetic learning was analyzed at class level for each set of classes. The Y-axis on the left projects the number of phylogenetic learning classes that had a higher metric value than those obtained by flat multi-class classification. The red line expresses this number, relative to the size of the corresponding set (Y-axis on the right).

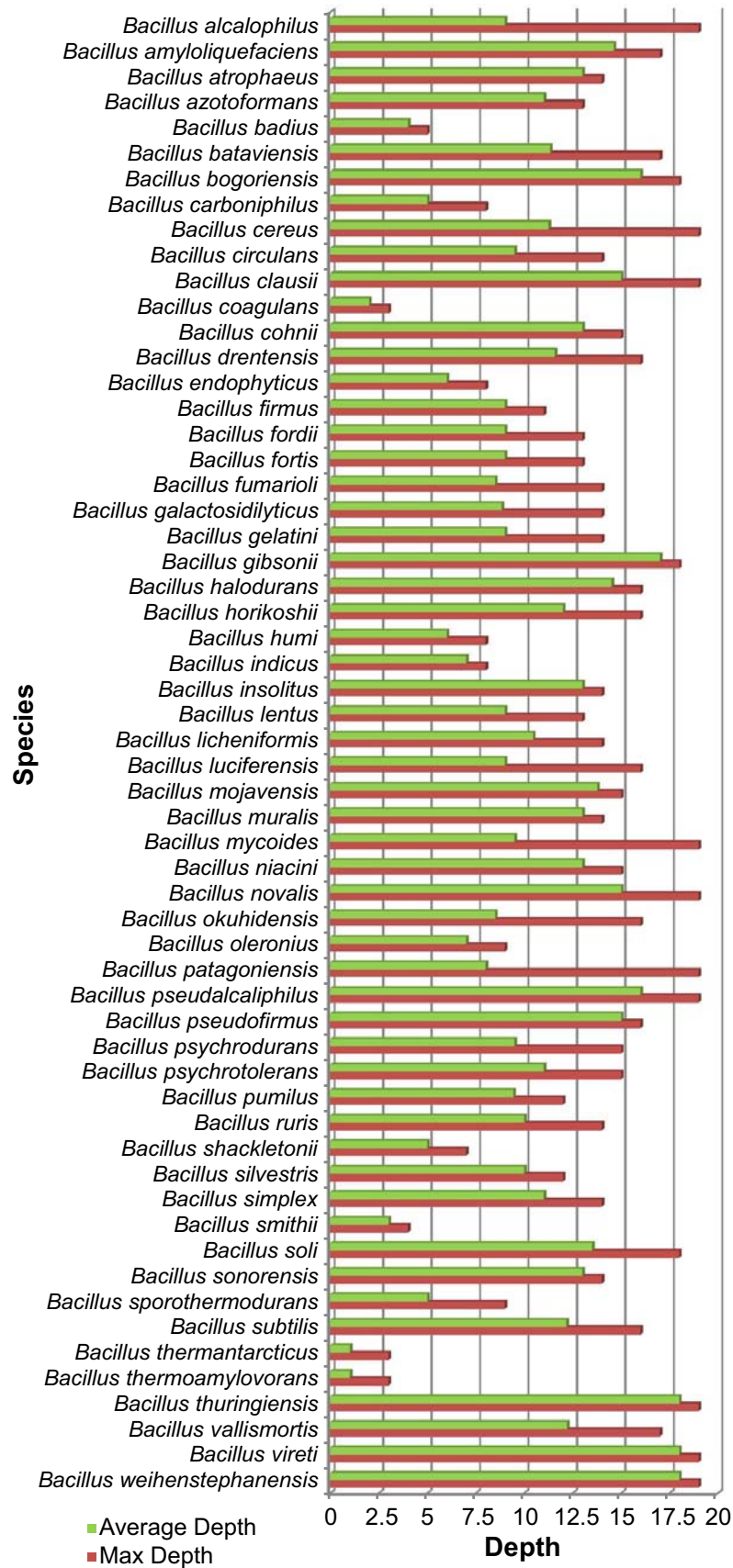


Figure 5.8: Average misclassification depth of phylogenetic learning based on a NJ tree. The average depth of the misclassified test points of each class is visualized for phylogenetic learning based on an NJ tree. Depth equals the number of nodes along the classification path until misclassification occurs (the corresponding node also included) and corresponds to the green bars. The maximum or correct depth is shown by the red bars. Maximum depth equals the number of nodes along the true phylogenetic path (final leaf included).

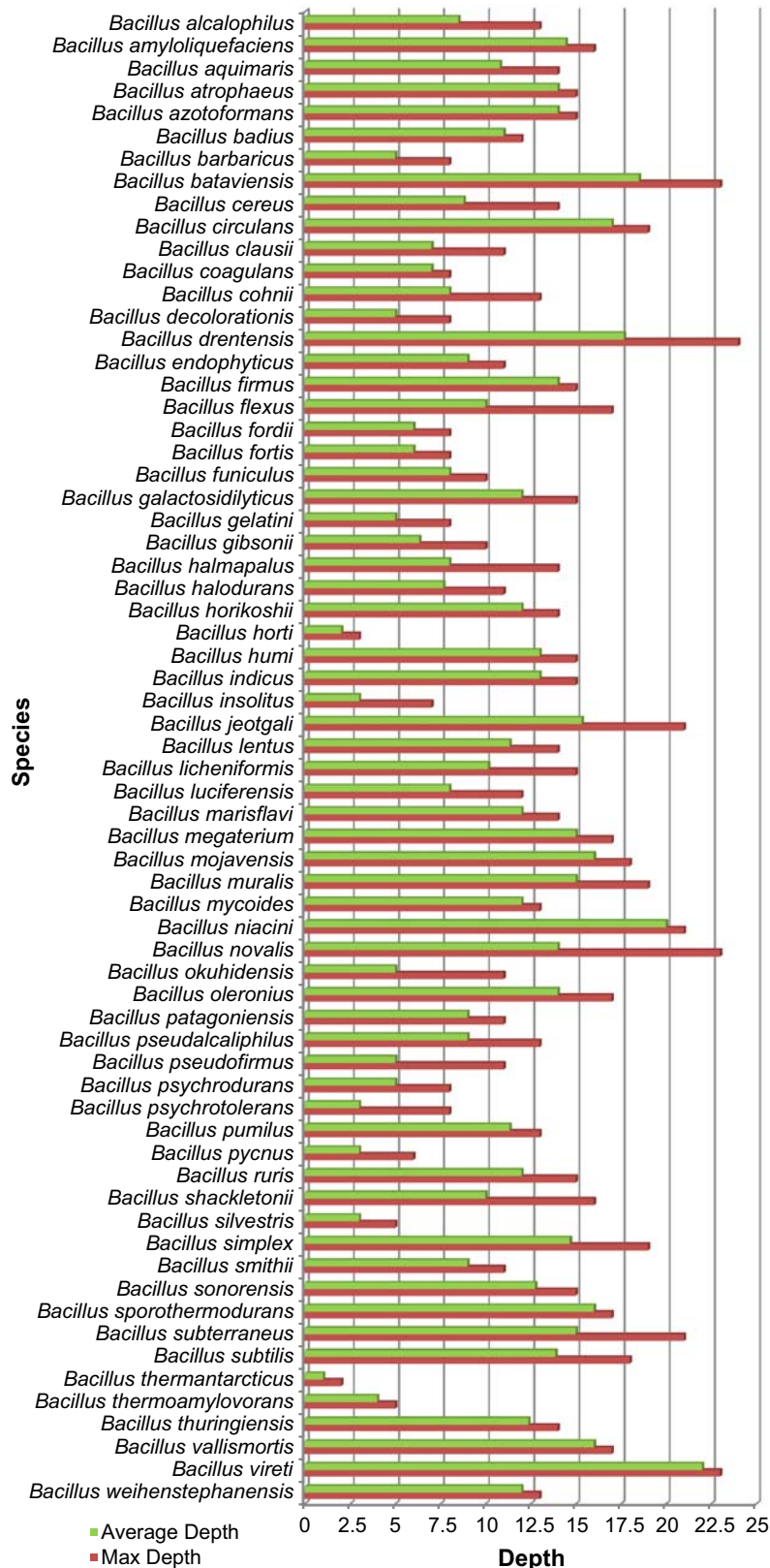


Figure 5.9: Average misclassification depth of phylogenetic learning based on a UPGMA tree. The average depth of the misclassified test points of each class are visualized for phylogenetic learning based on a UPGMA tree. Depth equals the number of nodes along the classification path until misclassification occurs (the corresponding node included) and corresponds to the green bars. The maximum or correct depth is shown by the red bars. Maximum depth equals the number of nodes along the true phylogenetic path (the final leaf included).

5.2.3 Publication

The research described in this section is submitted to the international peer-reviewed journal of BMC Bioinformatics with reference: B. Slabbinck, W. Waegeman, P. Dawyndt, P. De Vos and B. De Baets (2009). From learning taxonomies to phylogenetic learning: Integration of 16S rRNA gene data into FAME-based bacterial classification.

5.3 Putting Bacterial Species Identification into Context

In this section, we evaluate a first approach towards a pruning strategy for the phylogenetic learning approach and aim at an improved bacterial species identification. More specifically, by the Wilcoxon rank-sum test and a three colour scheme, a new tool was developed that allows for easy visualization of those bacterial species that were hard to distinguish from each other.

5.3.1 Statistically Significant Nodes

One of the most straightforward strategies for post-processing of the results obtained by phylogenetic learning is to highlight those tree nodes of which the underlying species and groups of species are possibly hard to distinguish from each other. In our case of hierarchically structured classes, this required the evaluation of the identification performance of the different models at the different tree nodes. When dealing with identification performance, though, evaluation also has to take into account the corresponding data set size for learning the model. As the data set size has a serious effect on the identification reliability, those nodes corresponding to a small data set size should also become highlighted. In this perspective, a good choice for statistical testing of the RF probability estimates is the non-parametric Wilcoxon rank-sum test (see also Subsection 1.3.3). Remember that, at each node, the RF probability estimates are resulting from the pooled test set (following cross-validation for performance estimation). In essence, we were confronted with identification scores of two classes, where, ideally, the scores of the first class equal zero, while those scores of the second class equal one (or conversely). Thus, this problem involved the detection of a significant difference in the identification scores. Scaled to our FAME data identification setting, the null hypothesis corresponded to similar probability estimates which, in other words, corresponded to FAME profiles that were hard to distinguish. Logically, the alternative hypothesis assumed significantly different probability estimates. An upper-tail Wilcoxon rank-sum test was executed, with normal approximation of the p -values in case of ties and large sample sizes (or number of probability estimates). All assumptions of the test were met, except for the shapes of the underlying distributions which were assumed identical. The significance level α was set at the conventional levels 0.05 and 0.01.

In the presented hierarchical setting, multiple hypotheses were tested which are correlated by the underlying data and the classes they represent. By applying the standard significance levels, statistical evaluation of all two-class classifications resulted in significantly different identification scores only. Or, this corresponded to significantly different FAME profiles in

all clusters. From a biological perspective, this is not a correct conclusion, as we showed in the previous two chapters that it was often hard to distinguish between the different species by FAME data. In other words, the significance levels were too weak. In the field of statistics, testing multiple hypotheses at the same time is called multiple hypothesis testing or the problem of multiplicity (Shaffer, 1995). This field is a key topic studied in microarray analysis where, for instance, multiple genes are analyzed for having an effect in a particular disease with respect to a control group. Each test has a specified type I error probability (i.e. rejecting the null hypothesis when it is actually true), but when many hypotheses are tested, the probability that at least some type I errors are committed increases with the number of hypotheses. This may have serious consequences if the set of conclusions must be evaluated as a whole. Numerous methods have been proposed for dealing with this problem and one popular method involves the calculation of adjusted p -values. Computing adjusted p -values can be performed by several methods that vary in the severity of the correction for multiplicity (Shaffer, 1995; Dudoit et al., 2002). One of the most popular multiplicity correction methods is the unweighted Bonferroni correction method. This method redefines the significance levels as $\frac{\alpha}{z}$, further denoted as α^* , with z the number of p -values or hypotheses. With regard to the p -values, the rejection rule $p_i \cdot z < \alpha$ is identical to the former rule, with $i = 1, \dots, N$ and N the number of p -values. In our setting, we had a slightly different setup in that a single observation (FAME data) was used to distinguish between hierarchically ordered species. Also, no global evaluation of the problem setting as a whole was pursued. Though, we evaluated the Bonferroni method for adjustment of the significance levels. Note that the unweighted Bonferroni correction is known to be very conservative, resulting in overcorrection and leading to a too strict adjustment of the p -values (Shaffer, 1995). Nonetheless, in this particular work, we pursued to inform microbiologists about those specific species that are hard to distinguish from each other based on FAME data by a specific machine learning method. As such, type I errors may be less severe than type II errors, i.e. falsely accepting significant output probability estimates. The converse conclusion would result in a serious flaw in decision-making. Because of these reasons, we accepted the unweighted Bonferroni correction, though, keeping its origin and drawbacks in mind.

In order to highlight significant branches in the phylogenetic tree, the two redefined significance levels were used. For classifiers with a p -value larger than 0.05^* , the corresponding nodes and the two splitted branches were highlighted in red, implying similar patterns in the FAME profiles. p -values situated in the interval $[0.01^*, 0.05^*]$ resulted in an orange highlighting and otherwise highlighting was done in green. This latter case implies two groups with significantly different probability estimates. The adjusted p -values are given at each corresponding node of the phylogenetic trees (bottom value). This three colour system was chosen to show a degree of significant difference between the output probability estimates, or thus between the FAME profiles of the different species.

5.3.2 Visualization and Evaluation

Phylogenetic learning was evaluated as a technique for analyzing the resolution of FAME analysis for species discrimination within the three genera *Bacillus*, *Paenibacillus* and *Pseu-*

domonas. The same parameter optimization and classification strategy was followed as in the previous section. We investigated the extraction of meta-information to improve the report of a FAME-based identification technique. This knowledge representation started from 16S rRNA gene NJ and UPGMA phylogenetic trees constructed for the three genera. For all 16S rRNA trees, the FAME-based species identification performance of the different models, as constructed on the different nodes, was visualized using the three colour system. Results are visualized in Figures 5.10-5.15.

When comparing the highlighted branches of the two trees of each genus, it could generally be concluded that the same branches were highlighted in orange or red, which corresponds to taxa with FAME profiles that were hard to distinguish, or even indistinguishable. The main highlighting was seen of terminal interior nodes and of branches corresponding to a leaf branched from a non-terminal interior node. Moreover, highlighting of the latter case was mostly in orange, while highlighted terminal interior nodes mostly coloured in red. No clear argument was found for this latter finding, which was influenced by two facts: the separability of the FAME profiles of the two groups and the number of profiles in the subtree. Given the recalculated significance levels, this again showed that FAME analysis performed well to distinguish a lot of species and most species groups from each other, but became problematic when two close relatives need to be separated. Another important influence on the results was posed by the number of profiles in each subtree. Near the leaves, a smaller data subset was handled, making the generalization over the data harder to achieve. Hence, lower and, thus, highlighted identification performances were obtained. Note that the same conclusion held for species branched from large subtrees.

In the case of the genus *Bacillus* (see Figures 5.10 and 5.11), when focusing on the terminal interior nodes with non-green highlighted branches, 15 nodes were highlighted in both the NJ and UPGMA tree. 12 identical nodes could be found, of which 10 had the same highlighting colour. Finding the same colouring should be expected because closely related species should cluster together by either tree inference method. When focusing on single highlighted leaves or species, splitted from non-terminal interior nodes, five UPGMA leaves and eight NJ leaves were highlighted as non-green. Four of these highlighted leaves occurred in both trees, however, not necessarily in the same phylogenetic structure. In case of the two known species groups *Bacillus cereus* group and *Bacillus subtilis* group unexpected results were, however, obtained. Especially, based on a RF model, all *Bacillus subtilis* group species could significantly be distinguished from the other member species. When looking at the classification results, statistics also showed that identification was relatively high. Note also that this group of species corresponded to a large number of FAME profiles. This confirms evidence from literature stating that some *B. subtilis* group species can be differentiated from each other based on FAME, i.e. *B. amyloliquefaciens*, *B. licheniformis* and *B. pumilus* (Vaerewijck et al., 2001; Coorevits et al., 2008). Important to remark here is that this result was not found in the flat multi-class classification models performed in the current chapter as well as in the previous chapter. For this group, phylogenetic learning clearly took advantage of the hierarchical learning setting. When focusing on the *Bacillus cereus* group, the *Bacillus mycoides* and *Bacillus weihenstephanensis* species, and the *Bacillus cereus* and *Bacillus thuringiensis* species corresponded to non-significant FAME

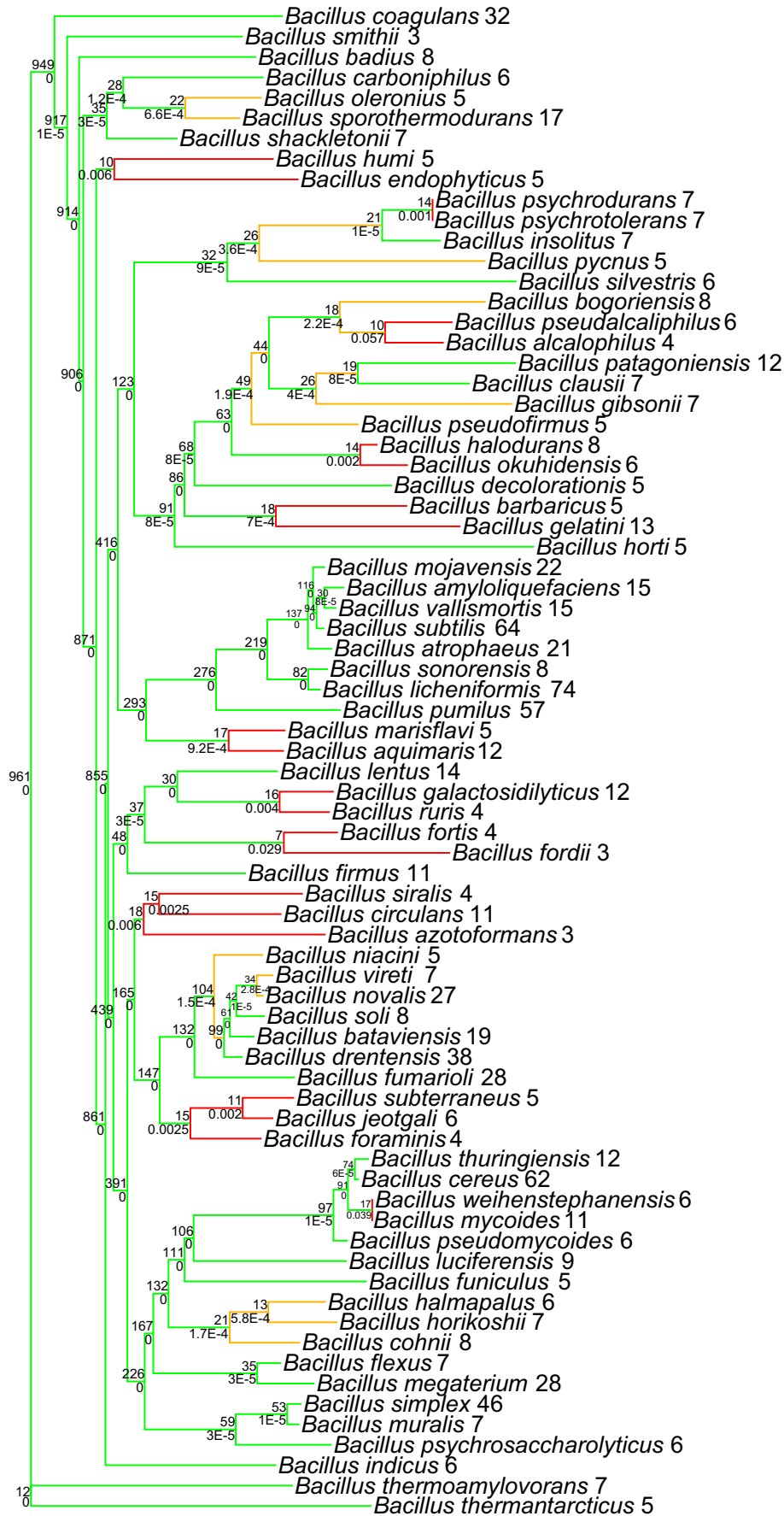


Figure 5.10: Statistical evaluation of phylogenetic learning for the genus *Bacillus* with a 16S rRNA gene NJ phylogenetic tree. Phylogenetic learning was performed with the 2008 *Bacillus* data set. The number of FAME profiles of each species is reported following the species name. For each node, the number of FAME profiles of the subtree (top value) is given, together with the p -values (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval [0.01*, 0.05*] and green colour to p -values below 0.01*.

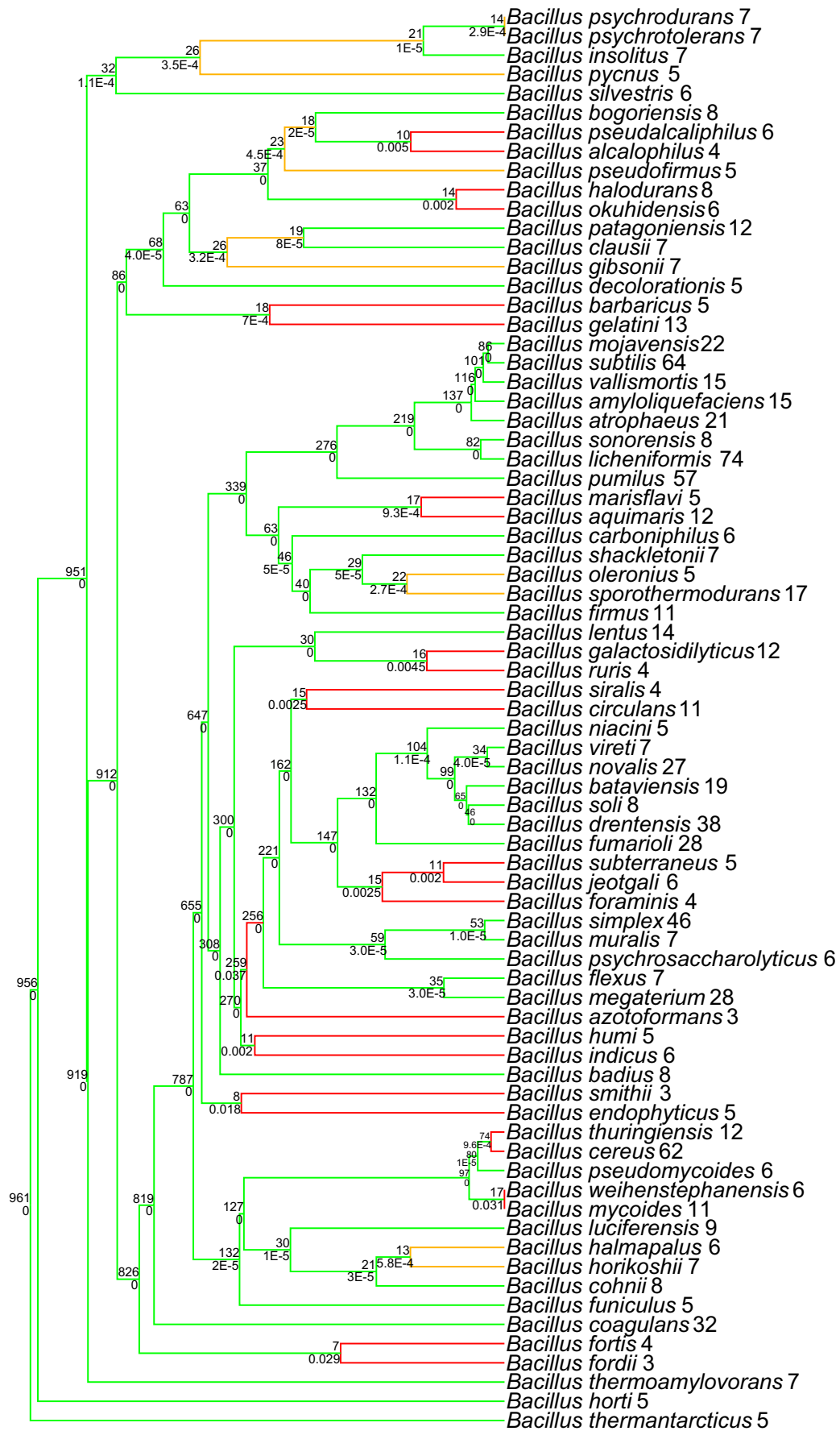


Figure 5.11: Statistical evaluation of phylogenetic learning for the genus *Bacillus* with a 16S rRNA gene UPGMA phylogenetic tree. Phylogenetic learning was performed with the 2008 *Bacillus* data set. The number of FAME profiles of each species is reported following the species name. For each node, the number of FAME profiles of the subtree (top value) is given, together with the p -value (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval [0.01*, 0.05*] and green colour to p -values below 0.01*.

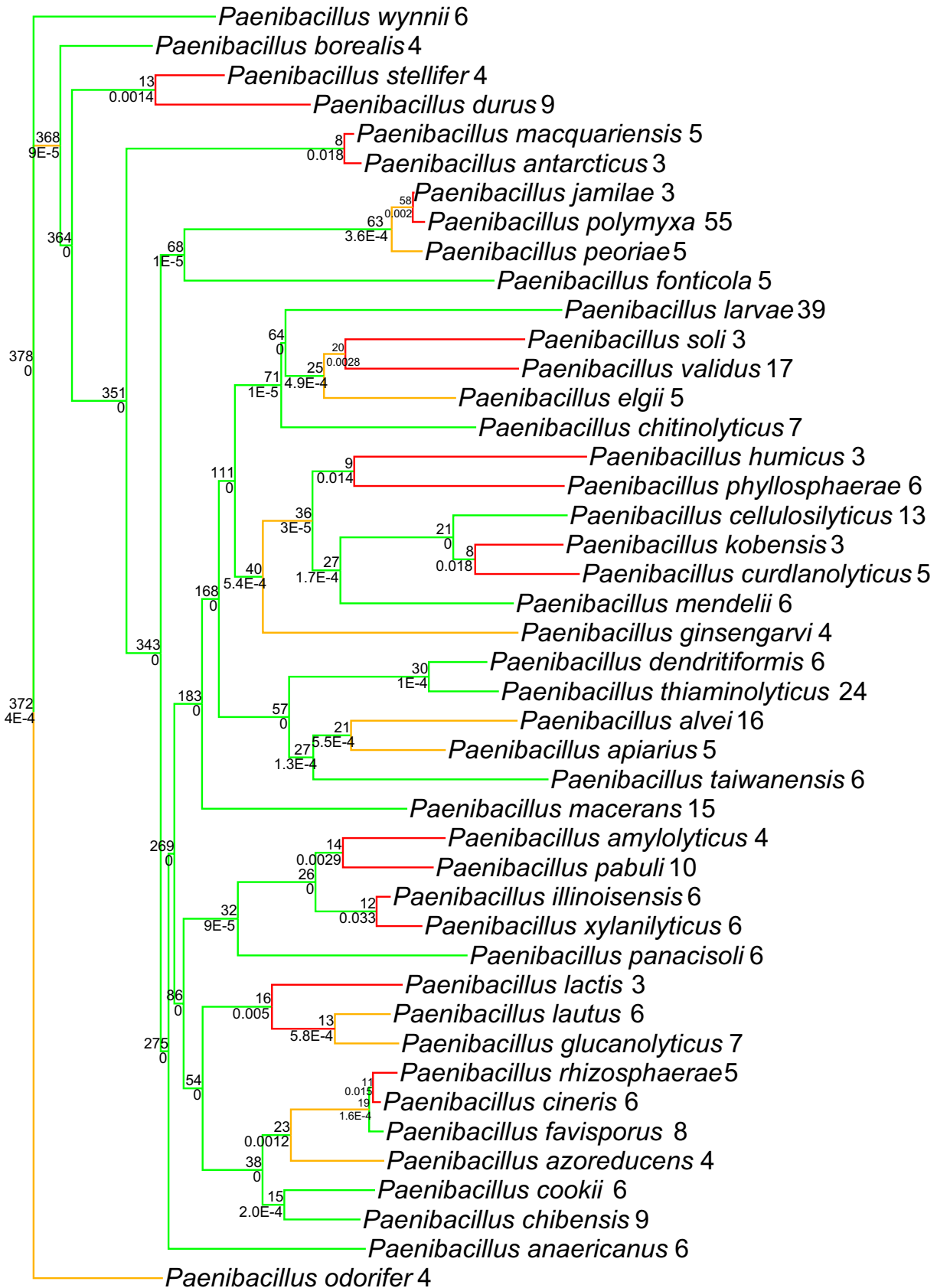


Figure 5.12: Statistical evaluation of phylogenetic learning for the genus *Paenibacillus* with a 16S rRNA gene NJ phylogenetic tree. Phylogenetic learning was performed with the 2008 *Paenibacillus* data set. The number of FAME profiles of each species is reported following the species name and for each node the number of FAME profiles of the subtree (top value) is given, together with the p -value (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval $[0.01^*, 0.05^*]$ and green colour to p -values below 0.01*.

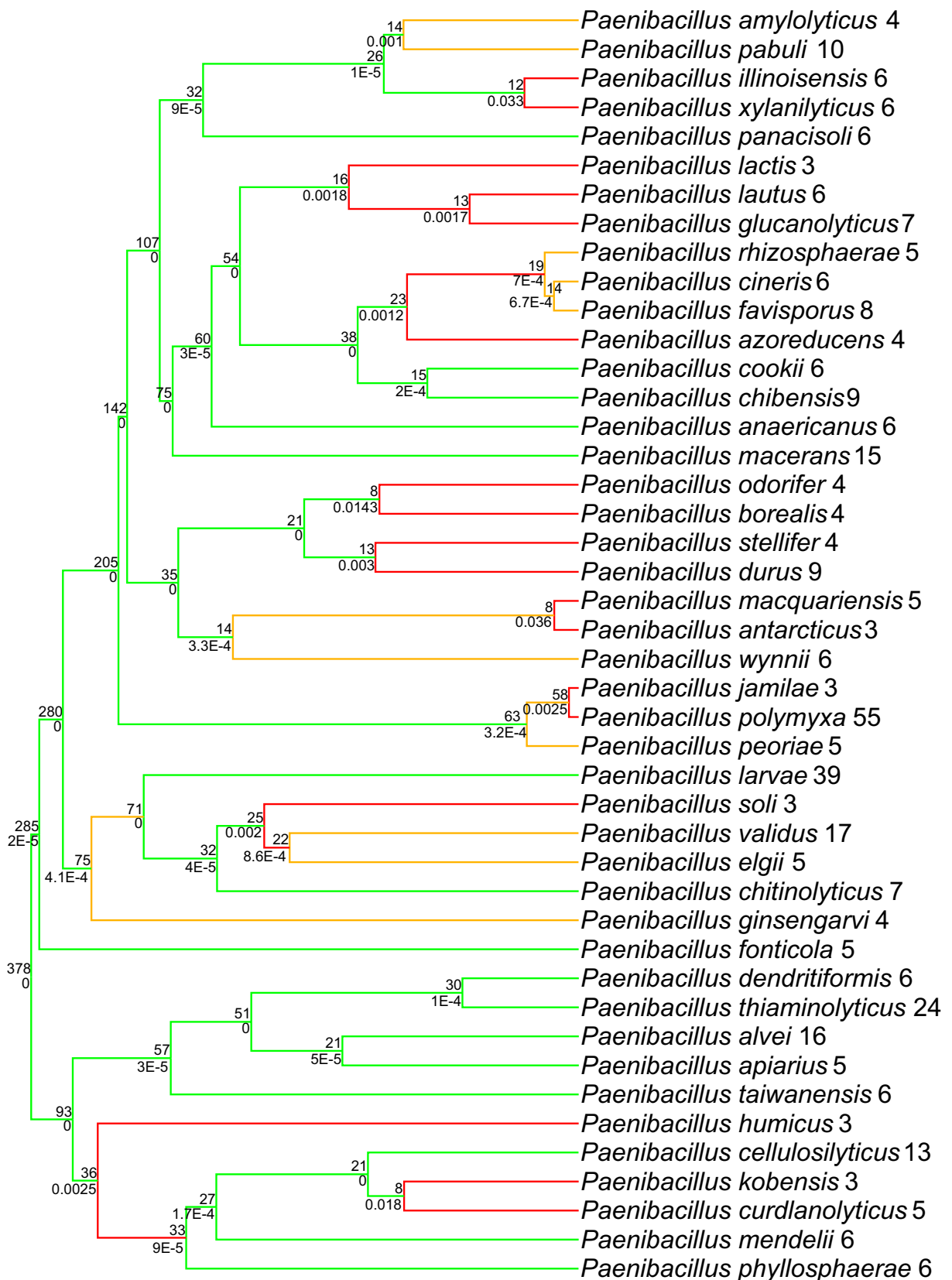


Figure 5.13: Statistical evaluation of phylogenetic learning for the genus *Paenibacillus* with a 16S rRNA gene UPGMA phylogenetic tree. Phylogenetic learning was performed with the 2008 *Paenibacillus* data set. The number of FAME profiles of each species is reported following the species name. For each node, the number of FAME profiles of the subtree (top value) is given, together with the p -value (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval $[0.01^*, 0.05^*]$ and green colour to p -values below 0.01*.

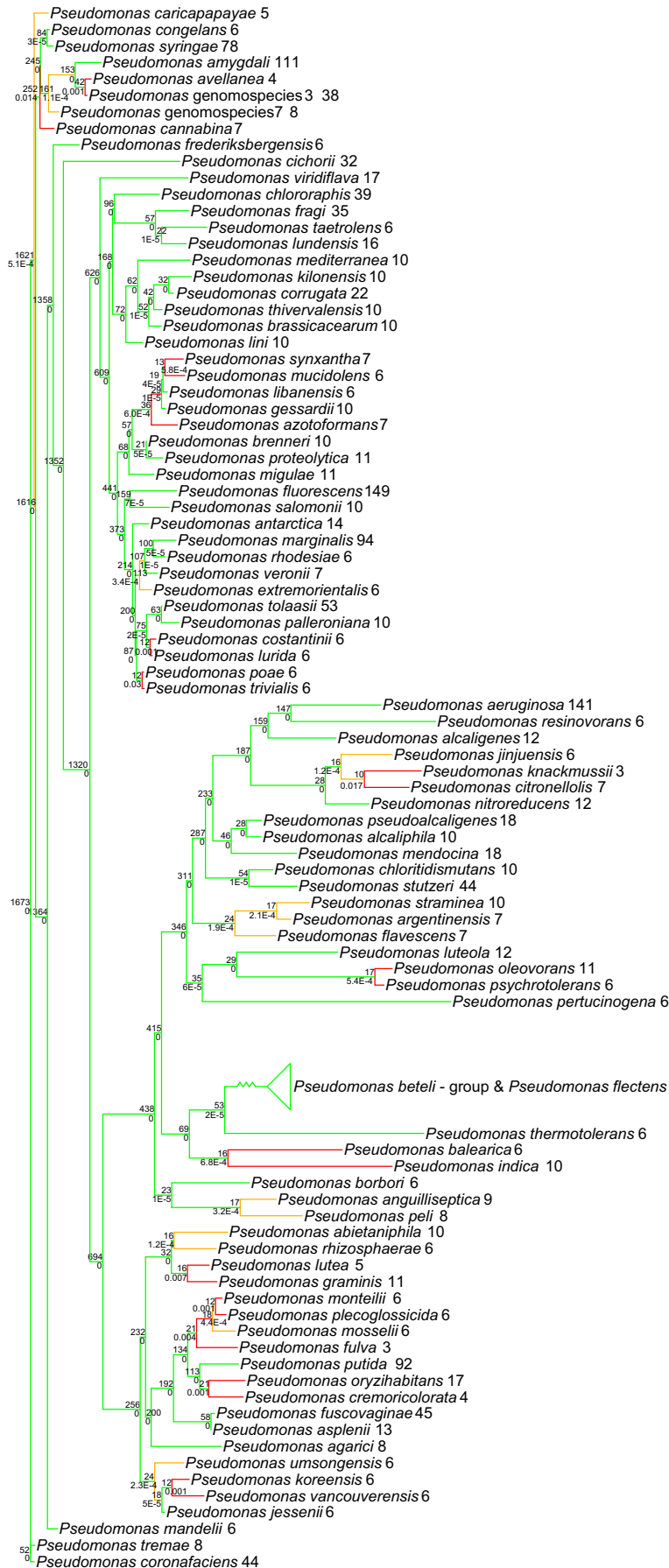


Figure 5.14: Statistical evaluation of phylogenetic learning for the genus *Pseudomonas* with a 16S rRNA gene NJ phylogenetic tree. Phylogenetic learning was performed with the 2008 *Pseudomonas* data set. The number of FAME profiles of each species is reported following the species name. For each node, the number of FAME profiles of the subtree (top value) is given, together with the *p*-value (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to *p*-values above 0.05*, orange colour to *p*-values in the interval [0.01*, 0.05*] and green colour to *p*-values below 0.01*.

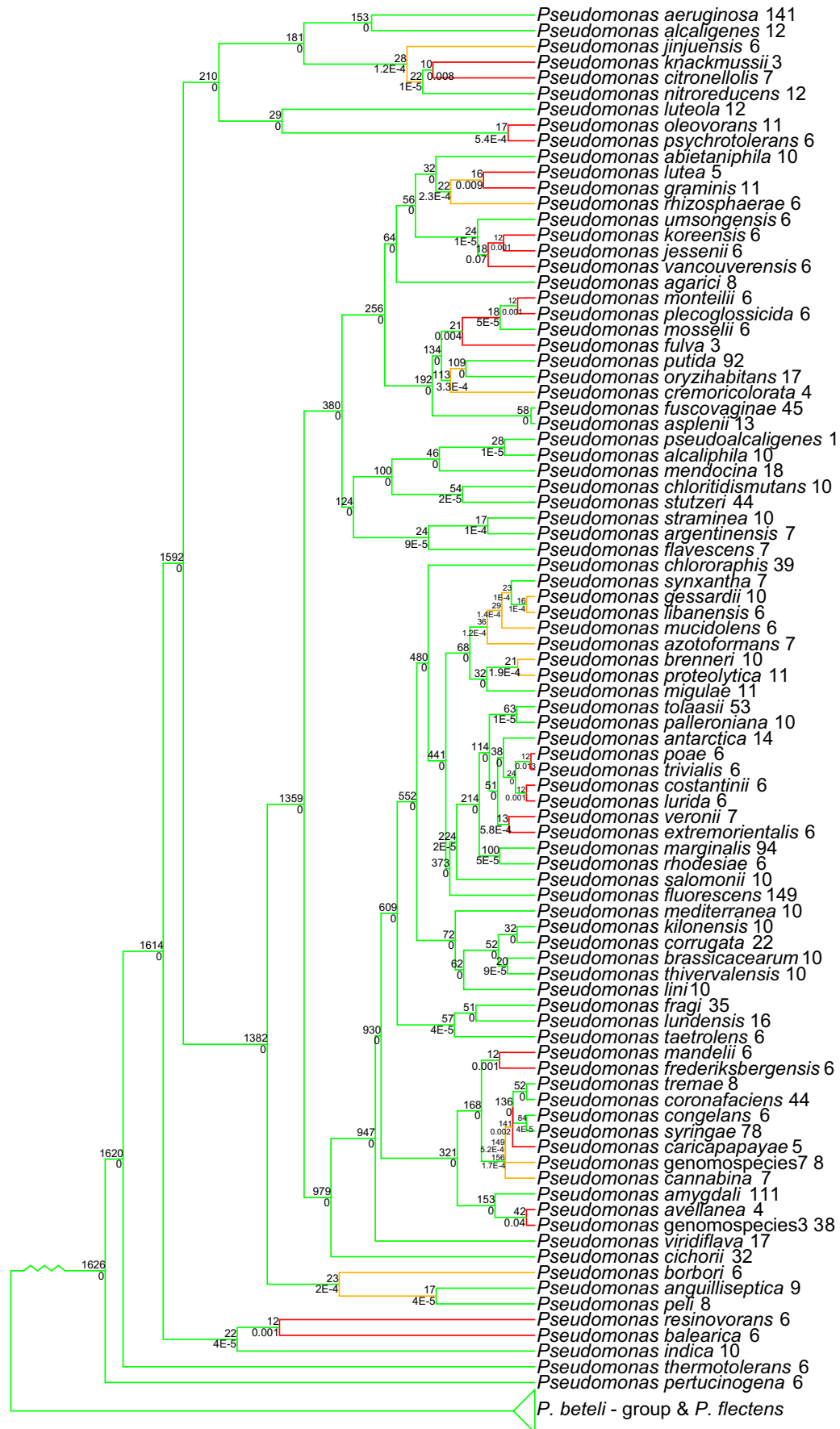


Figure 5.15: Statistical evaluation of phylogenetic learning for the genus *Pseudomonas* with a 16S rRNA gene UPGMA phylogenetic tree. Phylogenetic learning was performed with the 2008 *Pseudomonas* data set. The number of FAME profiles of each species is reported following the species name. For each node, the number of FAME profiles of the subtree (top value) is given, together with the p -value (bottom value). A Bonferroni correction was performed for the significance levels 0.05 and 0.01, respectively denoted as 0.05* and 0.01*. Red colour corresponds to p -values above 0.05*, orange colour to p -values in the interval [0.01*, 0.05*] and green colour to p -values below 0.01*.

profiles, except for the latter group in the NJ tree. Nonetheless, except for *Bacillus thuringiensis* relatively good identification statistics were obtained for all group species considered. The results for the genus *Paenibacillus* are shown in Figures 5.12 and 5.13. In the UPGMA tree, seven terminal interior nodes and four branched species were highlighted in red, together with three terminal interior nodes and four branched species with an orange highlighting. In the NJ tree, nine terminal interior nodes and one branched species were highlighted in red, together with two terminal interior nodes and five branched species in orange highlighting. For red highlighting, most groups also occurred in an identical phylogenetic structure in both trees, while for orange highlighting the coloured groups/species differed, mainly due to values around the defined thresholds. The results for the genus *Pseudomonas* are shown in Figures 5.14 and 5.15. In the UPGMA tree, 11 terminal interior nodes and two branched species were highlighted in red, together with two terminal interior nodes and eight branched species in orange highlighting. In the NJ tree, 11 terminal interior nodes and three branched species were highlighted in red, together with three terminal interior nodes and seven branched species in orange highlighting. Also in this case, most red highlighted groups were found in both trees but orange highlighting differed, again due to p -values near the threshold. Interestingly, most species of the *P. syringae* subcluster showed a non-green highlighting, showing again that it is hard to discriminate between species of this group based on FAME data. Also in the *P. putida* subcluster an observable highlighting occurred as about half of the species in this cluster became highlighted.

In collaboration with microbiologists who routinely perform FAME analysis, visual inspection of the highlighted trees led to the conclusion that most red and orange highlighted leaves or species were indeed known as difficult to distinguish. It is clear that the occurrence of identical groups of nodes and leaves in 16S rRNA phylogenetic trees resulting from different treeing methods imply a strong coherence or close relatedness between the corresponding leaves. All this consistent information proved to be highly convenient for enhancing the identification process of unlabeled FAME profiles. When generating a FAME profile identification report, the extracted knowledge could be further used to inform users about species that are moderately hard (orange colour) or hard (red colour) to distinguish from some other species or a group of species, or that too few data was at hand for the respective species to attain a reliable identification. This meta-information approach is an additional step in the direction of improved identification of bacterial species. By this highlighting system, meta-information combining taxonomic knowledge and relatedness in FAME profiles was proposed to enhance FAME-based identification of bacterial species by machine learning techniques.

5.4 Conclusions

In this chapter, we investigated taxonomic knowledge integration in multi-class classification models, using RF as machine learning technique. The 2008 *Bacillus* data set was considered which contained 74 species, 71 FAME peaks and 961 standard FAME profiles. In the previous chapters, we showed that FAME data does not allow to discriminate between all bacterial species. Supervised machine learning techniques showed to result in a moderate to good identification performance to distinguish between the FAME profiles of different *Bacillus* species.

Nonetheless, the classification models do not integrate any knowledge about the taxonomical relationships between the different species. In this whole concept of species identification, this information is quite important and can even be determinative in evaluating species discrimination. Two strategies were followed for knowledge integration concerning the taxonomic relationships between the different classes or species. First, divisive clustering with classifier performance as splitting criterion was considered to construct a FAME tree, resembling the hierarchical information hidden in the FAME profiles of the considered species. Due to combinatorial and computational issues, this experiment was restricted to the clustering of 15 species. Relatively good results were obtained as closely related species were retained out of the massive amount of computed clusters. Future research should look on how to resolve this problem. Also hierarchical classification for these 15 species was evaluated and a classification performance was obtained comparable to that obtained with flat multi-class classification. Secondly, we evaluated the approach of knowledge integration based on 16S rRNA gene data. In contrast to FAME data, the 16S rRNA gene does allow to discriminate the different bacterial species in most cases. Using quality controlled 16S rRNA gene sequences, phylogenetic trees were constructed for the type strains of the 74 considered *Bacillus* species, as validly published in March 2008. The two treeing algorithms under consideration were the NJ and UPGMA method. These trees were inferred from distance matrices computed from the aligned 16S rRNA sequences and corrected for the Jukes-Cantor evolution model. Hierarchically structured binary classifiers were trained at each node of the two phylogenetic trees to distinguish the FAME profiles corresponding to the two underlying branches. From a biological perspective, this should be a good starting point for classification, as it is our goal to distinguish between different bacterial species that are hierarchically structured in a taxonomy based on evolutionary relationships. When considering 16S rRNA gene phylogenetic trees as template for classification, a new approach for hierarchical classification was proposed. Herein, two types of data were combined for hierarchical classification: 16S rRNA gene sequences for defining the different classification tasks, and FAME data for classification of bacterial species as defined by these tasks. Because the 16S rRNA gene was used for phylogenetic analysis and this gene allows to discriminate most bacterial species, we call this approach phylogenetic learning. Phylogenetic learning with the NJ and UPGMA trees for the 74 species showed to be less accurate than flat multi-class classification. Phylogenetic learning with both trees also resulted in a similar identification performance. When classifier selection is at stake, preference should be given to the classifier with the best global performance and the lowest computational cost. Consequently, flat multi-class classification should be preferred over phylogenetic learning. However, as bacteria are structured in a taxonomy based on genotypic and phenotypic analysis, relevant information about this hierarchy is not used by flat classification methods. This knowledge was, however, integrated in the phylogenetic learning method. A clear advantage of this integration was seen when focusing on the incorrectly identified species in flat multi-class classification, as phylogenetic learning clearly took advantage of the binary classifier hierarchy to improve identification of these classes.

In summary, good strategies were found for knowledge integration of bacterial species identification based on FAME data. The next step to take in this taxonomic knowledge integration study was to develop a system for exploiting and visualizing the performance of the different

binary classifiers. The Wilcoxon rank-sum statistic was used to detect similar identification scores for FAME profiles of the two considered species. Two significance levels were chosen to map the resulting p -values to a three-colour highlighting scheme of red-orange-green. Adjusted significance levels were calculated with the unweighted Bonferroni method, similar to the multiplicity problem. Given this method and the recalculated levels, different interior nodes and leaves became highlighted. Red groups were consistently found in both the NJ and the UPGMA tree. These groups comprise species covering indistinguishable FAME profiles but also species with a number of FAME profiles too low for enabling a good generalization and identification. Orange groups were not consistently found, due to p -values near the defined thresholds. Ideally, this technique could be used as meta-information for further enhancing the identification report of FAME-based identification. Nonetheless, the method requires a further refining and evaluation. Ultimately, when using intelligent learning models, microbiologists will be able to resolve which groups of species are hard to distinguish from each other.

PART III

ONLINE DATABASE

CHAPTER 6

FAME-bank.net

Sharing knowledge is not about giving people something, or getting something from them.

That is only valid for information sharing.

Sharing knowledge occurs when people are genuinely interested in helping one another develop new capacities for action; it is about creating learning processes.

PETER SENGE

6.1 Introduction

First-line identification methods are a good option for rapid bacterial typing and narrowing down the bacterial spectrum. These identification methods do not only allow for identification at the genus level, but can even result in identification at the species and strain level. In this perspective, chemotaxonomic methods such as fatty acid methyl ester (FAME) analysis are often used because they are cheap, fast, automated and high-throughput. Following the introduction of gas chromatography by James and Martin in 1952, gas chromatographic (GC) fatty acid analysis of bacteria started around the first years of the 1960s by investigation of the species *Bacillus subtilis* and a species of the genus *Sarcina* (Akashi and Saito, 1960; Saito, 1960a,b; Kaneda, 1963a). For about 50 years now, FAME analysis has become one of the routine methods in many institutes for fast bacterial identification. Most laboratories perform bacterial FAME analysis using the commercial Sherlock Microbial Identification System (MIS) of the company MIDI Inc. (Newark, Delaware, USA), which, thus, has become the reference when performing FAME analysis. A serious implication of the routine use of this first-line bacterial identification method is that the high-throughput analysis has led to huge but private FAME databases. The joint FAME database of the Laboratory of Microbiology and the BCCMTM/LMG Bacteria Collection (LMG, Ghent University, Belgium) is a clear example. This laboratory started with whole-cell bacterial FAME analysis around 1989 and the resulting FAME database currently contains over 71,000 FAME profiles.

Gas chromatographic FAME analysis is a culture-dependent phenotypic method, implying that the data highly depends on growth and culture conditions. It is shown in literature that these conditions highly effect the gas chromatographic peak areas in a quantitative manner. A small effect was seen in peak presence (qualitatively) (Welch, 1991; Kämpfer, 2002). These findings

had a serious implication on the use of FAME analysis for bacterial identification and, consequently, standard growth and culture conditions are usually adopted, allowing for the comparison of FAME profiles. With their identification system, MIDI Inc. defines different protocols depending on growth atmosphere, isolation source, clinical impact, etc. As such, when analyzing the whole-cell fatty acid content of bacteria and setting up FAME databases, it is very important to maximally annotate the resulting FAME profiles. This is critical and a necessity, as comparing and sharing FAME profiles and doing numerical and computational FAME analysis only makes sense when identical growth and culture conditions were used for the generation of the profiles. Where numerical FAME analysis has been done for many years (O'Donnell et al., 1985; Eerola and Lechtonen, 1988; Kämpfer, 1994; Heyndrickx et al., 1996; Kämpfer, 1994; Vancanneyt et al., 1996), data mining and machine learning studies of the FAME data started only about 15 years ago (Ruggiero et al., 1993; Bertone et al., 1996; Giacomini et al., 2000, 2004), together with the research presented in the previous chapters of this dissertation. In these studies, intelligent identification models were constructed in which mathematical functions were learned to distinguish between different bacterial genera and/or species. However, with our research we showed a bottleneck in FAME research for genus-wide bacterial identification, namely the lack of data due to the private nature of FAME data storage. In order to perform large-scale studies and to upscale data mining and machine learning research on FAME data, cooperation between different institutes has become a necessity. A straightforward solution for this problem lies in the creation of a public FAME database, which is reported in this chapter.

In view of establishing a public FAME repository, it is important to note that a huge gap exists between genotypic and phenotypic databases. Due to massive technological improvements, sequencing has become very cheap and fast, and, as a consequence of this explosive trend and the simple annotation (i.e. set of all meta-information) of nucleotide sequences, deposits in the public nucleotide sequence databases of the INSDC (NCBI, EMBL and DDBJ) have known an exponential growth. The INSDC collaboration already exists for more than 18 years. In contrast, phenotypic methods mostly rely on commercial systems and their identification libraries, with the clear disadvantage of being costly, incomplete and not up-to-date with the current microbial taxonomy. Examples are biochemical testing (API/ID32, Biomérieux), metabolic fingerprinting (Biolog Microplates, Biolog and Vitek2, Biomérieux), MALDI-TOF analysis (MALDI Biotyper, Bruker Daltonics and SARAMIS, Anagnostec), resistance detection (Microscan, Siemens) and, of course, FAME analysis (Sherlock MIS, MIDI Inc.). Many more other techniques exist for phenotypic bacterial identification. While these methods are routinely used in almost every microbiological laboratory, the resulting data are privately stored in databases or computers, or simply published in print only (e.g. scientific journals and Bergey's Manual of Systematic Bacteriology). However, few public databases for sharing these phenotypic data exist. With the massive (high-throughput) generation of phenotypic data, the importance of polyphasic taxonomy (Vandamme et al., 1996) and the current bacterial species definition stating that phenotypic characteristics should agree with the 70% DNA-DNA hybridization threshold for species delineation (Wayne et al., 1987; Rosselló-Mora and Amann, 2001), it remains unclear why phenotypic databases have not been established before. It requires a lot more work

to maximally annotate the data instances, fingerprints or profiles, but with the technological improvements established in the last decade(s), it cannot be a hurdle too high to take. Regarding FAME analysis, a first initiative in this direction was taken by the Swedish culture collection CCUG (University of Göteborg), which makes its generated FAME data accessible on their website, next to genotypic data and other phenotypic data (CCUG, 2009). For gas chromatographic FAME analysis, CCUG follows the method as described by MIDI Inc. (Sasser, 1990). However, strain cultivation is done by different growth conditions, mainly because CCUG deals with a lot of fastidious organisms. For fastidious organisms, they use Chocolate agar in a complex formulation at different atmospheric conditions (CO₂, microaerophilic or anaerobic), the optimal temperature and differing growth durations. For non-fastidious organisms, a blood agar is used aerobically at 37°C and differing growth durations. Peaknaming is performed using the Sherlock MIS (Agilent) FAME standard as reference (Molin, 2004, 2008). CCUG provides their FAME profiles online with for each sample a summary of the retention times, ECL values, relative percentages and peak names for each detected peak. With each sample, also a strain number, a taxon name, a date and some remarks are possibly reported. However, no consistency is found in the manner of annotating the used growth and culturing conditions (personal communication with E. Moore, E. Falsen and K. Molin, CCUG). For identification of their FAME chromatograms, CCUG uses one single library based on the paper of Eerola and Lechtonen (1988). This, in contrast to Sherlock MIS. With the creation of FAME-bank, we take a first step towards a public database for sharing whole-cell bacterial FAME profiles.

6.2 Construction and Content

6.2.1 Database Schema and Implementation

As mentioned in the introduction, comparing, sharing and computationally analyzing FAME profiles requires identical growth conditions. Therefore, storing FAME profiles requires a careful annotation of the data. Because all our FAME profiles were generated using the Sherlock MIS system of MIDI Inc., the implemented annotation is based on that information reported by Sherlock MIS. This concerns the specific growth and culture conditions (growth medium, growth temperature, growth duration and growth atmosphere), depositor information, the used peak naming table and, if available, the identification library used. All information is stored in different tables of an Oracle relational database management system (Oracle Corporation, Redwood Shores, CA, USA). The FAME-bank schema is illustrated by an entity-relationship diagram in Figure 6.1. As the Sherlock MIS system is routinely used at the Laboratory of Microbiology (Ghent University, Belgium), we used this system as a basis for the construction of the database schema. The main table *profiles* covers all general information of a single FAME profile. This comprises strain number, species name, growth conditions, depositor information, provenance and additional information. Herein, provenance is implemented for logging purposes while additional information relates to non-standard profile information. Each FAME-bank profile is assigned a unique FAME-bank accession number ('FB' + eight digits). General peak information is also included and relates to the FAME profiles generated by Sherlock MIS,

namely the total peak area, the total named peak area, the percentage of named peaks, the corrected total named peak area (total amount) and other Sherlock MIS comments related to the quality of the FAME profile. These fields correspond to the information included in the Sherlock MIS FAME profile reports. The peaks of each profile are named by a particular peak naming table and are stored in a different table (*profile peaks*), together with corresponding information such as equivalent chain length (ECL), retention time and summed feature information. Notice that a peak naming table names a certain set of chromatogram peaks, and, thus, allows to express the peak areas as relative values. Stated otherwise, the areas of the named peaks should sum up to 100%. The corresponding peak naming table is saved in the table *peak naming tables* and, if available, the entries of the peak naming table are saved in the table *peak names*. Notice that, we have chosen to implement a 1–1 relationship between a FAME profile and a peak naming table. This implies that a FAME profile may be named by different peak naming tables but the resulting FAME profiles are stored as separate entries in the database. A FAME profile may also be identified by a particular identification library which reports one or more identification name and score for each profile. This identification information is saved in the table *profile identification* and the corresponding identification library is saved in a separate table (*identification libraries*). Furthermore, the name, email address and affiliation of each depositor are stored. In later stages, a login system based on login name and password will be installed for managing personal FAME profiles (more information below). Finally, for each medium, a name, description, owner, provenance and number is saved. These latter two tables are shared with the architecture of the StrainInfo webportal (Dawyndt, 2009).

6.2.2 Data Sources and Quality Control

At 28/06/2009, 3,149 FAME profiles were stored in the FAME-bank database, each uniquely described by a FAME-bank accession number. These profiles originate from two distinct sources. The major source is the private joint FAME database of the Laboratory of Microbiology and the BCCM™/LMG Bacteria Collection (Ghent University) from which 3,018 FAME profiles were exported into FAME-bank. Only FAME profiles from type strains are considered. A second source were 131 FAME profiles resulting from the research of Sikorski and Nevo (2005, 2007). For the latter profiles, see also Subsection 4.3.6.1. Of course, this is only a first step towards a valuable public database. Besides the Laboratory of Microbiology (Ghent University, Belgium), the scope of FAME-bank will also be extended towards other microbiology institutes and laboratories willingly to contribute to this FAME project. In this perspective, a critical issue on data sharing is quality control. FAME profiles are only valuable and useful if these are fully annotated with profiling and peak naming information. Consequently, import of FAME profiles will only be possible if profiles are annotated by growth and culture conditions, i.e. medium, temperature, duration and temperature, and corresponding to a particular peak naming method. More information on this topic is discussed below.

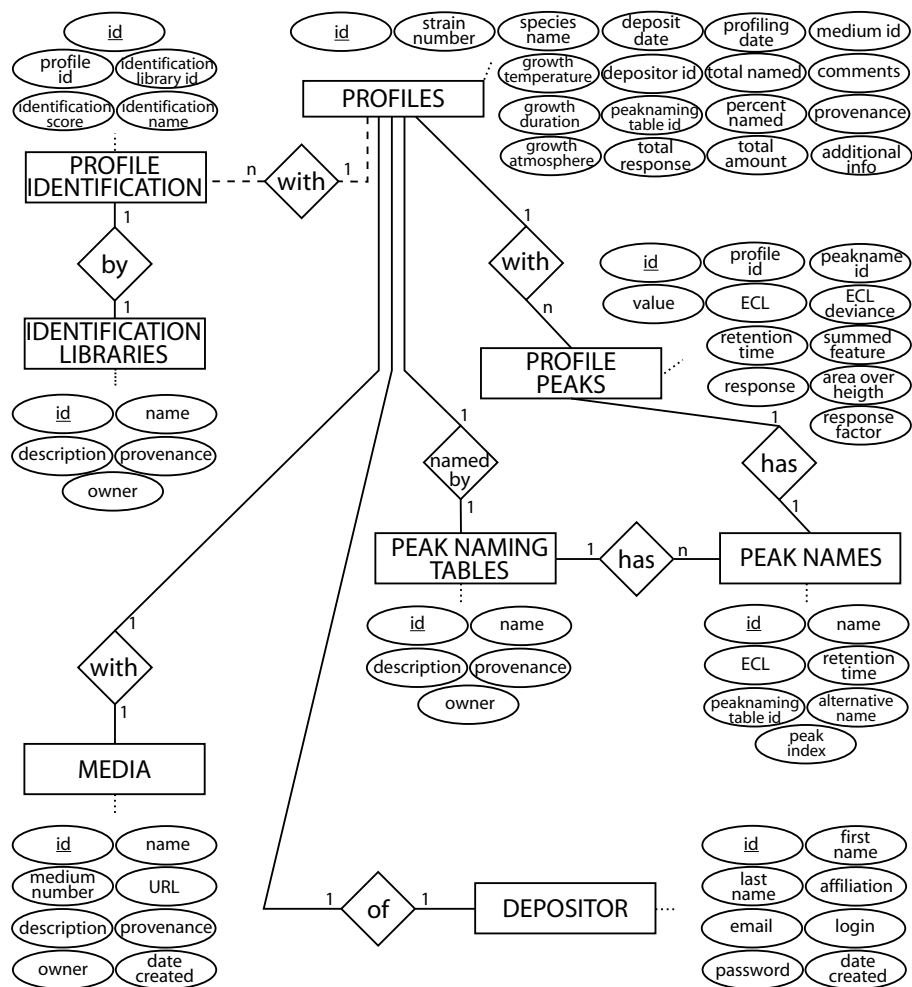


Figure 6.1: Entity-Relation Diagram of the FAME-bank database structure. Rectangular and oval entities correspond to table names and attributes, respectively. Attributes are connected to tables by dotted lines. Diamond squares represent relations between the different entities. Full lines represent at least one relationship between two entries of both entities, while dashed lines correspond to a possible relationship. The primary key of each table is underlined.

6.2.3 Web Interface

A web interface was developed on top of the FAME-bank database. Implementation of the application was based on a Java framework with Apache Struts 2 (web application), Spring (programming technology) and Hibernate (database connection). Hereby, the web interface allows for optimal visualization and searching of FAME profiles.

6.3 Utility and Discussion

6.3.1 User Interface

The FAME-bank main page allows users to easily search for FAME profiles and view the current status of FAME-bank. Background information on the FAME-bank project and a contact address can be found through an additional link. For future development also an upload link and login possibilities are provided (more information below). Two options are provided for

searching FAME profiles: a basic profile search and an advanced search tool. By the basic search, users are allowed to search for profiles based on a strain number, taxonomic name or accession number. A more advanced query builder allows for advanced searching based on peak prevalence, growth conditions, profiling and deposit dates, identifications and depositors. In the current version, users are supported in constructing their personal queries by simple combo-boxes. By submitting a particular search query, all search results are listed by their FAME-bank accession number. Additional information is provided by the fields: strain number, species name, peak naming table, profiling date and deposit date. Alternative sorting is possible through these fields.

Each search result can be viewed into more detail. All information about each profile is grouped and can be accessed by a tabbed header. The different tabs cover general information (overview tab), peak information (profile peaks tab), strain information (according tab), information on growth and culture conditions (sample conditions tab) and, if available, identification of the FAME profile (identification tab). The general information covers the accession number, deposit and profiling date and the name of the depositor. The profile peaks are visualized in a table and by two diagrams. In a first figure, the percentage of each peak is plotted in a bar diagram against its equivalent chain length. The second figure shows the different relative peak areas in a pie chart. If the original chromatogram is available, this figure will be included in this tab. Strain information covers the bacterial strain number and its species name. Links are provided to the StrainInfo.net webportal, in case more information regarding the particular strain or species is desired (Dawyndt, 2009). This mainly concerns a list of all strain numbers of a particular species, equivalent strain numbers, strain history and corresponding 16S rRNA sequences and literature. Growth temperature, duration, atmosphere and medium are reported in the sample conditions tab. If the FAME profile was identified by an identification system, the corresponding identification library together with the name and score of each identification result are reported in a final tab. Screenshots of the different tabs are visualized in Figure 6.2 for FAME profile FB00000647 which relates to the strain *Bacillus subtilis* LMG 7135^T. The corresponding Sherlock MIS FAME profile report of this strain is shown in Figure 2.11.

6.3.2 Use, Benefits and Future Development

With the presented database and web application, it is currently possible to search and analyze 3,149 fully annotated FAME profiles. Where the CCUG culture collection (University of Göteborg, Sweden) has put the FAME data of their strains online, searching online FAME profiles in a public FAME database was not possible before the launch of FAME-bank. In the current alpha version of FAME-bank, the web functionality only makes searching specific FAME profiles possible.

From the conclusions in the previous chapters, it is clear that private FAME databases and the absence of a public FAME database restricts the size of the generated data sets and, thus, the scope of the performed FAME research. Therefore, in the beta version of the web application, additional features will be implemented, allowing for an extended numerical and computational FAME analysis. This includes similarity searching, numerical taxonomy and identification of

FAME profile FB00000647

overview | profile peaks | strain information | sample conditions | identification

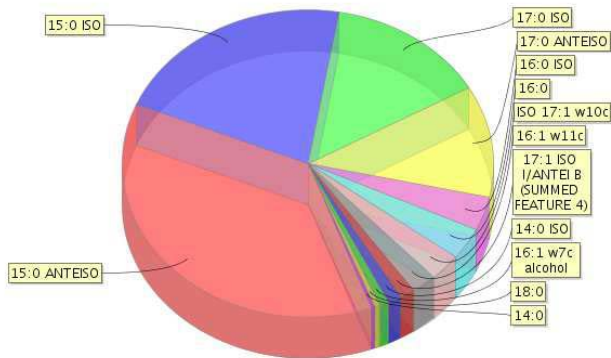
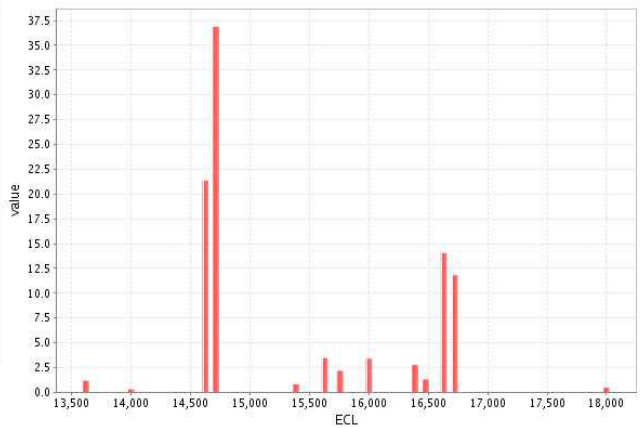
accession number **FB00000647**
 deposit date **26 Jun 2009**
 depositor [Laboratory of Microbiology and BCCM/LMG Bacteria Collection](#) (Ghent University)
 profiling date **05 Sep 2005**

(a) General overview

FAME profile FB00000647

overview | profile peaks | strain information | sample conditions | identification

peakname	value
14:0 ISO	1.19
14:0	0.32
15:0 ISO	21.37
15:0 ANTEISO	36.88
16:1 w7c alcohol	0.81
16:0 ISO	3.46
16:1 w11c	2.18
16:0	3.4
ISO 17:1 w10c	2.76
17:1 ISO I/ANTEI B (SUMMED FEATURE 4)	1.3
17:0 ISO	14.04
17:0 ANTEISO	11.81
18:0	0.47



(b) Profile peaks

FAME profile FB00000647

overview | profile peaks | strain information | sample conditions | identification

strain number **LMG 7135^T**
 species name ***Bacillus subtilis***
 ⓘ this strain number is reported by StrainInfo as [Bacillus subtilis subsp. subtilis](#)

(c) Strain Information

Figure 6.2: FAME profile FB00000647 of *Bacillus subtilis* LMG 7135^T. Screenshots are given for the different information tabs.

FAME profile FB00000647

overview	profile peaks	strain information	sample conditions	identification
medium name LMG Medium 185 growth atmosphere aerobic growth temperature 28C growth duration 24H				

(d) Sample Conditions

FAME profile FB00000647

overview	profile peaks	strain information	sample conditions	identification						
<table border="1"> <thead> <tr> <th>library</th> <th>identification</th> <th>score</th> </tr> </thead> <tbody> <tr> <td>TSBA50</td> <td>Bacillus subtilis</td> <td>0.873</td> </tr> </tbody> </table>					library	identification	score	TSBA50	Bacillus subtilis	0.873
library	identification	score								
TSBA50	Bacillus subtilis	0.873								

(e) Identification

Figure 6.2 continued.

FAME profiles by the machine learning models presented in this dissertation. In a first step, downloading FAME profiles will be made possible. Herein, the XML markup language allows for easy sharing and parsing of the data. We foresee to launch this new version soon. Ultimately, analysis will become possible from strain level to genus level and higher taxonomic ranks.

Where the first step towards a public FAME database is taken by the Laboratory of Microbiology (Ghent University, Belgium), a large qualitative database covering a wide spectrum of the bacterial landscape can only successfully be achieved by a community effort. At present, the number of fully annotated FAME profiles is rather small, but with only a few collaborations this number can grow extensively. In order to make other microbiologists participate in this FAME-bank project, an implementation of a FAME profile upload scheme is currently in development. Sharing data is, however, a critical issue, especially in the context of providing data of high quality. A good reference point are the INSDC nucleotide databases (INSDC, 2009) that contain a large number of sequences of low to bad quality. It has become that bad that new databases are established for providing quality checked nucleotide sequences available from the primary nucleotide databases. An example is the SILVA database which provides quality checked ribosomal nucleotide sequences (Pruesse et al., 2007). In establishing a FAME database, it is important to bear this data quality issue in mind. Consequently, as a main quality control, only fully annotated FAME profiles will be allowed for import in the database. This implies that all imported FAME profiles need to be provided with growth and culture conditions and a particular peak naming table. These restrictions are imposed as a comparison of these profiles is only meaningful when the FAME profiles can be put in their proper context. Moreover, as FAME peaks are valued by a relative percentage over the total named area, they should sum up to 100%. This property can be used as an additional quality control measure. Globally, it is even advisable to define a minimum information standard for a gas chromatographic FAME profile experiment, such as the minimum standards pursued by the MIBBI or Minimum Information

for Biological and Biomedical Investigations project. This project aims to increase the visibility of projects developing guidance for the reporting of aspects of biological and biomedical science; to encourage collaborative development between such projects and, where appropriate, to avoid duplication of effort or competition; and to promote the adoption of consensus guidance on reporting by journals and funders. Different minimal standards are already defined such as MIAME (Minimum Information About a Microarray Experiment), MIGS (Minimum Information about a Genome Sequence), MINSEQE (Minimum Information about a high-throughput SeQuencing Experiment), etc. (MIBBI, 2009). Currently, different import procedures are written and tested that mainly focus on an automated parsing of the text-annotated Sherlock MIS reports. Herein, it will be possible to import the different FAME profiles in a single step (or bulk upload). Moreover, as an additional control system, a login system is installed for which the user needs to register by giving his or her name, email and affiliation. In this way, the (blind) upload of FAME profiles of low quality is prevented. In the future, this login system will allow users to manage and analyze personal FAME profiles.

An interesting and popular feature available in many databases is matching data instances against the present database entries. Also in our case, finding the most similar FAME profiles in the database would be a valuable feature. In the beta version, a profile matching tool will be provided. Remark that this similarity calculation requires a totally different approach than, for instance, nucleotide matching, which is typically achieved by heuristical algorithms. For profile matching, a large choice of similarity measures can be implemented (e.g. Canberra distance, Euclidean distance, Mahalanobis distance etc.). Moreover, with the availability of more FAME profiles and a larger taxonomic scope, it will be possible to extend the research on developing intelligent FAME-based bacterial species identification models as described in this work. It is our purpose to make these models available to the community as an alternative FAME-based identification tool. In view of profile matching and identification, we will thus only allow to find similar FAME profiles among the different FAME-bank entries by profile matching and by intelligent computational models. It is, however, very important to mention and underscore that it is absolutely not our aim to re-identify stored FAME profiles by any commercial identification library. It is not our goal to fulfil the identification job of any commercial identification system. The core business of this FAME project is to make phenotypic data freely available and, as such, allow for a more large-scale and extended FAME analysis.

6.4 Conclusions

The number of public phenotypic databases is lagging far behind the number of genotypic databases. However, phenotypic analysis methods are routinely used in bacteriology and result in a massive amount of data. Moreover, besides the genotype, the phenotype is a major player in polyphasic taxonomy. It remains unclear why not even a small subset of these data has yet been shared online in a public database. With the FAME-bank project, we have started one of the first public phenotype databases by sharing whole-cell FAME profiles. Until now, FAME profiles were only privately stored on personal computers, servers or databases, making the scope of bacterial FAME studies limited. By establishing a public FAME database, searchable

without any further restrictions, we aimed at a public framework for FAME profile sharing and a more large-scale numerical and computational FAME analysis. A user friendly web application makes FAME-bank accessible world-wide. Besides making public FAME profiles searchable, the initial FAME-bank project seeks extensive further development that will allow to take FAME analysis studies to a higher level. For instance, by the development of custom identification libraries in any kind of niche of interest. Without becoming too idealistic, we attempt to make microbiologists enthusiastic about the open-access concept, especially when focusing on phenotypic data. With the improvements in (high-throughput) analysis technology and computer systems established in the last decades, networking becomes more and more important and critical. In this light, microbiology should not restrict its focus to nucleotide and protein databases but also reckon with phenotypic data.

6.5 Availability

FAME-bank is publicly and freely accessible by the website <http://www.fame-bank.net>. Searching and downloading FAME profiles is without any restriction, uploading of data will require a user to register for a FAME-bank account.

PART IV

GENERAL SUMMARY

CHAPTER 7

General Conclusions

In this chapter, the main conclusions of our research on FAME-based bacterial species identification are summarized. A brief overview is given of the different experiments, research strategies and results, together with a critical note on the strengths and weaknesses of the followed approaches. Finally, the main contributions to the scientific community are reported.

The main goal in this work was to investigate how the identification of bacterial species, based on FAME data, could be improved by the application of machine learning techniques. Machine learning allows for an intelligent computational identification and, given the limited resolution of FAME analysis for species identification, machine learning can maximally exploit the information and patterns present in the data. Hereby, machine learning allows for a more reliable identification than an identification solely based on similarity measures.

In general, the use of machine learning techniques for bacterial identification is still limited and, thus, an interesting field to exploit and investigate. This is especially true when focusing on phenotypic data. However, machine learning is still not a very popular field in microbiology. One reason is that machine learning techniques requires mathematical knowledge and, with regard to parameter optimization and good performance estimation, these techniques are not easily implemented and executed. Also, phenotypic data are mostly not publicly available and accessible to the community, making it hard to perform extended numerical and computational research. This also holds for FAME data. Research on FAME-based bacterial identification by machine learning techniques is still very limited. Moreover, an approach for up-to-date bacterial identification had not been investigated before, making it worthwhile to investigate a FAME-based species identification by machine learning techniques.

7.1 Data Sets

We limited our computational research to the analysis of the three bacterial genera *Bacillus*, *Paenibacillus* and *Pseudomonas*. These genera were chosen for three main reasons. First, the two genera *Bacillus* and *Paenibacillus* are phylogenetically closely related, though distant to the third genus, *Pseudomonas*. Second, for each of these genera, a reasonable number of FAME profiles was present in the joint database of the Laboratory of Microbiology and the BCCM™/LMG Bacteria Collection (Ghent University). Third, a profound expertise was avail-

able at the Laboratory of Microbiology. A manual procedure was installed for the creation of FAME data sets covering as many species of each genus as possible. This procedure was mainly related to the removal of FAME profiles of bad quality, resulting from non-standard growth and culture conditions, with a bad peak composition, of outliers and of profiles belonging to non-valid bacterial species. The resulting data sets were merged in a global genera data set and two data sets regarding different plant-pathogenic *Pseudomonas* strains were also constructed. For the three genus data sets, we initially chose to work with as many profiles as were present in the FAME database. By aiming at a genus-wide identification, it was necessary to generate hundreds of additional FAME profiles. A minimum of three FAME profiles per species was selected for data set creation, even though we kept in mind that the construction of identification models relying on such limited data is far from ideal. Results will become more objective and reliable when more FAME profiles will be included in the data sets. Moreover, extending research towards more genera will require the generation of a multitude of FAME profiles.

7.2 Data Analysis and Machine Learning

Three main research steps were investigated in this work. First, to gain initial insights in the FAME data, in what manner the FAME profiles of the different species could possibly be distinguished from each other, or how they were related, we performed a short data analysis. The second step handled a very straightforward problem in machine learning, namely the genus-wide classification of bacterial species based on the FAME data. The main purpose here was to investigate how FAME-based identification could be improved by several machine learning techniques for the species of different bacterial genera. In a third step, we focused on learning from FAME data in a taxonomic framework and on how learning could contribute to putting the identification results in this taxonomic context. For each of these three research topics, the goals, main achievements and general conclusions are reported.

7.2.1 Data Analysis

The main goal in this work focused on the research question of how to improve a particular identification problem by the application of machine learning techniques. In Chapter 3, a preliminary data analysis resulted in knowledge on the data composition, data patterns and relations between the different classes. Hereby, it was possible to analyze how well machine learning techniques could perform on the present data. Three particular standard approaches were investigated: the calculation of average FAME profiles, profile clustering and principal component analysis. These three analysis methods were performed for the three genus data sets and the latter method was also applied on the joint genera data set and the plant-pathogenic *Pseudomonas* data sets. From the average FAME profiles, it could clearly be concluded that the spectra contained core-genus, species-specific and strain-specific peaks which over the different species within one genus resulted in a broad spectrum of peaks. Average peak values and standard deviations showed possible species discriminations in a qualitative and/or quantitative manner. For clustering, we performed a basic peak and/or species clustering of the data, and

a TaxonGap analysis. Both methods showed close relations between different species of one genus, possible only allowing for a discrimination of species groups. Also, for one species, different FAME subgroups were found. Hereby, we further confirmed the results of different numerical studies on FAME data such as Stead (1992), Kämpfer (1994), Heyndrickx et al. (1996) and Vancanneyt et al. (1996). These findings also indicated that the calculation of similarity measures between FAME profiles will not always result in reliable identification results. The clustering results revealed a possible further restriction of the power of machine learning techniques for the classification of bacterial species, where the power was already confined by pursuing a classification of a large number of classes of which most classes were represented by a small number of data instances. Principal component analysis clearly showed a correlation between the many FAMEs. Biologically, this correlation could be related to the biochemical synthesis of fatty acids. Skree plots showed that the three genera could almost perfectly be separated from each other. For the species of each genus, the analysis confirmed the findings of the other data analysis methods, in that a lot of species were related to each other in FAME data, making it hard to discriminate between them. From biplots of the first principal components, we could also conclude that the species of the genus *Pseudomonas* were more related to each other than the species of the genera *Bacillus* and *Paenibacillus*. This genus also corresponded, however, to a lot more species. Finally, principal component analysis of the plant-pathogenic data sets also revealed highly related FAME patterns between plant-pathogenic *Pseudomonas* species. And, from first sight, the discrimination of plant-pathogenic *Pseudomonas* species from non-pathogenic *Pseudomonas* species seemed not very straightforward. In general, from these data analysis experiments, we concluded that the power of machine learning research for genus-wide FAME-based species identification could be confined by data pattern similarities between numerous species of a single genus. However, the trade-off between the large variability in the data and the presence of correlated peaks could possibly allow for a more flexible learning and model construction.

7.2.2 FAME-based Bacterial Species Classification

The data analysis performed in Chapter 3 clearly showed that species identification based on FAME data is not a straightforward task. Machine learning techniques, though, allow for an intelligent computational analysis by learning flexible species boundaries between the overlapping FAME data clouds. Hence, a challenging task was to investigate how machine learning techniques are capable of generalizing over the data of the different species and if the identification could be improved by this approach. The techniques artificial neural networks, support vector machines and random forests were considered for identification of the species of the three genera. Where species identification was already performed on a limited number of species of a small but different number of genera, we focused on species identification within one genus. This setup of genus-wide computational identification in a taxonomic setting was not investigated before. In a first setting, we applied artificial neural networks for the identification of *Bacillus* species. The main goal was to investigate different experimental strategies for up-scaling the machine learning research to different other genera. The performance on balanced

data sets was compared to that on imbalanced data sets, simple validation was compared to cross-validation and the combination of two different neural network activation functions were evaluated. Globally, a good identification performance was obtained for *Bacillus* species identification using imbalanced data sets and artificial neural network models validated by stratified cross-validation. For the considered neural network activation functions, we found that these functions should be selected from a set of activation functions. As artificial neural networks are quite parametrized, a lot more setups with different other parameters could be considered. Herein, different learning algorithms with different parameters and different architectures could be chosen. It was, however, our aim to investigate how an experimental setup should be installed in order to achieve a good and reliable species identification. Thus, with these experiments, we allowed for a promising further investigation of artificial neural network-based identification of bacterial species in different other genera. However, it is important to remark that due to a restrictive number of FAME profiles for a large number of species, better and more reliable results will be obtained by the inclusion of more data per species, which is supported by a better generalization obtained for each species. Finally, the data analysis experiments in Chapter 3 implied that the resolution of FAME data could possibly confine the performance of machine learning techniques. This is nicely confirmed in the neural network experiments for *Bacillus* species identification by focusing on two groups of closely related species, the *B. cereus* group and the *B. subtilis* group. Misidentified FAME profiles of species of those groups showed to be mainly identified as a member species of the corresponding group.

With the results of the *Bacillus* species identification experiments in mind, we took our machine learning research to a broader spectrum with the evaluation of three machine learning techniques for species identification within three bacterial genera. For these experiments, two strategies were investigated: a stratified setup with identification from genus level to species level and a straight species identification by which species of the three genera are distinguished using a single model. A better identification performance was achieved with the stratified approach for all three techniques. For this strategy, we could further conclude that the three machine learning techniques resulted in a nearly perfect genus identification. At species level, a moderate to high identification performance was achieved, keeping the limited discriminative power of FAME analysis in mind as well as ongoing discussions about the taxonomic positions of different *Bacillus*, *Paenibacillus* and *Pseudomonas* species. Data analysis experiments on the FAME data of the genus *Pseudomonas* showed that the resolution of FAME analysis was very limited. This was clearly observed in the different machine learning experiments that were confined in their power to discriminate between the different species. Generally, in the different classification experiments, random forests proved to be the best technique for identification within each of the three genera. Random forests is also very promising for upscaling the experiments to a wider bacterial spectrum as it has several advantages as opposed to artificial neural networks and support vector machines. Namely, random forests are robust against overfitting, correspond to a short computation time and require the optimization of only a small number of parameters. Hence, it is preferably to consider random forests in a stratified strategy as the machine learning technique for future experiments. The identification performance achieved with machine learning techniques was compared to those obtained by the commercial identification

system Sherlock MIS (MIDI Inc., USA). Even though the identification libraries of this system comprise a lot more genera and species, for the species of the three genera considered in both systems an improved identification was obtained by the machine learning approach. However, a more reliable comparison will be possible when more genera, species and strains become investigated. Identification was also evaluated by independent data sets. In the case of the *Bacillus simplex* data set and, when compared to the commercial system, identification could be greatly improved. With the analysis of the genus *Pseudomonas*, additional experiments were performed on two plant-pathogenic data sets. plant-pathogenic species could significantly be distinguished from non-plant-pathogenic species. In the case of a discrimination between the different plant-pathogenic species only a moderate performance was achieved. The presented machine learning approach for FAME-based bacterial species identification has several advantages. Machine learning techniques allow to learn from the data by generalizing over the data within one class or species. This learning is advantageous for non-linear separability problems like we dealt with in this study. This in contrast to comparing data points by similarity calculations. Another important advantage of the machine learning approach relates to the monthly changing bacterial taxonomy that requires identification methods to update their libraries. This problem is easily met by machine learning techniques by rapidly updating the data sets and by allowing a rapid retraining of the identification models. Moreover, in a stratified setting not all genus models need to be retrained when changes occur in a particular genus. However, disadvantages also exist. The results will become more reliable when more FAME profiles per species are included in the data sets. Besides more profiles, a better intra-species heterogeneity and, thus, generalization will be attained when also more strains are included. In this regard, it is important to emphasize that the current practice of publishing species with only a single strain has little discriminatory value. When pursuing a genus-wide species identification, more validly published species need to be included in the data sets as for *Bacillus* and *Paenibacillus* only half of the validly published species were considered. Increasing the number of strains, species and genera is, however, not as straightforward as presented due to the inability of many bacteria to grow under standard conditions or even to be culturable. Besides a genus-wide identification, we also worked in a taxonomic framework. No knowledge was integrated in the multi-class identification models regarding taxonomic relationships. Another important remark concerns the strategy of evaluating model performance based on the winner-take-all rule. By solely considering the highest output value, the other output scores and their differences are ignored. Besides this, the value of the highest output score is also ignored. Both sources can correspond to knowledge of major importance in the evaluation of the model performance. In the stratified setting, this would also result in a better interpretation of the identification results. Nevertheless, on the scale of the research performed in this study, no computational FAME analysis by means of machine learning techniques was performed so far.

7.2.3 Phylogenetic Learning

Globally, the different data analysis experiments and machine learning experiments showed that it was hard to distinguish most species from each other. This again underscores the fact that

the resolution of FAME analysis does not allow for a clear species discrimination but is rather related to the identification of species groups. A clear example was reported in the artificial neural network experiments on the species of the *Bacillus cereus* group and *Bacillus subtilis* group. A look at the misidentifications of these species showed that the corresponding identification was mainly due to a species of the same species group. The low performance on the identification of the different *Pseudomonas* species could also be attributed to closely related FAME patterns, as visualized in the data analysis experiments. Hence, in a following research topic, we investigated how we could integrate taxonomic and phylogenetic information in our machine learning models in order to further analyze the limited FAME resolution for species discrimination. As bacteria are hierarchically ordered in a taxonomy, we approached this integration by learning from these taxonomic relationships. Herein, the genus *Bacillus* was chosen as model genus. In the field of machine learning, one method for hierarchical classification is the approach of binary tree classifiers. A binary tree classifier typically builds a tree from the data of interest, constructs two-class classifiers on each node of the tree and identifies new data by putting these along the tree until identification in a particular leaf. We investigated this approach by the construction of a FAME tree using supervised clustering of the random forest identification scores. Due to the small sample sizes of several classes, agglomerative clustering would result in an unreliable initial tree reconstruction at leaf level. Divisive clustering was consequently chosen and analyzed in a small case study using a limited number of *Bacillus* species of which some were closely related. Good results were obtained and, with respect to classification, a performance was obtained similar to that obtained by flat multi-class classification. When scaled to all present species, this method was, however, infeasible due to unreasonable computation time. An alternative approach was found in the current species definition. While DNA-DNA hybridization values are not widely shared, the 16S rRNA gene sequence correlates with DNA-DNA hybridization and allows delineation of most bacterial species. Moreover, quality-controlled 16S rRNA sequences are freely available in the SILVA sequence database. Therefore, we investigated how 16S rRNA gene data could be combined with our FAME data. We chose to tackle the binary tree approach in view of the 16S rRNA data and the taxonomic framework. In other words, we inferred phylogenetic 16S rRNA gene sequence trees and used these trees as templates for binary tree classification. The NJ and UPGMA treeing methods were chosen with the Jukes-Cantor evolution model. It is important to mention that different treeing methods exist, each with their strengths and weaknesses, and a lot of discussion is going on about which treeing method is best for phylogenetic analysis. With a focus on binary tree classifiers that use rooted trees, we, therefore, did not consider a single tree but investigated the NJ and UPGMA methods for inference of rooted trees. In short, 16S rRNA gene sequences were used for defining the different classification tasks, and the FAME profiles were used for classification of bacterial species as defined by these tasks. Importantly, this approach of combining two data types for binary tree classification was not considered before and we called this approach phylogenetic learning. By the hierarchical identification scheme, we were able to analyze the identification results along a path of species groups and single species. For classification purposes, the method showed to be less performant than flat multi-class classification. However, the latter strategy does not incorporate any relevant phylogenetic

information and cannot provide detailed information on the resolution of FAME data to discriminate between related species and groups of species. Moreover, the identification of species incorrectly identified by flat multi-class classification was improved by phylogenetic learning. The performance of phylogenetic learning for both treeing methods was similar. It is clear that phylogenetic learning has distinct advantages when compared to flat multi-class classification. Moreover, statistical analysis of the identification scores at each node of the tree allowed to develop a system to exploit and visualize the performance of the different binary classifiers. The Wilcoxon rank-sum test on the random forest identification scores allowed to visualize the resolution of FAME analysis for species identification in the phylogenetic framework. A three-colour highlighting scheme was chosen for visualizing branches and nodes corresponding to p -values, according to two significance levels. Significance levels were adjusted following the Bonferroni correction, similar to its application in the statistical problem of multiplicity. Though the Bonferroni correction is considered a very conservative method, we rather allowed a more conservative penalisation than highlighting distinct FAME profiles as being similar. Experiments were performed for the three genera and identical groups were found in both the NJ and UPGMA trees. This highlighting information easily lends itself for enhancing the identification report of our machine learning models. Moreover, this approach could be used on any type of FAME-based identification method. Briefly summarized, good strategies were found for knowledge integration into FAME-based bacterial species identification models and these allow for a more integrated research by combining different data types. Moreover, the presented approach is easily generalized and extended towards a large bacterial spectrum, given the presence of high quality 16S rRNA gene sequences. In this study, we only considered the 16S rRNA gene, but it is clear that different other data could possibly be considered.

7.3 FAME-bank.net

In aiming at a good generalization and striving for a genus-wide identification, machine learning research was clearly restricted due to a lack of FAME data. More FAME profiles of more bacterial strains are needed for a better intra-species heterogeneity and more validly published species and genera are needed to extend the bacterial scope. Though, as with many phenotypic methods, FAME profiles are stored in private databases and are not shared among scientists. Analytical, numerical and computational research consequently remains dependent on research performed at single institutes and is further restricted to specific environmental, clinical and industrial niches. Generally, the resulting database of a single institute is taxonomically seen restricted, being an obstacle for scientific research. The ultimate solution lies in a public FAME database. While phenotypic analysis such as FAME analysis knows a long history and is routinely used in microbiology, it remains a question why the scientific community has not been sharing phenotypic data. In this work, we took the first step towards a public FAME database for sharing bacterial FAME profiles. A database architecture was constructed with a web application on top. All imported profiles were fully annotated with all required information for enabling a good interpretation of the data. This corresponds to growth and culture conditions, the used peak naming table and, possibly, one or more identification libraries. Deposit

and general profile information is also shared. Currently, the web application is still in its alpha phase and is, therefore, restricted to executing search queries. Nonetheless, once the beta version is launched, scientists will be able to import and download FAME profiles and perform computational FAME analysis. To ensure data of a high quality, only fully annotated profiles will be allowed for import.

With this FAME-bank.net project, we aim for an extended FAME research. However, the whole project stands with the willingness of the community to share its data. If the taxonomic scope becomes extended and the number of profiles grows, the performed research can be updated, leading to an extended research and more reliable identification results. Also, scientists will be able to construct custom identification libraries and perform FAME analysis in an extended bacterial scope. Hence, with this database we hope to allow for an extended and large-scale numerical and computational FAME analysis. In summary, by cooperation and extending this research in the future, the automated FAME-based identification tool for bacteria will become most valuable in microbiology and many related fields.

7.4 Main Contributions to the Community

The main goal of this dissertation was to evaluate how FAME-based bacterial species identification could be improved by the application of machine learning techniques. We evaluated three popular machine learning techniques on three different bacterial genera. Different setups were investigated regarding data set creation, validation, parameter settings and identification strategies. Moreover, knowledge integration was also investigated with the combination of FAME data with 16S rRNA gene data. Therefore, this work may contribute to the field of bacteriology by setting a first machine learning framework for bacterial species identification that can easily be extended towards a larger taxonomic scope and, eventually, other phenotypic data. The contribution to the machine learning community is rather small, except for describing an alternative approach to combine different data types in a single hierarchical classification model. In the following, the main contributions are briefly summarized:

- Data analysis was performed by different methods on a large number of species and strains and extended the numerical studies published in the last decade.
- Machine learning techniques improved bacterial species identification for the three genera. By learning from the data, machine learning techniques generalize over the data and result in intelligent species boundaries. Compared to similarity scoring, a more reliable identification was obtained.
- With the machine learning approach, we could easily tackle the problem of keeping pace with the monthly changing bacterial taxonomy by the ability of rapidly updating data sets and by retraining the models.
- With FAME analysis used as a first-line identification method, the machine learning framework is easily plugged into laboratory information management systems.
- Integration of taxonomic and phylogenetic knowledge in the models was achieved by combining different data types. Identification results can hierarchically be analyzed, al-

lowing for a better interpretation of the resolution of FAME analysis for species and species group discrimination.

- Post-processing of this technique allows to put the resolution of FAME data and bacterial species identification in a taxonomic context.
- The creation of the FAME-bank.net search portal can be regarded as a steppingstone towards a public sharing environment of bacterial FAME profiles. By this web portal, we ultimately aim at a deeper and wider data analysis and machine learning research. In this way, a further convergence becomes possible between the fields of bacteriology, computational analysis and machine learning.

CHAPTER 8

Future Perspectives

8.1 General

The most obvious future task is the extension of the computational research towards more bacterial genera and species. When aiming at a genus-wide identification strategy, a maximum number of species need to be integrated. At short term, this implies that many more FAME profiles need to be generated. Moreover, to achieve a good intra-species heterogeneity and, thus, a good model performance, a lot more strains need to be included as well. Regarding the integration of more bacterial genera, the machine learning task for prediction at genus level could become challenging, especially in the case of closely related genera, such as *Bacillus sensu lato*. A possible strategy here is to further stratify the identification strategy. However, with the described stratified strategy, the main problem lies in the use of only the highest output identification score of each FAME profile for directing the identification of each profile from genus level to species level. When further considering this strategy for FAME-based bacterial species identification with an extended scope at genus level, this approach could fail and an alternative scoring or weighing mechanism will become necessary to attain more correct identification results. A possible solution could consider the identification scores of the FAME profiles for their identification at both the genus and the species level.

For model evaluation in the multi-class setting, we applied the winner-take-all rule on the identification outputs to prevent the application of any arbitrary thresholding. The application of this rule possibly results in the loss of important information, which relates to the value of the corresponding output but also to the differences between the identification scores. Another future task should investigate alternative strategies for solving this problem.

Three machine learning methods in basic settings were investigated in this dissertation. Consequently, some techniques could be evaluated more deeply. This mainly comprises artificial neural networks and support vector machines. For neural networks, different architectures, different learning algorithms and different activation functions can be considered. In the case of support vector machines, different other kernels exist, together with different optimization techniques. An extended evaluation of these methods is possible. Regarding support vector machines, an interesting research topic could also be the development of custom kernel functions. Of course, many other machine learning techniques exist that could possibly result in a

promising identification. One example is boosting. Therefore, evaluation of additional techniques could consequently be another future task. The perspectives proposed in this paragraph are of main interest for machine learning purposes. It remains, however, questionable whether this research should focus on an endless evaluation of methods and settings.

In regard of plant pathogenicity, it would be interesting to further evaluate the plant pathogen vs. non-plant pathogen data analysis. Research could focus on the discovery of possible biomarkers and on the distinction between mushroom and plant pathogenicity. The analysis of different other genera, which are known to comprise plant-pathogenic species, would also be a valuable research topic.

Besides a further extension of the scientific research, it would also be highly interesting to use the developed identification methods for routine bacterial species identification, next to the commercial system Sherlock MIS. This could be made possible by the development of plugins for the LIMS system used at the Laboratory of Microbiology and the BCCM™/LMG Bacteria Collection.

8.2 Data Sets

Not all FAME profiles present in the joint FAME database of the Laboratory of Microbiology and the BCCM™/LMG Bacteria Collection are suitable for computational data analysis. Some profiles are not resulting from standard protocols, contain less than three or four FAME peaks, or even contain errors. For appropriate use of the database and, especially, for the selection of high-quality data sets, it is recommended to perform a quality check of the whole database or to provide additional fields or features to easily visualize the quality of the present profiles. The presented data selection procedure was a manual, tedious task. Therefore, it is recommended to develop an automated selection system in the LIMS system of the Laboratory of Microbiology, in which the proposed quality labels could play a crucial role. Of course, a final manual inspection will remain necessary.

In view of the expansion of the FAME data sets, it is important to mention that not all bacterial strains of a particular genus or species grow by standard growth and culture conditions. Protocols are already defined by the commercial Sherlock MIS system (Midi Inc., USA), though not all bacterial species are included in their backend libraries. When extending research towards numerous genera and species, agreements on conditions deviating from the standard are needed for an objective interpretation, analysis, comparison and sharing of the FAME profiles and the corresponding studies.

8.3 Data Analysis

Principal component analysis showed that about 90 to 95 percent of the variability in each FAME data set could be described by about five principal components. Together with the fact that FAME peaks are correlated, this knowledge indicates that certain FAME peaks in the data sets are redundant. This redundancy could have a serious impact on our classification exper-

iments. Therefore, one major future task is to evaluate the effect of feature selection on the performance of the different machine learning techniques. This study was not considered in this work, though it might have a positive effect on classification.

8.4 Phylogenetic Learning

An initial study was performed on the integration of taxonomic and phylogenetic information in species classification models, with the goal of a hierarchical analysis of the resolution of FAME data for species discrimination. A major future perspective is to extend and explore the different possibilities and approaches regarding this research. We mainly focused on the construction of binary tree classifiers built on a phylogenetic tree inferred from 16S rRNA gene sequences. Because binary tree classifiers start from rooted trees, we only considered NJ and UPGMA trees. Different other methods exist, although which method to prefer for phylogenetic tree inference has already been an item of discussion for many years. Besides the two considered methods, two other popular methods are the maximum parsimony and the maximum likelihood methods. These methods, however, infer unrooted trees. In the approach of binary classifiers, only rooted trees are handled. Consequently, rooting of these unrooted trees should be performed, most preferably by the integration of an outgroup. A second alternative to the considered trees are the exploitation of established phylogenies. Examples are the phylogenies resulting from the All-Species-Living-Tree project or the Taxonomic Outline of Bacteria and Archaea. The aim of the All-Species-Living-Tree project is to reconstruct a single 16S rRNA tree harboring all sequenced type strains of the hitherto classified species of Archaea and Bacteria (Yarza et al., 2008). The Taxonomic Outline of Bacteria and Archaea is a comprehensive taxonomy of the type strains of Bacteria and Archaea, based on the 16S rRNA gene phylogeny (Garrity et al., 2007). Besides this, machine learning research in unrooted trees could also be a challenging research topic. In the treeing methods, we chose to calculate the distances between the 16S rRNA sequences by the Jukes-Cantor evolution model. Different other evolution models exist and could be of possible interest for further research. Examples are the Kimura-2 model and others. Further research could also focus on other integration approaches than the binary tree classifier. With the popularity of kernel methods, a possible alternative could be to model 16S rRNA phylogenetic relationships by kernel functions. Of course, this whole research could also be oriented towards other data than 16S rRNA gene sequences. In this perspective, an interesting identification scheme is multi-locus sequence analysis and, in the long term, whole-genome sequence analysis. Generally, the concept of phylogenetic learning could also be considered on other data than FAME data.

The described approach of phylogenetic learning was ultimately developed for the visualization of the FAME resolution in a taxonomical framework. We applied statistics on the results of the many binary classifiers that were trained on the nodes of the considered phylogenetic trees, calculated p -values by statistical testing and applied a highlighting scheme defined by two significance levels adjusted by a Bonferroni correction, similarly as applied in the statistical problem of multiplicity. Before extending this research to a wider bacterial scope, this

approach should initially be further fine-tuned and evaluated, specifically in view of the multiplicity problem and the adjustment of p -values and/or significance levels.

The combination of different data types could also yield different possibilities for machine learning research. In the case of FAME data and 16S rRNA sequence data, one could evaluate how FAME profiles correlate with 16S rRNA sequences, if FAME profiles could be predicted given the 16S rRNA sequences (and conversely) and the approach of data fusion. With unrooted trees, an interesting suggestion for further machine learning research is to investigate classification in this phylogenetic framework and how this approach could relate to the research of this study.

In this topic on phylogenetic learning, we initially focused on the construction of a FAME tree by supervised clustering of the FAME profiles. This approach could also be promising for the evaluation and integration of the resolution of FAME analysis for species discrimination. However, the presented approach was computationally infeasible. Future research could focus on other methods and optimization techniques to solve this problem.

8.5 FAME-bank.net

A final perspective that needs special attention is data sharing. With FAME-bank, we tried to establish an environment for researchers to share their FAME data. This could not only lift the presented research in terms of larger data sets, but could also prove highly useful for the evaluation of the presented machine learning methods. With this database, an inter-laboratory comparison and analysis of FAME data will become possible. Actually, with this web portal, microbiologists can be made enthusiastic on the use of machine learning and computational analysis for identification purposes. Of course, besides a further development of the FAME-bank database, collaborations with other laboratories remain also a good option for extending FAME analysis research and these should also be further encouraged.

PART V

APPENDICES

APPENDIX A

Data Sets

A.1 Strain Tables

A.1.1 Data set 2006

Species name	Strains (Number of profiles)	EC
<i>B. alcalophilus</i>	LMG 7120 ^T (4)	
<i>B. amyloliquefaciens</i>	LMG 9814 ^T (7); LMG 12325 (1); LMG 12329 (1); LMG 12385 (1)	
<i>B. aquimaris</i>	LMG 23073 ^T (4)	
<i>B. arvi</i>	LMG 22165 ^T (6); R-16994 (1)	
<i>B. atrophaeus</i>	LMG 16797 ^T (1); LMG 8198 (9); LMG 8199 (5); LMG 17795 (4); LMG 17796 (2)	
<i>B. axarquiensis</i>	LMG 22476 ^T (5)	
<i>B. azotoformans</i>	LMG 9581 ^T (2); LMG 15443 (1)	
<i>B. badius</i>	LMG 7122 ^T (4); LMG 12332 (1); LMG 18004 (1); LMG 18005 (1); LMG 18006 (1)	
<i>B. barbaricus</i>	LMG 23067 (5)	
<i>B. bataviensis</i>	LMG 21833 ^T (7); LMG 21832 (1); R-15415 (1); R-15454 (2); R- 16296 (1); R-16308 (1); R-16321 (1); R-16324 (1); R-16325 (2); R-16336 (1); R-17019 (1)	
<i>B. bogoriensis</i>	LMG 22234 ^T (8)	
<i>B. carboniphilus</i>	LMG 18001 ^T (6)	
<i>B. cereus</i>	LMG 6923 ^T (7); LMG 2098 (1); LMG 6910 (1); LMG 6924 (1); LMG 8221 (9); LMG 8396 (2); LMG 9676 (1); LMG 12235 (1); LMG 12236 (1); LMG 12237 (1); LMG 12334 (1); LMG 12335 (3); LMG 12365 (9); LMG 13569 (4); LMG 14742 (4); LMG 17612 (1); LMG 18241 (2); LMG 18365 (2); LMG 18698 (1); LMG 22728 (1); LMG 22729 (1); LMG 22730 (1); LMG 22731 (1); LMG 22732 (1); LMG 22733 (1); R-20144 (1)	

Table A.1: Strain table corresponding to the 2006 *Bacillus* data set. Strain numbers and corresponding number of included FAME profiles are reported. Also, exceptional growth and culturing conditions are reported (column 'EC').

Table A.1 continued.

Species name	Strains (Number of profiles)	EC
<i>B. circulans</i> ¹	1) LMG 13261 ^T (5); LMG 13271 (1); LMG 16568 (1); LMG 16628 (1); LMG 16629 (2) 2) LMG 6927 (1); LMG 12238 (1); LMG 12343 (2) LMG 13265 (1); LMG 13266 (3); LMG 13267 (2); LMG 13268 (1); LMG 13270 (1); LMG 13272 (1); LMG 13273 (1); LMG 13274 (1); LMG 14421 (1); LMG 14422 (1); LMG 14423 (1); LMG 14424 (1); LMG 14636 (1); LMG 16560 (1); LMG 16561 (1); LMG 16564 (1); LMG 16565 (1); LMG 16567 (1); LMG 16570 (1); LMG 16583 (1); LMG 16585 (1); LMG 16593 (1); LMG 16594 (1); LMG 16595 (2); LMG 16624 (2); LMG 16626 (1); LMG 16627 (1); LMG 16630 (1); LMG 16634 (1); LMG 16635 (1); LMG 16637 (1); LMG 16638 (1); LMG 16640 (1); LMG 16641 (1); LMG 16645 (1); LMG 16646 (1); LMG 16647 (1); LMG 16649 (1); LMG 16650 (1); LMG 16693 (1); LMG 17441 (1); LMG 17442 (1); LMG 17444 (3); LMG 17447 (2); LMG 17449 (1); LMG 17464 (1); LMG 17465 (1); LMG 17469 (1); LMG 17470 (1); LMG 17472 (2); LMG 17473 (2); LMG 17486 (1); LMG 18014 (2); B0949 (1); R-36277 (1); R-36278 (1); R-36279 (1); R-36280 (1); R-36281 (1); R-36282 (1); R-36283 (1); R-36284 (1)	
<i>B. clausii</i>	LMG 17945 ^T (6); LMG 18518 (1)	
<i>B. coagulans</i>	LMG 6326 ^T (11); LMG 12345 (2); LMG 12346 (2); LMG 12398 (2); LMG 12399 (1); LMG 12400 (2); LMG 12401 (2); LMG 12402 (2); LMG 17451 (1); LMG 17452 (1); LMG 17453 (1); LMG 17456 (1); LMG 17457 (1); LMG 17474 (3); LMG 17475 (1); LMG 17476 (1); LMG 17477 (1); LMG 17478 (1); LMG 7376 (2)	
<i>B. cohnii</i>	LMG 16678 ^T (8)	
<i>B. decolorationis</i>	LMG 19507 ^T (4); R-5454 (1)	
<i>B. drentensis</i>	LMG 21831 ^T (26); LMG 21830 (1); R-15416 (1); R-15427 (1); R-15445 (1); R-16310 (3); R-16313 (1); R-16328 (2); R-16338 (1); R-16986 (1)	
<i>B. endophyticus</i>	LMG 21715 ^T (5)	
<i>B. farraginis</i>	LMG 22081 ^T (1); R-7343 (1); R-7148 (1)	
<i>B. firmus</i>	LMG 7125 ^T (5); LMG 12241 (1); LMG 12242 (1); LMG 12243 (1); LMG 12352 (1); R-15586 (1)	
<i>B. flexus</i>	LMG 11155 ^T (6)	
<i>B. fordii</i>	LMG 22142 (3)	
<i>B. fortis</i>	LMG 22079 ^T (1); LMG 22141 (2); R-7163 (1)	
<i>B. fumarioli</i>	R-10919 (1); R-13595 (1); R-13623 (1); R-13624 (1); R-13860 (1); R-13992 (1); R-14704 (1); R-14705 (1); R-14711 (1); R-16404 (1); R-19905 (1); R-19906 (1); R-19910 (1); R-20285 (1); R-20287 (1); R-20342 (1); R-20444 (1)	52°C
<i>B. funiculus</i>	LMG 22472 ^T (5)	
<i>B. fusiformis</i>	LMG 9816 ^T (6); LMG 17347 (2); B0661 (1)	
<i>B. galactosidilyticus</i>	LMG 17892 ^T (5); LMG 12353 (2); LMG 12396 (3); R-15577 (1); R-16004 (1)	
<i>B. gelatini</i>	LMG 21880 ^T (3); R-13476 (2); R-13565 (2); R-13588 (1); R-13635 (1); R-13810 (1); R-13864 (1); R-13975 (2)	52°C
<i>B. gibsonii</i>	LMG 17949 ^T (7)	
<i>B. halmapalus</i>	LMG 17950 ^T (6)	
<i>B. halophilus</i>	LMG 17942 ^T (3)	
<i>B. horikoshii</i>	LMG 17946 ^T (7)	
<i>B. horti</i>	LMG 18497 ^T (5)	
<i>B. humi</i>	LMG 22167 ^T (4); R-17036 (1)	
<i>B. indicus</i>	LMG 22858 ^T (6)	
<i>B. insolitus</i>	LMG 17757 ^T (3); LMG 17153 (3); B0433 (1)	
<i>B. jeotgali</i>	LMG 21653 ^T (6)	
<i>B. laevolacticus</i>	LMG 6923 ^T (9)	
<i>B. lentus</i> ²	LMG 16798 ^T (9); LMG 12354 (2); LMG 12359 (1); LMG 21649 (6); LMG 21758 (1); R-36285 (1)	

Table A.1 continued.

Species name	Strains (Number of profiles)	EC
<i>B. licheniformis</i>	LMG12363 ^T (8); LMG 6934 (1); LMG 7558 (5); LMG 7559 (1); LMG 7560 (4); LMG 7561 (2); LMG 7562 (3); LMG 7626 (1); LMG 7627 (3); LMG 7628 (1); LMG 7629 (1); LMG 7630 (1); LMG 7631 (3); LMG 7632 (2); LMG 7633 (1); LMG 7634 (1); LMG 7635 (1); LMG 7636 (1); LMG 7637 (2); LMG 12245 (1); LMG 12246 (1); LMG 12247 (1); LMG 12248 (1); LMG 12360 (1); LMG 12361 (1); LMG 12362 (1); LMG 17334 (1); LMG 17337 (1); LMG 17339 (1); LMG 17340 (1); LMG 17649 (1); LMG 17651 (1); LMG 17652 (1); LMG 17653 (1); LMG 17654 (1); LMG 17655 (1); LMG 17656 (1); LMG 17657 (1); LMG 17658 (1); LMG 17659 (1); LMG 17660 (1); LMG 17661 (1); LMG 17662 (1); LMG 17663 (1); LMG 18685 (1); R-1210 (1); R-6452 (1); R-6646 (1); R-6979 (1); R-7199 (1); R-7478 (1); R-21381 (2); R-21382 (2); R-21383 (2); R-21384 (2); R-21385 (2)	
<i>B. luciferensis</i>	LMG 18422 ^T (4); LMG 21400 (1); R-11670 (1); R-14109 (1); R-14110 (1); R-14111 (1)	
<i>B. malacitensis</i>	LMG 22477 ^T (5)	
<i>B. marisflavi</i>	LMG 23072 ^T (5)	
<i>B. megaterium</i>	LMG 7127 ^T (7); LMG 3585 (1); LMG 11162 (1); LMG 12249 (1); LMG 12250 (1); LMG 12252 (1); LMG 12253 (1); LMG 12254 (1); LMG 12255 (1); LMG 12408 (1); LMG 12409 (1); LMG 18670 (1); LMG 18686 (1); LMG 18687 (1); LMG 18688 (1); LMG 18705 (1); LMG 18710 (1); LMG 18714 (1); LMG 23147 (2)	
<i>B. mojavensis</i>	LMG 17797 ^T (4); LMG 17798 (4); R-28501 (1); R-28502 (1); R-28503 (1); R-28504 (1)	
<i>B. muralis</i>	LMG 20238 ^T (4); R-8204 (1); R-8210 (1); R-8251 (1)	
<i>B. mycooides</i>	LMG 7128 ^T (7); LMG 12256 (1)	
<i>B. neidei</i>	LMG 22737 (5)	
<i>B. niacini</i>	LMG 16677 ^T (5)	
<i>B. novalis</i>	LMG 21837 ^T (10); LMG 21836 (2); R-15418 (1); R-15446 (2); R-15450 (2); R-15453 (2); R-16295 (1); R-16297 (1); R-16309 (1); R-16340 (1); R-16342 (1); R-16345 (2); R-16347 (1)	
<i>B. okuhidensis</i>	LMG 22468 ^T (3)	
<i>B. oleronius</i>	LMG 17952 ^T (2); LMG 17882 (1); LMG 17884 (1); LMG 17887 (1)	
<i>B. patagoniensis</i>	LMG 23070 ^T (4)	
<i>B. pseudocaliphilus</i>	LMG 17951 ^T (6)	
<i>B. pseudofirmus</i>	LMG 17944 ^T (5)	
<i>B. pseudomycooides</i>	LMG 18993 ^T (6)	
<i>B. psychrodurans</i>	LMG 23063 ^T (5)	
<i>B. psychrosaccharolyticus</i>	LMG 9580 ^T (6)	
<i>B. psychrotolerans</i>	LMG 23062 ^T (6)	
<i>B. pumilus</i>	LMG 18928 ^T (17); LMG 3455 (3); LMG 8196 (3); LMG 8942 (2); LMG 10642 (3); LMG 12257 (1); LMG 12258 (2); LMG 12259 (3); LMG 12372 (1); LMG 12374 (1); LMG 12375 (1); LMG 12376 (4); LMG 12377 (1); LMG 18517 (9); LMG 18658 (1); LMG 18676 (1); B0296 (1); R-5334 (1); R-10579 (1); R-36286 (1); R-36308 (2); R-36309 (3); R-36310 (6); R-36311 (2)	
<i>B. pycnus</i>	LMG 21634 ^T (5)	
<i>B. ruris</i>	LMG 22866 ^T (3)	
<i>B. shackletonii</i>	LMG 18435 ^T (3); R-11667 (1); R-14112 (1); R-14113 (1); R-14114 (1)	
<i>B. silvestris</i>	LMG 18991 ^T (6)	
<i>B. simplex</i>	LMG 11160 ^T (8); LMG 12364 (2); LMG 17634 (1); LMG 17636 (1); LMG 17643 (1); LMG 18473 (2); LMG 18508 (1); LMG 19489 (1); B0405 (1); R-5269 (1); R-5275 (1); R-5282 (1); R-5307 (1); R-8191 (1); R-8193 (1); R-8202 (1); R-8207 (1); R-8208 (1); R-8214 (1); R-8215 (4); R-8216 (3); R-8218 (2); R-8220 (1); R-8225 (1); R-8231 (1); R-8234 (1); R-8253 (1); R-8254 (1); R-15936 (1); R-15943 (1)	

Table A.1 continued.

Species name	Strains (Number of profiles)	EC
<i>B. soralis</i>	LMG 22467 ^T (6)	
<i>B. smithii</i>	LMG 12526 ^T (3); LMG 6327 (2)	
<i>B. soli</i>	LMG 21838 ^T (5); LMG 21839 (1); R-16301 (1); R-16307 (1)	
<i>B. sonorensis</i>	LMG 21636 ^T (4)	
<i>B. sphaericus</i>	LMG 7134 ^T (13); LMG 18663 (1)	
<i>B. sporothermodurans</i>	LMG 17668 ^T (13); LMG 17895 (1); LMG 17896 (1); LMG 17897 (1); LMG 18460 (1); LMG 18461 (1); LMG 18462 (1); LMG 18463 (1); LMG 18464 (1); LMG 18465 (1); LMG 18466 (1); R-1952 (2)	
<i>B. subterraneus</i>	LMG 23065 (5)	
<i>B. subtilis</i>	LMG 12260 (1); LMG 12261 (2); LMG 12262 (1); LMG 12263 (4); LMG 12264 (1); LMG 12417 (1); LMG 13579 (1); LMG 17723 (1); LMG 17725 (6); LMG 18271 (4); LMG 2099 (5); LMG 3589 (1); LMG 3590 (1)	
<i>B. subsp. spizizenii</i>	LMG 19156 ^T (1); LMG 19155 (1); LMG 19545 (4); LMG 8197 (15)	
<i>B. subsp. subtilis</i>	LMG 7135 ^T (13); LMG 19154 (1)	
<i>B. thermoamylovorans</i>	LMG 18084 ^T (8)	52°C
<i>B. thermantarcticus</i>	LMG 23032 ^T (4)	52°C
<i>B. thuringiensis</i>	LMG 7138 ^T (6); LMG 12265 (1); LMG 12266 (1); LMG 12267 (2); LMG 12268 (1); LMG 12269 (1)	
<i>B. vallismortis</i>	LMG 18725 ^T (5); LMG 17799 (2); LMG 17800 (4); R-28507 (1); R-28553 (1)	
<i>B. velezensis</i>	LMG 22478 ^T (5)	
<i>B. vireti</i>	LMG 21834 ^T (5); R-15428 (1); R-15441 (1)	
<i>B. weihenstephanensis</i>	LMG 18989 ^T (6)	

¹ *Bacillus circulans* is a very heterogeneous group of strains. The strains reported in point 1 are considered to be *Bacillus circulans*. The strains reported in point 2 are assigned to *Bacillus circulans* but further revision is needed.

² The strain LMG 21649 is annotated as *Bacillus lentus* and as *Bacillus halodurans* by different culture collections (Dawyndt, 2009).

A.1.2 Data sets 2008

Species name	Strains (Number of profiles)	EC
<i>B. alcalophilus</i>	LMG 7120 ^T (4)	
<i>B. amyloliquefaciens</i>	LMG 9814 ^T (7); LMG 12325 (1); LMG 12329 (1); LMG 12385 (1); LMG 22478 (5)	
<i>B. aquimaris</i>	R-38158 (4); R-38159 (8)	
<i>B. atrophaeus</i>	LMG 16797 ^T (1); LMG 8198 (5); LMG 8199 (2); LMG 17795 (4); LMG 17796 (2); R 38160 (2); R-38161 (1); R-38846 (2); R-38847 (2)	
<i>B. azotoformans</i>	LMG 9581 ^T (2); LMG 15443 (1)	
<i>B. badius</i>	LMG 7122 ^T (4); LMG 12332 (1); R-1167 (1); R-1168 (1); R-1202 (1)	
<i>B. barbaricus</i>	LMG 23067 ^T (5)	
<i>B. bataviensis</i>	LMG 21833 ^T (7); LMG 21832 (1); R-15415 (1); R-15454 (2); R-16296 (1); R-16308 (1); R-16321 (1); R-16324 (1); R-16325 (2); R-16336 (1); R-17019 (1)	
<i>B. bogoriensis</i>	LMG 22234 ^T (8)	
<i>B. carboniphilus</i>	LMG 18001 ^T (6)	
<i>B. cereus</i>	LMG 6923 ^T (7); LMG 6910 (1); LMG 6924 (1); LMG 8221 (11); LMG 8396 (2); LMG 9005 (4); LMG 9676 (1); LMG 12235 (1); LMG 12236 (1); LMG 12237 (1); LMG 12334 (1); LMG 12335 (3); LMG 12365 (9); LMG 14742 (4); LMG 17612 (1); LMG 18241 (2); LMG 18365 (2); LMG 18698 (1); LMG 22728 (1); LMG 22729 (1); LMG 22730 (1); LMG 22731 (1); LMG 22732 (1); LMG 22733 (1); R-2896 (1); R-20144 (1); R-38162 (1)	
<i>B. circulans</i>	LMG 13261 ^T (7); LMG 16568 (1); LMG 16628 (1); LMG 16629 (2)	
<i>B. clausii</i>	LMG 17945 ^T (6); LMG 18518 (1)	
<i>B. coagulans</i>	LMG 6326 ^T (5); LMG 7376 (2); LMG 12345 (2); LMG 12346 (2); LMG 12398 (2); LMG 12399 (1); LMG 12400 (2); LMG 12401 (2); LMG 12402 (1); LMG 17451 (1); LMG 17452 (1); LMG 17453 (1); LMG 17456 (1); LMG 17457 (1); LMG 17474 (3); LMG 17475 (1); LMG 17476 (1); LMG 17477 (1); LMG 17478 (1); R-38163 (1)	
<i>B. cohnii</i>	LMG 16678 ^T (8)	
<i>B. decolorationis</i>	LMG 19507 ^T (4); R-5454 (1)	
<i>B. drementensis</i>	LMG 21831 ^T (26); LMG 21830 (1); R-15416 (1); R-15427 (1); R-15445 (1); R-16310 (3); R-16313 (1); R-16328 (2); R-16338 (1); R-16986 (1)	
<i>B. endophyticus</i>	LMG 21715 ^T (5)	
<i>B. firmus</i>	LMG 7125 ^T (6); LMG 12241 (1); LMG 12242 (1); LMG 12243 (1); LMG 12352 (1); R 15586 (1)	
<i>B. flexus</i>	LMG 11155 ^T (7)	
<i>B. foraminis</i>	LMG 23174 ^T (4)	
<i>B. fortis</i>	LMG 22079 ^T (1)	
<i>B. fordii</i>	LMG 22142 (3)	
<i>B. fortis</i>	LMG 22141 (2); R-7163 (1)	
<i>B. fumarioli</i>	LMG 19448 ^T (2); LMG 17492 (1); LMG 18409 (1); LMG 18418 (1); R-38164 (1); R 38165 (1); R-38166 (1); R-38167 (1); R-38168 (1); R-38169 (1); R-38170 (1); R 38171 (1); R-38172 (1); R-38173 (1); R-38174 (1); R-38175 (1); R-38176 (1); R 38177 (2); R-38178 (1); R-38179 (1); R-38180 (1); R-38181 (1); R-38182 (1); R 38183 (2); R-38184 (1)	52°C

Table A.2: Strain table corresponding to the 2008 *Bacillus*, *Paenibacillus* and *Pseudomonas* data sets. Strain numbers and corresponding number of included FAME profiles are reported. Also, exceptional growth and culturing conditions are reported (column ‘EC’). Regarding the *Pseudomonas* strains, the plant-pathogenic strain reallocation by Gardan et al. (1999) is followed and integrated in the table. For this genus, strains with one or more plant-pathogenic strains are denoted by superscript ‘p’.

Table A.2 continued.

Species name	Strains (Number of profiles)	EC
<i>B. funiculus</i>	R-38185 (5)	
<i>B. galactosidilyticus</i>	LMG 17892 ^T (5); LMG 12353 (2); LMG 12396 (3); R-15577 (1); R-16004 (1)	
<i>B. gelatini</i>	LMG 21880 ^T (3); R-13476 (2); R-13565 (2); R-13588 (1); R-13635 (1); R-13810 (1); R 13864 (1); R-13975 (2)	52°C
<i>B. gibsonii</i>	LMG 17949 ^T (7)	
<i>B. halmapalus</i>	LMG 17950 ^T (6)	
<i>B. halodurans</i>	LMG 7121 ^T (2); LMG 21649 (6)	
<i>B. horikoshii</i>	LMG 17946 ^T (7)	
<i>B. horti</i>	LMG 18497 ^T (5)	
<i>B. humi</i>	LMG 22167 ^T (4); R-17036 (1)	
<i>B. indicus</i>	LMG 22858 ^T (6)	
<i>B. insolitus</i>	LMG 17757 ^T (3); LMG 17153 (3); R-38186 (1)	
<i>B. jeotgali</i>	LMG 21653 ^T (2); R-38187 (2); R-38188 (2)	
<i>B. lentus</i>	LMG 16798 ^T (9); LMG 12354 (2); LMG 12359 (1); LMG 21758 (1); R-36285 (1)	
<i>B. licheniformis</i>	LMG 12363 ^T (6); LMG 6934 (1); LMG 7558 (4); LMG 7559 (1); LMG 7560 (3); LMG 7561 (2); LMG 7626 (1); LMG 7628 (1); LMG 7629 (1); LMG 7630 (1); LMG 7632 (2); LMG 7633 (1); LMG 7634 (1); LMG 7636 (1); LMG 7637 (2); LMG 12245 (1); LMG 12246 (1); LMG 12247 (1); LMG 12248 (1); LMG 12360 (1); LMG 12361 (1); LMG 12362 (1); LMG 17334 (1); LMG 17337 (1); LMG 17339 (1); LMG 17340 (1); LMG 17649 (1); LMG 17651 (1); LMG 17652 (1); LMG 17653 (1); LMG 17654 (1); LMG 17655 (1); LMG 17656 (1); LMG 17657 (1); LMG 17658 (1); LMG 17659 (1); LMG 17661 (1); LMG 17662 (1); LMG 17663 (1); LMG 18685 (1); R 1210 (1); R-6452 (1); R-6646 (1); R-6979 (1); R-7199 (1); R-7478 (1); R-15573 (1); R 38189 (1); R-38190 (1); R-38191 (2); R-38192 (1); R-38193 (2); R-38194 (1); R 38195 (2); R-38196 (1); R 38197 (1); R-38848 (1); R-38849 (1)	
<i>B. luciferensis</i>	LMG 18422 ^T (4); LMG 21400 (1); R-11670 (1); R-14109 (1); R-14110 (1); R-14111 (1)	
<i>B. marisflavi</i>	LMG 23072 ^T (5)	
<i>B. megaterium</i>	LMG 7127 ^T (7); LMG 11162 (1); LMG 12249 (1); LMG 12250 (1); LMG 12252 (1); LMG 12253 (1); LMG 12254 (1); LMG 12255 (1); LMG 12408 (1); LMG 12409 (1); LMG 18670 (1); LMG 18686 (1); LMG 18687 (1); LMG 18688 (1); LMG 18705 (1); LMG 18710 (1); LMG 18714 (1); LMG 23147 (2); R-1092 (2); R-38198 (1)	
<i>B. mojavensis</i>	LMG 17797 ^T (4); LMG 22476 (5); LMG 22477 (5); R-28501 (1); R-28502 (1); R 28503 (1); R-28504 (1); R-38850 (2); R-38851 (2)	
<i>B. muralis</i>	LMG 20238 ^T (4); R-8204 (1); R-8210 (1); R-8251 (1)	
<i>B. mycooides</i>	LMG 7128 ^T (7); LMG 12256 (1); R-2892 (1); R-2893 (1); R-2895 (1)	
<i>B. niacini</i>	LMG 16677 ^T (5)	
<i>B. novalis</i>	LMG 21837 ^T (10); LMG 21836 (2); R-15418 (1); R-15446 (2); R-15450 (2); R-15453 (2); R-16295 (1); R-16297 (1); R-16309 (1); R-16340 (1); R-16342 (1); R-16345 (2); R 16347 (1)	
<i>B. okuhidensis</i>	LMG 22468 ^T (6)	
<i>B. oleronius</i>	LMG 17952 ^T (2); LMG 17882 (1); LMG 17884 (1); LMG 17887 (1)	
<i>B. patagoniensis</i>	LMG 23070 ^T (8); R-38852 (2); R-38864 (2)	
<i>B. pseudalcaliphilus</i>	LMG 17951 ^T (6)	
<i>B. pseudofirmus</i>	LMG 17944 ^T (5)	
<i>B. pseudomycooides</i>	LMG 18993 ^T (6)	
<i>B. psychrodurans</i>	LMG 23063 ^T (7)	
<i>B. psychrosaccharolyticus</i>	LMG 9580 ^T (6)	
<i>B. psychrotolerans</i>	LMG 23062 ^T (7)	
<i>B. pumilus</i>	LMG 18928 ^T (17); LMG 3455 (3); LMG 8196 (3); LMG 10642 (3); LMG 12257 (1); LMG 12259 (3); LMG 12372 (1); LMG 12374 (1); LMG 12375 (1); LMG 12376 (4); LMG 12377 (1); LMG 18517 (9); LMG 18658 (1); LMG 18676 (1); LMG 21165 (1); R 5334 (1); R-33429 (1); R-36286 (1); R-38199 (1); R-38200 (1); R-38853 (1); R 38854 (1)	

Table A.2 continued.

Species name	Strains (Number of profiles)	EC
<i>B. pycnus</i>	LMG 21634 ^T (5)	
<i>B. ruris</i>	LMG 22866 ^T (3); LMG 22867 (1)	
<i>B. shackletonii</i>	LMG 18435 ^T (3); R-11667 (1); R-14112 (1); R-14113 (1); R-14114 (1)	
<i>B. silvestris</i>	LMG 18991 ^T (6)	
<i>B. simplex</i>	LMG 11160 ^T (8); LMG 12364 (2); LMG 17634 (1); LMG 17636 (1); LMG 17643 (1); LMG 18473 (3); LMG 18508 (1); LMG 19489 (1); LMG 21002 (3); LMG 22045 (1); LMG 22046 (1); R-5275 (1); R-5307 (1); R-8191 (1); R-8193 (1); R-8202 (1); R-8207 (1); R-8208 (1); R-8214 (1); R-8215 (4); R-8218 (2); R-8220 (1); R-8225 (1); R-8231 (1); R 8234 (1); R-8253 (1); R-8254 (1); R-15936 (1); R-15943 (1); R-38201 (1)	
<i>B. siralis</i>	LMG 22467 ^T (4)	
<i>B. smithii</i>	LMG 12526 ^T (1); LMG 6327 (2)	
<i>B. soli</i>	LMG 21838 ^T (5); LMG 21839 (1); R-16301 (1); R-16307 (1)	
<i>B. sonorensis</i>	LMG 21636 ^T (4); R-28505 (1); R-28506 (1); R-28548 (1); R-28552 (1)	
<i>B. sporothermodurans</i>	LMG 17668 ^T (6); LMG 17895 (1); LMG 17896 (1); LMG 17897 (1); LMG 18460 (1); LMG 18461 (1); LMG 18462 (1); LMG 18463 (1); LMG 18464 (1); LMG 18465 (1); LMG 18466 (1); R-1952 (1)	
<i>B. subterraneus</i>	R-38855 ^T (5)	
<i>B. subtilis</i>	LMG 2099 (5); LMG 3589 (1); LMG 3590 (1); LMG 12260 (1); LMG 12262 (1); LMG 12263 (2); LMG 12264 (1); LMG 12417 (1); LMG 13579 (1); LMG 17723 (1); R 38203 (2); R-38204 (2); R-38205 (2); R-38856 (1); R-38857 (1); R-38858 (1); R 38859 (1); R-38860 (2); R-38861 (2)	
<i>B. subtilis</i> subsp. <i>spizizenii</i>	LMG 19156 ^T (1); LMG 8197 (15); LMG 19155 (1); LMG 19545 (4)	
<i>B. subtilis</i> subsp. <i>subtilis</i>	LMG 7135 ^T (13); LMG 19154 (1)	
<i>B. thermantarcticus</i>	LMG 23032 ^T (5)	52°C
<i>B. thermoamylovorans</i>	LMG 18084 ^T (7)	52°C
<i>B. thuringiensis</i>	LMG 7138 ^T (6); LMG 12265 (1); LMG 12266 (1); LMG 12267 (2); LMG 12268 (1); LMG 12269 (1)	
<i>B. vallismortis</i>	LMG 18725 ^T (7); LMG 17799 (2); R-28507 (1); R-28553 (1); R-38862 (2); R-38863 (2)	
<i>B. vireti</i>	LMG 21834 ^T (5); R-15428 (1); R-15441 (1)	
<i>B. weihenstephanensis</i>	LMG 18989 ^T (6)	
<i>Pa. alvei</i>	LMG 13253 ^T (3); LMG 13254 (1); LMG 13255 (1); LMG 13256 (1); LMG 13258 (1); LMG 13260 (1); LMG 16907 (1); LMG 16912 (1); LMG 16913 (1); LMG 16914 (1); LMG 16915 (1); LMG 17051 (1); LMG 17052 (1); LMG 17053 (1)	
<i>Pa. amylolyticus</i>	LMG 21767 ^T (4)	
<i>Pa. anaericanus</i>	LMG 23658 ^T (2); LMG 23878 (4)	48h
<i>Pa. antarcticus</i>	LMG 22078 ^T (3)	48h
<i>Pa. apiarius</i>	LMG 17433 ^T (3); LMG 17434 (2)	
<i>Pa. azoreducens</i>	LMG 21668 ^T (4)	
<i>Pa. borealis</i>	LMG 21603 ^T (4)	
<i>Pa. cellulolyticus</i>	LMG 22232 ^T (9); R-38865 (2); R-38866 (2)	
<i>Pa. chibensis</i>	LMG 14457 ^T (8); R-38867 (1)	
<i>Pa. chitinolyticus</i>	LMG 18047 ^T (7)	
<i>Pa. cineris</i>	LMG 18439 ^T (3); LMG 21976 (3)	
<i>Pa. cookii</i>	LMG 18419 ^T (4); LMG 18437 (2)	
<i>Pa. curdolanolyticus</i>	LMG 23061 ^T (5)	
<i>Pa. dendritiformis</i>	LMG 21716 ^T (6)	
<i>Pa. durus</i>	LMG 18446 ^T (2); LMG 14658 (5); LMG 14659 (1); LMG 14661 (1)	
<i>Pa. elgii</i>	LMG 24465 ^T (5)	48h
<i>Pa. favisporus</i>	LMG 20987 ^T (3); LMG 20989 (1); R-38868 (2); R-38869 (2)	
<i>Pa. fonticola</i>	LMG 23577 ^T (5)	
<i>Pa. ginsengarvi</i>	LMG 23815 ^T (4)	
<i>Pa. glucanolyticus</i>	LMG 12239 ^T (4); LMG 12240 (1); LMG 12395 (2)	

Table A.2 continued.

Species name	Strains (Number of profiles)	EC
<i>Pa. humicus</i>	LMG 23886 ^T (3)	
<i>Pa. illinoisensis</i>	LMG 18051 ^T (6)	
<i>Pa. jamilae</i>	LMG 21667 ^T (3)	
<i>Pa. kobensis</i>	LMG 18049 ^T (3)	
<i>Pa. lactis</i>	LMG 21940 ^T (3)	
<i>Pa. larvae</i>	LMG 9820 ^T (5); LMG 14425 (1); LMG 14426 (1); LMG 14427 (3); LMG 14428 (4); LMG 15974 (7); LMG 16214 (1); LMG 16215 (1); LMG 16241 (1); LMG 16242 (1); LMG 16243 (1); LMG 16244 (1); LMG 16245 (1); LMG 16246 (1); LMG 16247 (2); LMG 16249 (2); LMG 16250 (2); LMG 16251 (2); LMG 16252 (2)	
<i>Pa. lautus</i>	LMG 11157 ^T (2); LMG 14015 (2); LMG 14669 (1); R-38870 (1)	
<i>Pa. macerans</i>	LMG 13281 ^T (4); LMG 6325 (1); LMG 13282 (1); LMG 13283 (1); LMG 13284 (1); LMG 13285 (1); LMG 13286 (1); LMG 13288 (1); LMG 18690 (1); LMG 21891 (3)	
<i>Pa. macquariensis</i>	LMG 6935 ^T (2); LMG 13290 (2); LMG 13291 (1)	
<i>Pa. mendelii</i>	LMG 23002 ^T (6)	48h
<i>Pa. odorifer</i>	LMG 19079 ^T (4)	
<i>Pa. pabuli</i>	LMG 15970 ^T (3); LMG 12394 (1); LMG 13292 (1); LMG 14016 (1); LMG 14017 (1); LMG 14671 (1); R-38871 (1); R-38872 (1)	
<i>Pa. panacisoli</i>	LMG 23405 ^T (6)	
<i>Pa. peoriae</i>	LMG 14832 ^T (2); LMG 16104 (1); LMG 16108 (1); LMG 16109 (1)	
<i>Pa. phyllosphaerae</i>	LMG 22192 ^T (6)	
<i>Pa. polymyxa</i>	LMG 13294 ^T (6); LMG 6320 (1); LMG 6321 (1); LMG 11619 (1); LMG 11623 (1); LMG 11647 (1); LMG 11649 (2); LMG 11724 (1); LMG 13295 (1); LMG 13297 (1); LMG 13298 (1); LMG 13301 (1); LMG 21892 (6); R-2386 (1); R-2472 (1); R-2507 (1); R 38873 (1); R-38874 (1); R-38875 (1); R-38876 (1); R-38877 (1); R-38878 (1); R 38879 (1); R-38880 (1); R-38881 (1); R-38882 (1); R-38883 (1); R-38884 (1); R 38885 (1); R-38886 (1); R-38887 (1); R-38888 (1); R-38889 (1); R-38890 (1); R 38891 (1); R-38892 (1); R-38893 (1); R-38894 (1); R-38895 (1); R-38896 (1); R 38897 (1); R-38898 (2); R-38899 (1)	
<i>Pa. rhizosphaerae</i>	LMG 21955 ^T (5)	
<i>Pa. soli</i>	LMG 23604 ^T (3)	
<i>Pa. stellifer</i>	LMG 22679 ^T (4)	
<i>Pa. taiwanensis</i>	LMG 23799 ^T (6)	
<i>Pa. thiaminolyticus</i>	LMG 17412 ^T (7); LMG 16908 (2); LMG 16916 (1); LMG 16917 (1); LMG 16918 (1); LMG 16919 (1); LMG 16920 (1); LMG 16921 (1); LMG 16922 (1); LMG 16923 (1); LMG 16924 (1); LMG 16925 (1); LMG 16926 (1); LMG 17406 (1); LMG 17407 (1); LMG 17409 (1); LMG 17410 (1)	
<i>Pa. validus</i>	LMG 11161 ^T (2); LMG 9817 (1); LMG 14018 (1); LMG 14019 (1); LMG 14020 (1); LMG 14468 (2); LMG 14469 (2); LMG 14470 (1); LMG 14663 (1); LMG 14664 (1); LMG 14665 (1); LMG 14666 (1); LMG 14668 (1); LMG 14717 (1)	
<i>Pa. wynnii</i>	LMG 22176 ^T (2); R-16774 (1); R-16780 (1); R-16781 (1); R-22540 (1)	
<i>Pa. xylanilyticus</i>	LMG 21957 ^T (6)	
<i>P. abietaniphila</i>	LMG 20220 ^T (10)	

Table A.2 continued.

Species name	Strains (Number of profiles)	EC
<i>P. aeruginosa</i>	LMG 1242 ^T (18); LMG 1272 (1); LMG 1274 (1); LMG 5031 (1); LMG 5032 (2); LMG 6395 (11); LMG 8029 (13); LMG 9009 (6); LMG 10268 (2); LMG 10269 (2); LMG 10270 (1); LMG 10639 (4); LMG 10643 (1); LMG 12121 (1); LMG 12228 (1); LMG 13757 (1); LMG 13771 (1); LMG 13802 (1); LMG 13836 (1); LMG 13842 (1); LMG 13883 (1); LMG 13909 (2); LMG 14071 (1); LMG 14072 (1); LMG 14073 (1); LMG 14076 (1); LMG 14077 (1); LMG 14078 (1); LMG 14079 (1); LMG 14080 (1); LMG 14081 (1); LMG 14082 (1); LMG 14083 (1); LMG 14084 (1); LMG 14085 (1); LMG 14741 (1); LMG 15153 (4); LMG 18574 (1); LMG 18585 (2); LMG 18591 (1); LMG 18600 (2); LMG 18616 (1); LMG 18619 (1); LMG 18629 (1); LMG 21144 (1); LMG 21145 (1); LMG 23160 (2); R-11747 (1); R-11748 (1); R-14056 (1); R-14057 (1); R-16141 (3); R-16146 (1); R-16150 (3); R-16159 (2); R-16165 (2); R-16944 (1); R 16960 (1); R-17312 (1); R-17322 (1); R-17395 (1); R-17420 (1); R-17437 (1); R 17440 (1); R-17769 (1); R-17773 (1); R-17930 (1); R-17935 (1); R-17946 (1); R 17951 (1); R-17955 (1); R-38900 (3); R-38901 (2); R-38902 (1); R-8844 (1)	
<i>P. agarici</i> ^P	LMG 2112 ^T (5); LMG 2110 (1); LMG 2113 (1); LMG 2115 (1)	
<i>P. alcaligenes</i>	LMG 1224 ^T (6); LMG 6353 (1); LMG 6355 (1); R-17306 (1); R-17774 (1); R-17929 (1); R-38903 (1)	
<i>P. alcaliphila</i>	LMG 23134 ^T (10)	
<i>P. amygdali</i> ^P	LMG 2220 (2); LMG 5694 (2); LMG 5695 (1); LMG 5696 (1)	
	<i>P. savastanoi</i> pv. <i>glycinae</i> : LMG 5066 (5); LMG 5171 (1)	
	<i>P. savastanoi</i> pv. <i>phaseolica</i> : LMG 2245 (6)	
	<i>P. savastanoi</i> pv. <i>savastanoi</i> : LMG 2209 (3); LMG 5011 (1); LMG 5154 (2); LMG 5187 (1); LMG 5385 (1); LMG 5387 (1); LMG 5389 (1); LMG 5487 (1); LMG 6766 (1); LMG 6767 (2); LMG 17565 (1); LMG 17570 (1); LMG 17571 (1); LMG 17572 (1); LMG 17573 (1); LMG 17574 (1); LMG 17575 (1); LMG 17576 (1); LMG 17577 (1); LMG 17578 (1); LMG 17579 (1); LMG 17580 (1); LMG 17581 (1); LMG 17582 (1); LMG 17583 (1); LMG 17584 (1); LMG 17585 (1); LMG 17586 (1); LMG 17587 (1); LMG 21151 (1); LMG 21152 (1); LMG 21153 (1); LMG 21154 (1); LMG 21155 (1); R-4611 (1); R-4612 (1); R-4614 (1); R-4615 (1); R-4617 (1); R-4618 (1); R-4619 (1); R-4620 (1); R-4621 (1); R-4622 (1); R-4623 (1); R-4624 (1); R-38904 (1); R 38905 (1); R-38906 (1)	
	<i>P. syringae</i> pv. <i>ciccaronei</i> : LMG 5541 (3)	
	<i>P. syringae</i> pv. <i>eriobotryae</i> : LMG 2184 (3); LMG 5654 (1)	
	<i>P. syringae</i> pv. <i>lachrymans</i> : LMG 5070 (2); LMG 5172 (1); R-38907 (1)	
	<i>P. syringae</i> pv. <i>mellae</i> : LMG 5072 (2); LMG 5073 (1)	
	<i>P. syringae</i> pv. <i>mori</i> : LMG 5074 (2); LMG 5562 (1)	
	<i>P. syringae</i> pv. <i>morsprunorum</i> : LMG 2222 (1); LMG 5075 (5); R-38908 (1); R 38909 (1)	
	<i>P. syringae</i> pv. <i>myricae</i> : LMG 5668 (2); LMG 5669 (1)	
	<i>P. syringae</i> pv. <i>sesami</i> : LMG 2289 (3); LMG 5489 (1)	
	<i>P. syringae</i> pv. <i>tabaci</i> : LMG 5192 (1); LMG 5393 (4)	
	<i>P. syringae</i> pv. <i>ulmi</i> : LMG 5094 (3)	
<i>P. anguilliseptica</i>	LMG 21629 ^T (6); R-38910 (1); R-38911 (1); R-38912 (1)	
<i>P. antarctica</i>	LMG 22709 ^T (11); LMG 23832 (3)	
<i>P. argentinensis</i>	LMG 22563 ^T (5); LMG 22564 (2)	
<i>P. asplenii</i> ^P	LMG 2137 ^T (10); R-38913 (1); R-38914 (2)	
<i>P. avellanae</i> ^P	<i>P. syringae</i> pv. <i>theae</i> : LMG 5092 (3); LMG 5687 (1)	
<i>P. azotoformans</i>	LMG 21611 ^T (7)	
<i>P. balearica</i>	LMG 18376 ^T (6)	
<i>P. beteli</i> ^P	LMG 978 ^T (6)	
<i>P. borbori</i>	LMG 23199 ^T (6)	
<i>P. boreopolis</i>	LMG 979 ^T (7)	

Table A.2 continued.

Species name	Strains (Number of profiles)	EC
<i>P. brassicacearum</i>	LMG 21623 ^T (10)	
<i>P. brenneri</i>	LMG 23068 ^T (10)	
<i>P. cannabina</i> ^P	LMG 5096 ^T (4); LMG 2150 (1); LMG 5650 (2)	
<i>P. caricapapayae</i> ^P	LMG 2152 ^T (2); LMG 2153 (1); LMG 5051 (1); LMG 5375 (1)	
<i>P. chloritidismutans</i>	LMG 23064 ^T (10) (Should be enclosed in <i>P. stutzeri</i>)	
<i>P. chlororaphis</i> subsp. <i>aurantiaca</i>	LMG 21630 ^T (6)	
<i>P. chlororaphis</i> subsp. <i>aureofaciens</i>	LMG 1245 ^T (10); LMG 16909 (1); LMG 5832 (10); R-16169 (2); R-38915 (2)	
<i>P. chlororaphis</i> subsp. <i>chlororaphis</i>	LMG 5004 ^T (6); R-16943 (1); R-18031 (1)	
<i>P. cichorii</i> ^P	LMG 2162 ^T (3); LMG 1248 (1); LMG 2163 (1); LMG 2164 (1); LMG 2165 (1); LMG 5052 (1); LMG 5055 (1); LMG 24427 (1); LMG 24428 (2); LMG 24429 (1); LMG 24440 (1); R-25254 (1); R-25295 (1); R-25315 (1); R-26431 (1); R-26451 (1); R 27702 (1); R-28087 (1); R-29002 (1); R-29006 (1); R-29260 (1); R-31789 (1); R 36300 (1); R-36301 (1); R-36302 (1); R-36303 (1); R-36801 (1); R-36802 (1); R 36804 (1)	
<i>P. cissicola</i> ^P	LMG 21719 ^T (6); LMG 2168 (2)	
<i>P. citronellolis</i>	LMG 18378 ^T (7)	
<i>P. congelans</i>	LMG 21466 ^T (6)	
“ <i>P. coronafaciens</i> ” ^P	LMG 13190 ^T (3); LMG 2330 (1); LMG 5030 (3); LMG 5060 (2); LMG 5081 (5)	
	<i>P. syringae</i> pv. <i>garcae</i> : LMG 5064 (1); LMG 5065 (1)	
	<i>P. syringae</i> pv. <i>oryzae</i> : LMG 10912 (5); LMG 10913 (2); LMG 10914 (2); LMG 10915 (4); LMG 10916 (3); LMG 10917 (4); LMG 10918 (2); LMG 10919 (2); LMG 10920 (4)	
<i>P. corrugata</i> ^P	LMG 2172 ^T (9); LMG 1276 (3); LMG 2173 (4); LMG 5036 (2); LMG 5037 (1); LMG 5038 (2); R-17963 (1)	
<i>P. costantinii</i> ^P	LMG 22119 ^T (6)	
<i>P. cremoricolorata</i>	R-38951 ^T (4)	
<i>P. extremorientalis</i>	LMG 19695 ^T (6)	
<i>P. flavescens</i> ^P	LMG 18387 ^T (7)	
<i>P. flectens</i> ^P	LMG 2187 ^T (2); LMG 2186 (4)	
<i>P. fluorescens</i> ^P	LMG 1794 ^T (17); LMG 1244 (5); LMG 1799 (2); LMG 2189 (1); LMG 5167 (3); LMG 5168 (5); LMG 5822 (3); LMG 5825 (2); LMG 5830 (12); LMG 5831 (1); LMG 5833 (1); LMG 5849 (1); LMG 5916 (2); LMG 5938 (4); LMG 5939 (20); LMG 5940 (2); LMG 6812 (2); LMG 7207 (2); LMG 7216 (1); LMG 7220 (1); LMG 14561 (1); LMG 14562 (1); LMG 14563 (1); LMG 14564 (1); LMG 14565 (1); LMG 14566 (1); LMG 14567 (1); LMG 14568 (1); LMG 14569 (1); LMG 14570 (1); LMG 14571 (1); LMG 14573 (1); LMG 14574 (1); LMG 14575 (1); LMG 14576 (1); LMG 14577 (1); LMG 14673 (4); LMG 14674 (6); LMG 14675 (5); R-16143 (1); R 16177 (1); R-16178 (2); R-16185 (3); R-16930 (1); R-16955 (1); R-16956 (1); R 17303 (1); R-17333 (1); R-17397 (1); R-17400 (1); R-17414 (1); R-17441 (1); R 17797 (1); R-17927 (1); R-17932 (1); R-17956 (1); R-38916 (1); R-38917 (1); R 38918 (1); R-38919 (1); R-38920 (1); R-38921 (1); R-38922 (1); R-38923 (1); R 38924 (1); R-38925 (2)	
<i>P. fragi</i>	LMG 2191 ^T (12); LMG 5919 (4); LMG 5920 (1); R-35697 (2); R-35701 (2); R-35703 (2); R-35705 (2); R-35706 (1); R-35706 t2 (1); R-35709 (2); R-35710 (2); R-35717 (2); R 35719 (2)	
<i>P. frederiksbergensis</i>	LMG 19851 ^T (6)	
<i>P. fulva</i>	LMG 11722 ^T (3)	
<i>P. fuscovaginae</i> ^P	LMG 2158 ^T (8); LMG 2192 (12); LMG 5097 (7); LMG 5742 (6); LMG 12424 (1); LMG 12426 (1); LMG 12428 (1); R-1256 (1); R-1302 (1); R-1341 (1); R-1774 (1); R 1775 (1); R-1776 (1); R-1778 (1); R-1779 (1); R-1789 (1)	
<i>P. geniculata</i>	LMG 2195 ^T (6)	

Table A.2 continued.

Species name	Strains (Number of profiles)	EC
<i>P. genomospecies3</i> ^P	<i>P. syringae</i> pv. <i>antirrhini</i> : LMG 5057 (4); LMG 5377 (1) <i>P. syringae</i> pv. <i>apii</i> : LMG 2132 (5); LMG 5058 (1) <i>P. syringae</i> pv. <i>berberidis</i> : LMG 2146 (1); LMG 2147 (2) <i>P. syringae</i> pv. <i>delphinii</i> : LMG 5003 (1); LMG 5381 (3) <i>P. syringae</i> pv. <i>maculicola</i> : LMG 5071 (1); LMG 5559 (1) <i>P. syringae</i> pv. <i>passiflorae</i> : LMG 5185 (2); LMG 5671 (1) <i>P. syringae</i> pv. <i>persicae</i> : LMG 5184 (2); LMG 5566 (1) <i>P. syringae</i> pv. <i>primulae</i> : LMG 2252 (2); LMG 5680 (1) <i>P. syringae</i> pv. <i>ribicola</i> : LMG 2276 (3) <i>P. syringae</i> pv. <i>tomato</i> : LMG 5093 (3); LMG 5155 (1) <i>P. syringae</i> pv. <i>viburni</i> : LMG 2351 (2)	
<i>P. genomospecies7</i> ^P	<i>P. syringae</i> pv. <i>helianthi</i> : LMG 5067 (3); LMG 5558 (1) <i>P. syringae</i> pv. <i>tagetis</i> : LMG 5090 (3); LMG 5684 (1)	
<i>P. gessardii</i>	LMG 21604 ^T (10)	
<i>P. graminis</i>	LMG 21661 ^T (11)	
<i>P. hibiscicola</i> ^P	LMG 980 ^T (6)	
<i>P. indica</i>	LMG 23066 ^T (10)	
<i>P. jessenii</i>	LMG 21605 ^T (6)	
<i>P. jinjuensis</i>	LMG 21316 ^T (6)	
<i>P. kilonensis</i>	LMG 21624 ^T (10)	
<i>P. knackmussii</i>	LMG 23759 ^T (3)	
<i>P. koreensis</i>	LMG 21318 ^T (6)	
<i>P. libanensis</i>	LMG 21606 ^T (6)	
<i>P. lini</i>	LMG 21625 ^T (9); R-16937 (1)	
<i>P. lundensis</i>	LMG 13517 ^T (6); R-35702 (2); R-35711 (2); R-35721 (2); R-35723 (2); R-35724 (2)	
<i>P. lurida</i>	LMG 21995 ^T (4); R-38926 (1); R-38927 (1)	
<i>P. lutea</i>	LMG 21974 ^T (5)	
<i>P. luteola</i>	LMG 7041 ^T (6); LMG 5946 (2); R-16151 (1); R-16174 (2); R-17793 (1)	
<i>P. mandelii</i>	LMG 21607 ^T (6)	
<i>P. marginalis</i> ^P	LMG 5175 (2); LMG 5850 (1); LMG 6466 (2); LMG 6469 (1); LMG 6481 (1); LMG 6482 (1); LMG 6802 (1); LMG 6804 (5); LMG 6815 (1); LMG 14572 (1); R-16967 (1); R 17331 (1); R-17403 (1); R-17958 (1) <i>P. marginalis</i> pv. <i>alfalfae</i> : LMG 2214 (4); LMG 5039 (1); LMG 5040 (8) <i>P. marginalis</i> pv. <i>marginalis</i> : LMG 2210 ^T (9); LMG 2215 ^T (2); LMG 1243 (2); LMG 2211 (3); LMG 2212 (1); LMG 5170 (1); LMG 5173 (1); LMG 5174 (6); LMG 5176 (1); LMG 5177 (2); LMG 5178 (1); LMG 5180 (1); LMG 5181 (1); R-38928 (1); R-38929 (1); R-38930 (1); R-38931 (1); R-38932 (1); R-38933 (2); R-38934 (2); R-38935 (1); R 38936 (1); R-38937 (2); R-38938 (1) <i>P. marginalis</i> pv. <i>pastinacae</i> : LMG 2238 (7); LMG 5042 (3); LMG 5043 (3); LMG 5044 (3)	
<i>P. mediterranea</i> ^P	LMG 23075 ^T (10)	
<i>P. mendocina</i>	LMG 1223 ^T (8); LMG 5941 (3); LMG 6396 (1); R-9506 (1); R-9506 (1); R-16952 (1); R 17393 (1); R-17766 (1); R-17947 (1)	
<i>P. migulae</i>	LMG 21608 ^T (7); LMG 23195 (4)	
<i>P. monteilii</i>	LMG 21609 ^T (6)	
<i>P. mosselii</i>	LMG 21539 ^T (6)	
<i>P. mucidolens</i>	LMG 2223 ^T (6)	
<i>P. nitroreducens</i>	LMG 21614 ^T (5); LMG 20221 (6); LMG 21143 (1)	
<i>P. oleovorans</i>	LMG 2229 ^T (8); R-17305 (1); R-17338 (1); R-17791 (1)	
<i>P. oryzihabitans</i>	LMG 7040 ^T (6); LMG 5947 (2); LMG 18583 (3); LMG 18596 (1); LMG 18605 (2); LMG 18628 (1); R-16188 (1); R-17434 (1)	
<i>P. palleroniana</i>	LMG 23076 ^T (10)	
<i>P. peli</i>	LMG 23201 ^T (7); R-8840 (1)	
<i>P. pertucinogena</i>	LMG 1874 ^T (3); LMG 1875 (3)	

Table A.2 continued.

Species name	Strains (Number of profiles)	EC
<i>P. pictorum</i>	LMG 981 ^T (8)	
<i>P. plecoglossicida</i>	LMG 21750 ^T (6)	
<i>P. poae</i>	LMG 21465 ^T (6)	
<i>P. proteolytica</i>	LMG 22710 ^T (11)	
<i>P. pseudoalcaligenes</i>	LMG 1225 ^T (6); LMG 2854 (3); LMG 5516 (2); LMG 5517 (3); LMG 6036 (2); LMG 6037 (1); R-17968 (1)	
<i>P. psychrotolerans</i>	LMG 21977 ^T (6)	
<i>P. putida</i>	LMG 2257 ^T (14); LMG 1246 (3); LMG 2171 (2); LMG 2232 (1); LMG 2258 (3); LMG 2259 (1); LMG 5834 (1); LMG 5835 (10); LMG 9070 (2); LMG 14678 (2); LMG 14680 (4); LMG 14681 (2); LMG 14682 (3); LMG 14683 (1); LMG 16118 (1); LMG 16206 (1); LMG 16335 (4); LMG 18566 (1); LMG 18615 (1); R-4940 (1); R 4945 (1); R-4946 (1); R-4972 (1); R-4973 (1); R-4974 (1); R-4975 (1); R-4976 (1); R 16190 (2); R-16946 (1); R-16950 (1); R-16962 (2); R-17311 (1); R-17326 (1); R 17423 (1); R-17426 (1); R-17442 (1); R-17760 (1); R-17801 (1); R-17941 (1); R 17954 (1); R-38939 (1); R-38940 (2); R-38941 (2); R-38942 (3); R-38943 (3); R 38944 (1); R-38945 (1)	
<i>P. resinovorans</i>	LMG 2274 ^T (6)	
<i>P. rhizosphaerae</i>	LMG 21640 ^T (6)	
<i>P. rhodesiae</i>	LMG 17764 ^T (2); LMG 17765 (2); LMG 17766 (2)	
<i>P. salomonii</i> ^P	LMG 22120 ^T (10)	
<i>P. straminea</i>	LMG 21615 ^T (7); LMG 11723 (3)	
<i>P. stutzeri</i>	LMG 11199 ^T (9); LMG 1228 (7); LMG 2243 (2); LMG 2332 (1); LMG 2839 (2); LMG 5838 (1); LMG 6397 (1); LMG 10652 (1); LMG 14935 (2); LMG 18520 (1); LMG 18521 (1); LMG 18794 (1); R-16158 (2); R-16171 (1); R-16175 (1); R-16929 (1); R-16932 (1); R-16945 (1); R-17321 (1); R-17336 (1); R-17415 (1); R-17416 (1); R 17765 (1); R-17777 (1); R-17781 (1); R-17960 (1)	
<i>P. synxantha</i>	LMG 2190 ^T (4); R-38946 (1); R-38947 (1)	
<i>P. syringae</i> ^P	R-16948 (1); R-17971 (1)	
	<i>P. syringae</i> pv. <i>aceris</i> : LMG 2106 (3)	
	<i>P. syringae</i> pv. <i>aptata</i> : LMG 5059 (3); LMG 5095 (2); LMG 5532 (1)	
	<i>P. syringae</i> pv. <i>atrofaciens</i> : LMG 5533 (1)	
	<i>P. syringae</i> pv. <i>dysoxyli</i> : LMG 5062 (2); LMG 5542 (1)	
	<i>P. syringae</i> pv. <i>japonica</i> : LMG 5068 (2); LMG 5069 (1)	
	<i>P. syringae</i> pv. <i>lachrymans</i> : (should be enclosed in <i>P. amygdali</i>) LMG 21245 (2); LMG 21246 (2); LMG 21247 (2)	
	<i>P. syringae</i> pv. <i>lapsa</i> : LMG 2206 (2); LMG 5006 (1)	
	<i>P. syringae</i> pv. <i>panici</i> : LMG 2367 (4)	
	<i>P. syringae</i> pv. <i>papulans</i> : LMG 5076 (2); LMG 5077 (1)	
	<i>P. syringae</i> pv. <i>philadelphia</i> : (should be enclosed in <i>P. genomospecies3</i>) R-38950 (1)	
	<i>P. syringae</i> pv. <i>pisi</i> : LMG 5009 (1); LMG 5079 (2)	
	<i>P. syringae</i> pv. <i>porri</i> : (should be enclosed in " <i>P. coronafaciens</i> ") R-38949 (1)	
	<i>P. syringae</i> pv. <i>syringae</i> : LMG 1247 ^T (6); LMG 2230 (7); LMG 2231 (6); LMG 5082 (2); LMG 5083 (1); LMG 5086 (1); LMG 5087 (1); LMG 5189 (1); LMG 5190 (1); LMG 5570 (6); LMG 6108 (1); LMG 12643 (1); LMG 12648 (2)	
	<i>P. syringae</i> pv. <i>tomato</i> : (should be enclosed in <i>P. genomospecies3</i>) LMG 21249 (2)	
	<i>P. syringae</i> pv. <i>zizaniae</i> : (should be enclosed in <i>P. coronafaciens</i>) R-38948 (1)	
<i>P. taetrolens</i>	LMG 2336 ^T (6)	
<i>P. thermotolerans</i>	LMG 21284 ^T (6)	45°C
<i>P. thivervalensis</i>	LMG 21626 ^T (10)	

Table A.2 continued.

Species name	Strains (Number of profiles)	EC
<i>P. tolaasii</i> ^P	LMG 2342 ^T (5); LMG 2339 (4); LMG 2345 (7); LMG 2346 (2); LMG 2829 (2); LMG 6634 (1); LMG 6635 (1); LMG 6642 (1); LMG 12211 (3); LMG 12212 (4); LMG 12213 (2); LMG 12214 (2); LMG 12215 (2); LMG 12216 (2); LMG 12217 (2); LMG 12218 (2); LMG 12219 (2); LMG 12220 (3); LMG 18141 (2); LMG 18142 (2); LMG 18143 (2)	
<i>P. tremae</i> ^P	LMG 22121 ^T (8)	
<i>P. trivialis</i>	LMG 21464 ^T (6)	
<i>P. unsongensis</i>	LMG 21317 ^T (6)	
<i>P. vancouverensis</i>	LMG 20222 ^T (6)	
<i>P. veronii</i>	LMG 17761 ^T (4); LMG 17762 (1); LMG 23196 (2)	
<i>P. viridiflava</i> ^P	LMG 2352 ^T (4); LMG 2353 (2); LMG 5101 (6); LMG 5331 (1); LMG 6480 (1); LMG 12647 (1); R-17407 (1); R-17802 (1)	

A.2 Average FAME Profiles: Major Constituents

The following tables concern only the data sets of March 2008.

Average FAME peaks in the genus <i>Bacillus</i>	p. 218
Average FAME peaks in the genus <i>Paenibacillus</i>	p. 220
Average FAME peaks in the genus <i>Pseudomonas</i>	p. 222

Table A.3 continued.

Species name	C _{17:0}	C _{17:0} anteiso	C _{17:0} iso	C _{18:0}	C _{18:1} w7c	C _{18:1} w9c	iso C _{15:1} at 5	iso C _{17:1} w10c	summed feature 2	summed feature 3	summed feature 4
<i>B. atcatophilus</i>		7.798 (2.256)	3.215 (0.405)	0.115 (0.252)				0.858 (0.591)			0.465 (0.313)
<i>B. amyloliquefaciens</i>	0.038 (0.147)	7.095 (3.339)	9.287 (2.808)	0.115 (0.252)			0.040 (0.155)		0.045 (0.128)		0.274 (0.572)
<i>B. aquimaris</i>		3.500 (0.486)	0.581 (0.286)	0.066 (0.228)				0.682 (1.938)			0.833 (0.469)
<i>B. atropurpureus</i>		14.490 (5.997)	5.805 (1.190)			0.357 (0.618)		1.396 (0.438)			1.201 (0.327)
<i>B. bacilliformans</i>		1.323 (2.292)									
<i>B. badius</i>		2.414 (0.345)	2.869 (0.409)	0.105 (0.297)	0.250 (0.707)		0.728 (0.109)	4.856 (1.123)		0.330 (0.572)	3.211 (0.325)
<i>B. barbaricus</i>		2.030 (0.264)	0.802 (0.156)	0.064 (0.143)				0.380 (0.217)	0.388 (0.408)	3.805 (0.488)	0.832 (0.039)
<i>B. bataviensis</i>		1.205 (0.358)	1.566 (0.791)	2.789 (1.298)	0.051 (0.153)	1.327 (0.952)		1.926 (0.709)		2.244 (1.269)	0.794 (0.277)
<i>B. bogoriensis</i>		3.688 (0.496)	3.713 (0.868)					1.984 (0.295)		0.658 (0.734)	0.494 (0.411)
<i>B. carboniphilus</i>		1.227 (0.287)	2.192 (0.391)	0.225 (0.551)				6.552 (0.426)		0.823 (0.411)	0.823 (0.411)
<i>B. cereus</i>		0.383 (0.938)	6.016 (2.162)	0.275 (1.908)	0.116 (0.675)			3.580 (1.464)	2.261 (0.910)	1.354 (0.062)	2.502 (0.216)
<i>B. circulans</i>		3.606 (0.764)	1.191 (0.314)	0.123 (0.281)				0.544 (0.815)	0.445 (0.396)	1.201 (4.584)	0.043 (0.142)
<i>B. clausii</i>		5.034 (2.772)	9.556 (4.674)	0.219 (0.373)						0.254 (0.463)	0.254 (0.463)
<i>B. coagulans</i>		27.894 (7.808)	2.758 (1.958)	0.300 (0.877)						9.398 (2.308)	
<i>B. colnii</i>	0.080 (0.153)	3.656 (0.638)	2.718 (0.272)	1.424 (3.184)		0.550 (0.737)		8.529 (4.870)	0.144 (0.205)	1.354 (0.062)	2.502 (0.216)
<i>B. decolorationis</i>		11.272 (1.393)	1.919 (1.122)	0.635 (3.036)		2.923 (2.697)	0.005 (0.032)	4.662 (2.364)	0.011 (0.070)	1.201 (4.584)	0.110 (0.246)
<i>B. dretnensis</i>	0.092 (0.178)	1.072 (0.611)	2.657 (1.238)	1.919 (1.122)						0.981 (0.503)	0.981 (0.503)
<i>B. endophyticus</i>		4.282 (0.428)	0.857 (0.496)							0.156 (0.349)	0.156 (0.349)
<i>B. firmus</i>		3.495 (0.044)	0.875 (0.761)	1.859 (0.428)						3.936 (2.146)	3.936 (2.146)
<i>B. flexus</i>		2.166 (1.233)	1.859 (0.428)							5.233 (0.785)	5.233 (0.785)
<i>B. foraminis</i>	0.062 (0.125)	0.790 (0.640)	0.265 (0.347)	2.085 (0.318)	0.075 (0.150)	2.692 (3.112)		0.502 (0.734)		0.739 (0.619)	0.739 (0.619)
<i>B. fordii</i>	0.570 (0.591)	6.683 (0.116)	5.313 (0.571)	1.107 (0.163)		1.630 (0.229)		5.139 (0.304)		1.617 (0.243)	1.617 (0.243)
<i>B. fortis</i>	0.753 (0.693)	7.447 (1.510)	2.255 (1.572)	0.430 (0.638)		0.098 (0.195)		1.787 (0.219)		0.777 (0.040)	0.777 (0.040)
<i>B. clausii</i>	0.023 (0.121)	15.074 (3.305)	15.205 (2.744)	0.117 (0.492)				1.840 (1.760)		0.319 (0.675)	0.319 (0.675)
<i>B. fannarioli</i>		4.136 (0.299)	5.606 (0.676)					1.210 (0.177)		0.232 (0.215)	0.232 (0.215)
<i>B. faniculus</i>		4.659 (2.203)	0.895 (0.355)	0.613 (0.586)	0.824 (2.855)				0.122 (0.441)	2.207 (7.644)	
<i>B. galactosidiphilus</i>	0.228 (0.386)	9.958 (1.567)	12.570 (0.910)								
<i>B. gelatini</i>		0.909 (0.749)	1.800 (0.388)	0.240 (0.414)				8.172 (7.093)		3.888 (2.199)	3.888 (2.199)
<i>B. gibsonii</i>		4.560 (2.794)	5.212 (2.154)	0.110 (0.311)							
<i>B. halimipalans</i>		7.974 (3.551)	4.580 (2.998)								
<i>B. halodurans</i>		6.481 (2.001)	2.074 (0.713)								
<i>B. horikoshii</i>		1.384 (0.491)	7.662 (0.598)	0.236 (0.528)							
<i>B. horri</i>		2.278 (2.931)	0.658 (0.180)								
<i>B. humi</i>		1.257 (0.370)	0.993 (0.144)								
<i>B. indicus</i>		0.543 (1.436)	0.863 (2.283)								
<i>B. insolitus</i>		2.757 (0.922)	1.440 (0.466)	0.372 (0.593)							
<i>B. jeongii</i>	0.046 (0.174)	3.831 (1.990)	3.733 (4.733)	0.187 (0.424)		0.022 (0.083)	0.038 (0.094)	8.706 (0.549)		10.156 (2.436)	10.156 (2.436)
<i>B. lentus</i>		9.329 (2.287)	7.461 (1.257)	0.085 (0.438)				11.382 (1.854)		1.238 (0.229)	1.238 (0.229)
<i>B. licheniformis</i>		3.237 (0.612)	0.811 (0.314)					0.534 (0.606)		1.156 (0.881)	1.156 (0.881)
<i>B. luciferensis</i>		10.230 (0.272)	0.898 (0.109)					2.768 (0.317)		2.857 (0.141)	2.857 (0.141)
<i>B. marisflavi</i>		1.373 (1.033)	0.491 (0.634)					0.236 (0.624)		0.177 (0.469)	0.177 (0.469)
<i>B. megerium</i>	0.044 (0.234)	10.610 (2.677)	9.603 (1.435)	0.137 (0.223)				7.777 (1.415)		1.089 (1.063)	1.089 (1.063)
<i>B. mojavensis</i>	0.005 (0.021)	0.644 (0.451)	0.527 (0.362)	0.044 (0.133)				0.833 (1.117)		0.664 (2.486)	0.664 (2.486)
<i>B. muratis</i>		0.069 (0.229)	5.506 (1.427)					1.453 (0.797)		0.992 (0.577)	0.992 (0.577)
<i>B. mycoides</i>		0.912 (0.540)	2.258 (0.745)					0.014 (0.043)		0.837 (0.130)	0.837 (0.130)
<i>B. niacini</i>	0.112 (0.250)	1.497 (0.633)	1.056 (1.888)	0.240 (0.329)	0.240 (0.329)	1.130 (2.527)		0.030 (0.067)		1.537 (0.520)	1.537 (0.520)
<i>B. novatis</i>	0.006 (0.029)	7.125 (0.766)	3.787 (0.384)	0.661 (1.888)		0.011 (0.056)	0.053 (0.095)	1.466 (0.563)		0.196 (0.276)	0.196 (0.276)
<i>B. okuhidensis</i>		17.382 (0.699)	2.140 (0.441)	0.058 (0.143)				0.874 (1.897)	0.112 (0.405)	0.752 (0.501)	0.752 (0.501)
<i>B. oleronius</i>		7.188 (1.338)	5.242 (0.978)	0.103 (0.246)				8.840 (1.679)	0.723 (0.868)	7.815 (0.781)	7.815 (0.781)
<i>B. paenagionensis</i>	0.047 (0.162)	5.260 (4.144)	3.908 (2.129)							0.560 (1.878)	0.560 (1.878)
<i>B. pseudocitriciphilus</i>		4.026 (1.002)	0.522 (0.490)								
<i>B. pseudofirmus</i>		0.958 (0.785)	5.887 (0.354)								
<i>B. pseudomycolides</i>		2.043 (1.074)	0.427 (0.325)	0.540 (0.393)							
<i>B. psychrodurans</i>	1.033 (0.456)	0.707 (0.527)	0.570 (0.292)								
<i>B. psychrosaccharophilicus</i>		3.504 (2.146)	4.689 (1.781)	0.127 (0.336)							
<i>B. psychrotolerans</i>		0.707 (0.527)	1.012 (0.805)								
<i>B. pumilus</i>		0.536 (0.336)	1.096 (0.092)	0.012 (0.064)							
<i>B. pyrenis</i>		7.765 (5.179)	2.005 (1.341)	0.254 (0.568)							
<i>B. raris</i>	1.040 (0.761)	17.199 (2.348)	0.561 (0.527)	0.489 (1.293)							
<i>B. shackletonii</i>		1.250 (1.004)	2.258 (1.118)								
<i>B. silvestris</i>		1.910 (1.090)	1.012 (0.805)								
<i>B. simplex</i>		2.428 (0.247)	1.538 (0.157)	0.613 (0.163)							
<i>B. stralis</i>		16.353 (3.731)	12.847 (1.336)	0.377 (0.652)							
<i>B. smithii</i>		1.674 (0.297)	2.870 (0.527)	0.026 (0.074)							
<i>B. soli</i>		12.185 (1.681)	9.808 (0.496)	0.200 (0.234)							
<i>B. sonorensis</i>		13.665 (4.592)	2.506 (0.794)	0.385 (0.656)							
<i>B. sponothermodurans</i>	0.171 (0.321)	4.096 (0.620)	3.044 (0.704)			0.028 (0.114)					
<i>B. subterraneus</i>		10.486 (3.257)	10.227 (3.050)	0.157 (0.332)							
<i>B. subtilis</i>		8.907 (3.524)	11.819 (4.343)	1.101 (1.616)							
<i>B. thermocampylovarans</i>	0.287 (0.373)	13.358 (1.658)	22.326 (6.846)	0.432 (0.794)							
<i>B. thermocantarcticus</i>		0.178 (0.421)	4.893 (1.308)	0.287 (0.993)							
<i>B. thuringiensis</i>		9.033 (3.843)	10.666 (4.887)	0.159 (0.211)							
<i>B. vallis mortis</i>		2.401 (0.453)	5.591 (0.612)	0.079 (0.208)							
<i>B. vireti</i>		0.132 (0.323)	5.538 (0.843)								
<i>B. weltheistephanensis</i>				3.714 (9.802)							
									2.771 (0.358)	10.744 (1.632)	0.469 (0.433)
									1.853 (4.920)	2.555 (6.755)	0.590 (0.270)
									0.320 (0.504)	7.100 (1.012)	0.307 (0.751)

Species name	C _{12:0}	C _{13:0 anteiso}	C _{13:0 iso}	C _{14:0}	C _{14:0 iso}	C _{15:0 anteiso}	C _{15:0 iso}	C _{16:0}
<i>Pa. alvei</i>		0.053 (0.120)	0.174 (0.258)	3.357 (1.168)	0.636 (0.582)	50.352 (6.171)	11.734 (4.106)	13.051 (3.785)
<i>Pa. amylohydrolyticus</i>		0.380 (0.255)	0.080 (0.160)	15.185 (1.367)	3.822 (0.127)	44.520 (4.024)	5.298 (0.726)	11.410 (1.435)
<i>Pa. anaericanus</i>	0.983 (0.573)	0.423 (0.220)	0.063 (0.155)	15.900 (5.613)	3.738 (1.148)	29.252 (8.008)	6.380 (3.773)	34.072 (12.262)
<i>Pa. antarcticus</i>		0.223 (0.193)	0.267 (0.231)	17.443 (0.743)	1.567 (0.145)	36.770 (6.062)	11.657 (2.041)	13.883 (8.831)
<i>Pa. apitarius</i>		0.042 (0.094)		1.660 (0.538)	1.050 (0.163)	50.500 (1.629)	13.546 (0.810)	4.944 (1.599)
<i>Pa. azoreducens</i>	1.443 (0.510)	0.070 (0.140)		8.498 (1.991)	1.795 (0.268)	37.800 (2.853)	6.617 (0.301)	20.233 (1.761)
<i>Pa. borealis</i>	0.102 (0.205)			18.120 (1.084)	4.925 (0.552)	38.845 (3.108)	13.988 (1.459)	10.293 (2.615)
<i>Pa. cellulohydrolyticus</i>		1.190 (0.581)		2.096 (0.711)	3.962 (0.456)	49.095 (2.151)	1.734 (1.074)	8.372 (1.791)
<i>Pa. chinensis</i>	0.361 (0.457)			5.159 (1.924)	1.659 (0.302)	42.892 (2.882)	7.156 (0.988)	18.011 (3.067)
<i>Pa. chitinolyticus</i>				1.827 (0.350)	5.007 (0.568)	58.441 (1.082)	6.814 (0.649)	5.403 (1.007)
<i>Pa. cinereus</i>			0.555 (0.610)	2.120 (0.884)	1.197 (0.383)	38.652 (8.817)	28.580 (22.404)	8.397 (7.436)
<i>Pa. cookii</i>	0.608 (0.503)	0.035 (0.086)		3.967 (1.141)	1.353 (0.064)	41.778 (2.590)	5.960 (0.716)	10.463 (3.503)
<i>Pa. curdlanolyticus</i>		0.072 (0.100)		2.218 (1.095)	3.306 (1.178)	55.768 (4.037)	10.768 (6.408)	8.750 (1.164)
<i>Pa. dendritiformis</i>	0.028 (0.069)	0.032 (0.078)	0.028 (0.063)	1.288 (0.251)	0.300 (0.148)	48.213 (1.544)	7.063 (0.655)	6.847 (1.109)
<i>Pa. durus</i>		0.710 (0.420)		8.100 (2.813)	3.992 (2.602)	49.390 (11.200)	3.658 (1.903)	20.407 (4.916)
<i>Pa. elgii</i>				1.442 (0.421)	3.414 (0.516)	56.202 (1.766)	11.218 (0.739)	4.630 (1.074)
<i>Pa. favisporus</i>	0.044 (0.124)			2.013 (0.248)	1.190 (0.177)	51.846 (1.342)	4.174 (1.801)	12.289 (2.373)
<i>Pa. fonticola</i>				7.312 (1.675)	5.128 (0.954)	36.274 (1.685)	3.850 (2.124)	27.708 (2.107)
<i>Pa. ginsengarvii</i>				0.463 (0.364)	5.553 (0.554)	42.117 (0.752)	15.352 (0.540)	3.510 (0.305)
<i>Pa. gluconohydrolyticus</i>				2.534 (1.094)	1.184 (0.189)	55.129 (4.612)	6.003 (2.445)	16.824 (6.440)
<i>Pa. humicus</i>				1.520 (0.151)	2.843 (0.229)	55.247 (0.917)	17.797 (0.452)	4.347 (0.280)
<i>Pa. ilinoisensis</i>		0.150 (0.135)	0.215 (0.236)	4.965 (0.507)	3.083 (0.260)	49.867 (2.685)	10.345 (1.735)	9.558 (2.670)
<i>Pa. jamilae</i>		0.148 (0.233)		1.933 (0.405)	2.933 (0.306)	61.540 (1.400)	5.820 (0.310)	5.837 (0.310)
<i>Pa. kobensis</i>			0.060 (0.104)	0.810 (0.704)	1.457 (1.319)	58.000 (6.282)	3.983 (1.485)	6.583 (4.890)
<i>Pa. lactis</i>			0.038 (0.134)	3.773 (0.828)	2.603 (0.250)	38.920 (1.449)	12.467 (0.667)	23.237 (1.695)
<i>Pa. larvae</i>	0.019 (0.097)			2.036 (1.429)	0.483 (0.539)	38.024 (5.666)	15.981 (4.537)	10.381 (4.190)
<i>Pa. lautus</i>				2.737 (0.335)	1.487 (0.231)	49.213 (6.660)	5.827 (1.584)	24.528 (5.054)
<i>Pa. macerans</i>	0.872 (0.692)	0.061 (0.170)		6.653 (2.291)	4.383 (1.530)	31.066 (3.172)	8.883 (5.321)	19.934 (6.282)
<i>Pa. macquariensis</i>	0.072 (0.161)	2.036 (0.971)	0.394 (0.391)	6.460 (2.961)	2.094 (0.325)	57.910 (7.690)	15.458 (2.919)	9.758 (3.695)
<i>Pa. mendelii</i>				2.322 (0.129)	2.972 (0.326)	55.033 (0.966)	15.245 (1.379)	8.203 (0.297)
<i>Pa. odorifer</i>			0.190 (0.220)	4.740 (0.427)	4.280 (0.169)	44.373 (6.601)	15.965 (2.286)	14.098 (0.907)
<i>Pa. pabuli</i>		0.081 (0.171)		5.520 (2.342)	3.786 (0.371)	58.486 (6.042)	8.698 (2.003)	8.612 (3.298)
<i>Pa. panactisoli</i>		0.180 (0.091)	0.028 (0.069)	7.818 (1.084)	2.778 (0.261)	51.990 (2.568)	6.757 (0.456)	20.437 (2.067)
<i>Pa. peoriae</i>		0.292 (0.273)	0.806 (0.276)	2.752 (1.146)	3.320 (1.663)	52.886 (3.673)	12.342 (1.465)	5.840 (0.878)
<i>Pa. phyllospphaerae</i>		0.022 (0.053)		1.892 (0.119)	2.182 (0.318)	56.177 (1.834)	7.035 (1.190)	10.970 (1.685)
<i>Pa. polymyxa</i>		0.016 (0.085)	0.041 (0.215)	2.214 (0.565)	1.704 (0.595)	62.415 (5.587)	7.765 (1.899)	6.849 (2.445)
<i>Pa. rhizospherae</i>				3.262 (0.319)	1.350 (0.126)	48.198 (0.929)	10.720 (0.653)	12.546 (1.463)
<i>Pa. soli</i>		0.070 (0.121)		2.660 (0.411)	4.050 (1.911)	66.143 (7.099)	4.043 (0.514)	7.380 (3.295)
<i>Pa. stellifer</i>	0.030 (0.060)	0.103 (0.118)	0.427 (0.039)	10.068 (1.455)	4.897 (0.304)	23.133 (1.147)	11.355 (0.954)	40.053 (1.279)
<i>Pa. taiwanensis</i>		0.817 (0.986)	0.620 (0.724)	1.168 (0.132)	2.863 (0.290)	49.688 (3.596)	11.138 (0.651)	2.665 (0.383)
<i>Pa. thiaminolyticus</i>	0.186 (0.190)	0.082 (0.244)	0.025 (0.087)	2.865 (1.051)	0.764 (0.270)	46.291 (5.640)	8.649 (8.457)	9.474 (4.271)
<i>Pa. validus</i>		0.018 (0.075)		2.026 (0.458)	3.095 (0.308)	52.101 (2.330)	12.568 (2.219)	5.671 (1.566)
<i>Pa. wynnii</i>				5.662 (1.280)	5.025 (1.277)	32.452 (2.942)	5.057 (0.859)	34.660 (7.473)
<i>Pa. xylanolyticus</i>		0.142 (0.164)	0.202 (0.235)	4.433 (0.803)	2.673 (0.326)	48.178 (3.056)	10.085 (1.575)	12.505 (1.818)

Table A.4: Average FAME peaks in the genus *Paenibacillus*. Averages and standard deviations of the peaks with a prevalence in more than ten species.

Table A.4 continued.

Species name	C _{16:0} iso	C _{16:1} w11c	C _{16:1} w7c alcohol	C _{17:0}	C _{17:0} anteiso	C _{17:0} iso	C _{18:0}	iso C _{17:1} w10c	summed feature 4
<i>Pa. alvei</i>	2.803 (1.456)	6.486 (1.722)	0.608 (0.580)	0.114 (0.311)	4.870 (1.626)	3.534 (1.636)		0.756 (0.663)	0.891 (0.610)
<i>Pa. anyolyticus</i>	1.750 (0.271)	16.837 (0.676)	0.177 (0.355)		0.448 (0.519)	0.092 (0.185)			
<i>Pa. anaericanus</i>	4.025 (1.162)	1.582 (1.863)			1.125 (0.136)	1.512 (0.322)	0.342 (0.433)		
<i>Pa. antarcticus</i>	0.833 (0.140)	15.900 (1.975)	0.270 (0.235)		0.390 (0.096)	0.497 (0.240)			
<i>Pa. apiarius</i>	5.034 (0.577)	3.878 (0.820)	1.234 (0.206)		7.730 (1.094)	4.974 (0.303)		2.204 (0.269)	1.910 (0.264)
<i>Pa. azoreducens</i>	9.675 (1.417)				9.918 (0.769)	3.438 (0.398)			
<i>Pa. borealis</i>	7.942 (1.388)	1.420 (1.417)			2.033 (0.302)	2.325 (0.368)			
<i>Pa. cellulolyticus</i>	23.718 (0.860)	0.411 (0.348)	1.447 (0.685)		6.490 (0.785)	0.571 (0.481)	0.548 (1.232)		
<i>Pa. chibensis</i>	10.804 (1.455)	0.886 (0.627)			8.812 (0.419)	3.869 (0.563)	0.320 (0.770)		
<i>Pa. chitinolyticus</i>	11.259 (1.172)	3.849 (0.677)	1.081 (0.280)		4.243 (0.254)	2.019 (0.202)			
<i>Pa. cineris</i>	6.330 (5.972)	1.612 (1.019)	0.227 (0.352)		6.878 (3.530)	3.790 (0.626)		0.054 (0.144)	0.457 (0.509)
<i>Pa. cookii</i>	12.733 (0.781)				19.668 (1.600)	3.433 (0.548)		1.207 (1.335)	
<i>Pa. curdellanolyticus</i>	13.040 (6.159)		0.138 (0.201)	0.060 (0.134)	3.618 (1.083)	2.234 (0.731)			
<i>Pa. dendritiformis</i>	3.912 (0.044)	5.185 (0.440)	0.755 (0.105)	0.718 (0.171)	15.152 (1.256)	5.650 (0.389)	0.258 (0.293)	1.940 (0.071)	2.660 (0.476)
<i>Pa. durus</i>	5.318 (3.076)				1.976 (1.211)	0.244 (0.370)	0.109 (0.216)		
<i>Pa. elgii</i>	7.202 (0.613)	6.346 (0.965)	1.608 (0.905)		4.722 (0.345)	2.766 (0.172)		0.448 (0.615)	
<i>Pa. favisporus</i>	11.424 (1.550)	0.219 (0.311)			14.555 (0.929)	2.094 (0.797)	0.155 (0.438)		
<i>Pa. fonticola</i>	10.130 (1.050)	4.320 (1.227)	0.068 (0.152)	0.066 (0.148)	2.416 (0.445)	0.952 (1.066)	0.114 (0.255)		
<i>Pa. ginsengarvi</i>	22.270 (0.818)			0.450 (0.525)	4.732 (0.360)	5.752 (0.412)			
<i>Pa. glucanolyticus</i>	4.139 (0.609)	5.303 (0.899)	0.221 (0.294)		5.551 (1.132)	2.139 (1.288)	0.353 (0.934)	0.427 (0.420)	0.193 (0.331)
<i>Pa. humicus</i>	10.797 (0.384)				4.390 (0.300)	2.583 (0.201)	0.320 (0.554)		
<i>Pa. illinoisensis</i>	4.042 (0.608)	10.773 (0.876)	0.075 (0.184)		3.030 (0.401)	3.402 (0.544)	0.100 (0.245)	0.398 (0.459)	
<i>Pa. jamilae</i>	10.103 (1.040)	2.840 (0.249)	0.800 (0.044)		6.017 (0.951)	2.053 (0.248)		0.137 (0.237)	
<i>Pa. kobensis</i>	9.797 (7.247)				8.177 (1.198)	2.137 (2.125)	3.510 (3.277)		
<i>Pa. lactis</i>	5.210 (0.130)	3.847 (0.792)	0.290 (0.062)	0.671 (1.072)	4.563 (0.241)	4.807 (0.609)	0.100 (0.173)	0.067 (0.115)	
<i>Pa. larvae</i>	6.081 (2.803)	0.182 (0.802)			14.254 (4.502)	7.657 (2.529)	0.117 (0.301)	0.100 (0.445)	0.026 (0.162)
<i>Pa. lautus</i>	4.140 (1.087)	3.712 (1.946)			5.355 (0.831)	3.003 (1.284)			
<i>Pa. macerans</i>	14.261 (4.240)	0.137 (0.529)			7.912 (2.644)	5.707 (2.539)	0.079 (0.305)		
<i>Pa. macquariensis</i>	1.652 (0.354)	1.934 (2.285)			0.594 (0.405)	0.578 (0.366)			
<i>Pa. mendelii</i>	9.123 (0.771)	0.515 (0.566)			3.588 (0.175)	3.003 (0.289)			
<i>Pa. odorifer</i>	5.040 (0.819)	3.010 (1.047)	0.090 (0.180)		1.893 (0.521)	2.958 (0.340)	3.368 (6.156)		
<i>Pa. pabuli</i>	3.957 (1.301)	6.490 (3.087)	0.332 (0.550)		2.031 (0.510)	1.563 (0.481)	0.171 (0.361)	0.272 (0.481)	
<i>Pa. panacisoli</i>	3.508 (0.510)	2.098 (0.390)	0.035 (0.086)	0.080 (0.124)	2.688 (0.312)	1.370 (0.113)	0.228 (0.279)		
<i>Pa. peoriae</i>	6.950 (2.267)	2.462 (1.338)	0.400 (0.518)	0.056 (0.125)	5.196 (1.566)	5.434 (2.021)	0.638 (0.882)	0.624 (0.216)	
<i>Pa. phyllosphaerae</i>	8.507 (0.893)	1.835 (0.300)	1.828 (0.112)		6.417 (0.488)	1.547 (0.174)	0.197 (0.330)		1.395 (0.111)
<i>Pa. polymyxa</i>	6.144 (1.601)	2.554 (0.934)	0.309 (0.382)		6.575 (1.515)	3.031 (1.471)	0.025 (0.092)	0.266 (0.333)	0.066 (0.166)
<i>Pa. rhizosphaerae</i>	9.352 (1.202)	0.818 (0.237)			9.542 (0.464)	4.208 (0.237)			
<i>Pa. soli</i>	8.690 (0.580)			0.087 (0.150)	2.517 (2.372)	0.633 (0.616)			
<i>Pa. strelifer</i>	6.543 (0.527)				1.012 (0.095)	1.952 (0.272)	0.430 (0.476)		
<i>Pa. taiwanensis</i>	9.055 (1.432)	2.783 (0.451)	2.220 (0.301)	0.447 (0.350)	8.122 (1.186)	3.783 (0.941)		2.398 (0.319)	2.077 (0.478)
<i>Pa. thiaminolyticus</i>	3.890 (1.186)	6.669 (2.982)	0.690 (0.443)	0.067 (0.142)	12.993 (3.069)	3.757 (1.804)	0.062 (0.131)	1.167 (0.556)	2.189 (1.159)
<i>Pa. validus</i>	6.427 (0.931)	6.130 (1.040)	2.365 (0.336)		4.098 (0.591)	3.182 (0.843)		0.819 (0.346)	0.566 (0.185)
<i>Pa. wynnii</i>	4.875 (0.655)	6.973 (4.772)	0.622 (0.484)		0.780 (0.630)	1.217 (0.946)	2.175 (1.393)		
<i>Pa. xylanolyticus</i>	4.170 (0.272)	9.318 (1.112)	0.047 (0.114)		3.507 (0.658)	4.092 (0.605)	0.165 (0.404)	0.488 (0.261)	

Species name	C _{8:0} 3OH	C _{10:0}	C _{10:0} 3OH	C _{11:0} 3OH	C _{11:0} iso 3OH	C _{12:0}	C _{12:0} 2OH	C _{12:0} 3OH	C _{12:1} 3OH	C _{13:0} iso	C _{14:0}	C _{15:0} anteiso	fa15:0 iso
<i>P. abietaniphila</i>	0.074 (0.141)	0.161 (0.116)	4.739 (1.196)	0.023 (0.058)	0.036 (0.199)	4.406 (1.457)	4.856 (0.993)	5.726 (0.768)	0.274 (0.313)	0.007 (0.046)	0.398 (0.034)	0.160 (1.345)	0.022 (0.050)
<i>P. aernigmosa</i>	0.008 (0.038)	0.167 (0.356)	4.966 (2.333)			3.883 (1.155)	6.073 (2.545)	5.820 (1.771)	0.143 (0.327)		0.881 (0.716)		0.457 (3.732)
<i>P. agarici</i>		0.016 (0.046)	5.231 (0.673)			2.386 (0.414)	6.661 (0.990)	5.639 (0.690)			1.170 (0.449)		
<i>P. altigenes</i>		2.404 (1.509)	5.323 (1.498)	0.279 (0.150)	0.177 (0.173)	6.264 (1.719)	0.070 (0.077)	3.874 (2.245)	0.039 (0.066)	0.242 (0.041)	3.881 (1.287)		0.193 (0.164)
<i>P. atlaphila</i>		0.236 (0.051)	6.348 (1.616)	0.419 (0.145)	0.301 (0.072)	10.382 (1.723)	0.146 (0.047)	5.552 (1.116)		0.005 (0.030)	0.271 (0.023)	0.067 (0.586)	0.022 (0.198)
<i>P. amygdali</i>		0.049 (0.067)	4.397 (1.427)		0.253 (1.079)	5.180 (1.300)	3.561 (0.096)	5.014 (1.365)	0.178 (0.308)		0.234 (0.348)		0.008 (0.023)
<i>P. anguillicolapica</i>		0.848 (0.286)	6.583 (2.405)	0.007 (0.020)	0.154 (0.216)	5.352 (1.271)	0.039 (0.061)	6.244 (2.157)			0.764 (0.028)		
<i>P. antarctica</i>		0.096 (0.044)	6.406 (1.126)	0.061 (0.069)	0.061 (0.069)	2.212 (0.291)	7.718 (1.305)	6.622 (1.074)			0.326 (0.028)		
<i>P. argentinensis</i>		0.171 (0.018)	4.874 (0.747)	0.211 (0.036)	0.027 (0.047)	12.243 (1.480)	0.153 (0.028)	5.577 (0.604)	0.025 (0.049)	0.006 (0.015)	0.288 (0.066)		
<i>P. asplenii</i>	0.006 (0.022)	0.087 (0.081)	5.598 (0.755)			1.839 (0.344)	7.548 (1.110)	5.987 (0.906)			0.128 (0.148)		
<i>P. avellanica</i>		4.447 (1.628)				5.840 (1.614)	3.938 (1.366)	5.562 (1.997)			0.267 (0.122)		
<i>P. asoniformans</i>		0.083 (0.085)	5.626 (2.810)		0.500 (0.043)	8.422 (0.638)	0.057 (0.063)	3.738 (0.370)			0.845 (0.110)	14.543 (3.834)	0.248 (0.041)
<i>P. balnearia</i>		0.187 (0.031)	3.137 (0.293)	0.137 (0.030)	2.313 (0.508)	4.427 (0.429)	0.110 (0.033)	3.347 (0.394)	0.268 (0.090)	0.385 (0.040)	0.578 (0.053)		33.247 (2.369)
<i>P. borebori</i>		0.157 (0.032)	4.505 (0.929)	0.097 (0.071)	1.814 (0.402)	0.009 (0.023)	3.289 (0.735)	3.660 (0.579)	0.129 (0.096)	0.020 (0.031)	0.403 (0.076)		
<i>P. brassiacaeorum</i>	0.050 (0.056)	0.699 (0.167)	0.129 (0.092)			4.768 (0.708)	6.061 (1.820)	6.975 (1.375)		1.009 (0.106)	0.866 (0.252)	3.921 (0.315)	46.411 (3.153)
<i>P. bremeri</i>		0.204 (0.044)	5.895 (1.316)	0.018 (0.057)	0.012 (0.038)	2.731 (0.232)	7.904 (1.647)	7.074 (1.369)			0.541 (0.057)		
<i>P. camubina</i>		0.114 (0.086)	6.207 (1.455)			5.520 (1.451)	3.629 (0.909)	5.103 (1.587)			0.206 (0.123)		
<i>P. caricapapuae</i>		0.033 (0.042)	3.954 (1.204)			5.600 (0.326)	3.644 (0.718)	4.944 (0.904)			0.068 (0.093)		
<i>P. chloridismutans</i>		0.040 (0.055)	4.282 (0.787)	0.020 (0.045)	0.120 (0.022)	9.229 (1.188)	0.104 (0.049)	3.660 (0.579)		0.129 (0.021)	1.022 (0.061)		
<i>P. chlorographis</i>	0.150 (0.186)	0.232 (0.043)	3.777 (0.641)	0.023 (0.056)	0.035 (0.058)	1.758 (0.810)	7.833 (2.165)	8.584 (2.688)	2.075 (1.915)	0.502 (0.049)	0.532 (0.192)	0.043 (0.191)	0.051 (0.221)
<i>P. cichorii</i>	0.018 (0.060)	0.132 (0.093)	3.941 (1.060)			5.422 (1.231)	3.697 (1.121)	5.178 (1.248)	0.270 (0.337)	0.415 (0.320)	1.470 (1.011)	5.275 (3.294)	24.753 (15.405)
<i>P. cissicola</i>		0.212 (0.234)	2.624 (1.696)			4.303 (0.664)	4.821 (0.985)	5.059 (0.908)			0.783 (0.062)		
<i>P. citranthalis</i>		0.077 (0.064)	3.953 (0.682)			5.727 (0.604)	3.925 (0.631)	5.395 (0.603)			0.332 (0.029)		
<i>P. conglans</i>		0.014 (0.049)	5.070 (1.303)	0.025 (0.039)	0.070 (0.120)	5.405 (0.706)	3.442 (0.975)	4.824 (1.316)			0.402 (0.775)		
<i>*P. coronificans*</i>		0.303 (0.261)	6.096 (1.086)			5.310 (1.258)	4.735 (1.142)	5.681 (1.031)	0.585 (0.643)		0.145 (0.142)		
<i>P. corrigata</i>	0.247 (0.372)	0.067 (0.078)	5.675 (0.828)			2.085 (0.139)	8.343 (1.540)	6.593 (1.091)	0.018 (0.045)		0.602 (0.065)		
<i>P. costaninii</i>	0.028 (0.044)	0.123 (0.062)	6.887 (1.131)	0.052 (0.105)	0.268 (0.054)	2.020 (0.075)	10.940 (2.207)	7.970 (1.478)			0.333 (0.005)		
<i>P. cremoricolorata</i>		0.123 (0.098)	4.738 (0.963)			2.747 (0.155)	7.015 (1.850)	6.108 (1.335)			0.430 (0.031)		
<i>P. extremorientalis</i>		0.121 (0.065)	4.729 (0.928)	0.110 (0.052)		10.601 (2.087)	0.101 (0.073)	4.764 (0.984)			0.720 (0.104)		
<i>P. flavescens</i>		0.115 (0.127)	5.866 (3.409)			1.993 (1.565)	1.323 (0.044)	0.020 (0.049)		1.387 (2.151)	2.325 (0.546)	8.420 (13.075)	5.163 (8.019)
<i>P. floccens</i>	0.035 (0.239)	0.084 (0.135)	6.688 (1.682)	0.002 (0.017)	0.007 (0.033)	3.237 (1.352)	6.633 (2.485)	6.104 (1.733)	0.166 (0.486)	0.002 (0.022)	0.421 (0.121)	0.004 (0.034)	0.002 (0.029)
<i>P. fluorescens</i>		0.067 (0.078)	9.880 (1.370)	0.003 (0.015)	0.183 (0.166)	3.875 (0.892)	6.697 (1.741)	6.676 (1.534)			0.599 (0.355)		
<i>P. fragi</i>	0.417 (0.058)		3.707 (3.241)			2.810 (0.349)	6.203 (0.350)	9.382 (1.253)	6.962 (1.439)		0.363 (0.071)		
<i>P. frederiksbergensis</i>			3.707 (3.241)			0.427 (0.383)	8.597 (1.612)	5.953 (1.135)			0.903 (0.138)		
<i>P. fulva</i>		0.106 (0.322)	6.649 (2.061)	0.025 (0.040)		1.946 (0.605)	7.202 (1.599)	5.895 (1.425)	0.255 (0.461)		0.416 (0.345)	0.015 (0.103)	0.011 (0.075)
<i>P. fuscovaginatae</i>		0.933 (0.170)	0.262 (0.045)			5.570 (1.007)	0.025 (0.061)	4.153 (0.765)		0.367 (0.085)	4.588 (0.495)	0.015 (0.103)	0.011 (0.075)
<i>P. gentaculata</i>		0.086 (0.084)	4.513 (1.830)	0.021 (0.064)		5.686 (0.896)	3.666 (1.282)	5.354 (1.861)			0.229 (0.107)	9.685 (1.082)	30.452 (0.592)
<i>P. genomospecies3</i>		0.091 (0.059)	4.295 (1.390)	0.013 (0.035)		1.789 (0.254)	4.174 (1.378)	5.514 (1.679)			0.210 (0.031)	0.012 (0.076)	0.009 (0.058)
<i>P. genomospecies7</i>		0.280 (0.088)	4.512 (0.879)			5.102 (0.529)	8.621 (1.980)	7.050 (1.516)	0.022 (0.047)		0.821 (0.055)	0.014 (0.044)	0.046 (0.154)
<i>P. gossardii</i>	0.018 (0.039)	0.010 (0.033)	3.455 (0.696)	0.033 (0.056)			3.699 (0.822)	5.192 (1.114)			0.292 (0.111)	12.393 (2.433)	36.222 (2.621)
<i>P. graminis</i>		0.678 (0.098)	0.190 (0.025)	0.025 (0.042)			0.038 (0.063)	3.297 (0.372)		0.460 (0.093)	3.183 (0.231)		
<i>P. hibiscicola</i>	0.102 (0.122)	0.770 (0.489)	7.280 (2.512)	0.457 (0.180)		10.594 (1.162)	0.187 (0.058)	4.968 (1.113)	1.778 (1.157)	0.215 (0.169)	0.637 (0.037)		
<i>P. indica</i>	0.020 (0.049)	0.158 (0.036)	6.225 (1.897)	0.022 (0.053)		5.362 (0.769)	4.853 (1.608)	6.398 (1.709)	0.102 (0.080)		0.435 (0.116)		
<i>P. jessenii</i>		0.207 (0.032)	5.278 (0.724)	0.403 (0.033)		7.297 (0.617)	2.072 (0.424)	5.187 (0.824)			0.622 (0.092)	0.212 (0.518)	0.237 (0.580)

Table A.5: Average FAME peaks in the genus *Pseudomonas*. Averages and standard deviations of the peaks with a prevalence in more than twenty species.

Table A.5 continued.

Species name	C _{16:0}	C _{16:0 iso}	C _{16:1 w5c}	C _{17:0}	C _{17:0 cyclo}	C _{17:0 iso}	C _{17:1 w8c}	C _{18:0}	C _{18:1 w7c}	C _{19:0 cyclo w8c}	summ. feat. 2	summ. feat. 3	unknown 11.799	unknown 12.484
<i>P. abietaniphila</i>	28.244 (2.541)		0.022 (0.047)	0.184 (0.176)	11.203 (7.338)	0.071 (0.093)	0.066 (0.106)	0.482 (0.106)	11.851 (1.347)	0.471 (0.432)	0.080 (0.136)	26.362 (8.295)	0.030 (0.181)	0.154 (0.139)
<i>P. aegrinosa</i>	21.847 (3.854)	0.015 (0.124)	0.005 (0.022)	0.243 (0.175)	0.683 (1.146)	0.081 (0.402)	0.468 (1.643)	0.513 (0.501)	34.085 (9.079)	0.834 (0.828)	0.082 (0.854)	17.196 (3.558)		0.051 (0.109)
<i>P. agrosi</i>	34.099 (1.699)			0.344 (0.181)	11.110 (4.031)	0.030 (0.085)		0.450 (0.219)	7.161 (0.956)	0.111 (0.165)		25.547 (4.936)		
<i>P. alcaligenes</i>	15.844 (5.953)	0.033 (0.087)	0.094 (0.142)	0.600 (0.713)	0.288 (0.999)	0.083 (0.136)	0.892 (0.490)	0.133 (0.220)	30.392 (6.352)			27.005 (4.788)	0.110 (0.089)	
<i>P. alcaliphila</i>	12.180 (1.634)	0.389 (0.155)	0.020 (0.042)	0.813 (0.130)		0.506 (0.084)	1.653 (0.305)	0.454 (0.121)	34.649 (3.897)	0.042 (0.384)	0.011 (0.075)	23.295 (9.915)	0.295 (0.074)	0.013 (0.041)
<i>P. angulata</i>	25.952 (2.842)	0.011 (0.064)		0.051 (0.087)	0.887 (1.423)	0.213 (0.307)	0.011 (0.043)	0.949 (0.633)	15.951 (2.679)		0.020 (0.042)	36.241 (2.673)		0.006 (0.024)
<i>P. anguillicepica</i>	18.981 (2.649)	0.102 (0.211)	0.023 (0.046)	0.327 (0.324)		0.361 (0.389)	0.159 (0.173)	0.566 (0.931)	28.792 (5.238)	0.338 (0.212)		25.160 (2.137)	5.237 (1.472)	0.052 (0.085)
<i>P. antarctica</i>	26.474 (1.596)		0.008 (0.029)	0.473 (0.040)	6.319 (3.334)	0.142 (0.095)	0.070 (0.099)	0.710 (0.167)	15.556 (1.540)			26.576 (3.407)		0.026 (0.046)
<i>P. argentinensis</i>	16.089 (0.931)	0.070 (0.069)		0.473 (0.040)		0.083 (0.078)	0.703 (0.082)	0.376 (0.041)	35.416 (3.392)		0.196 (0.032)	31.944 (0.102)	0.196 (0.032)	0.099 (0.072)
<i>P. asplenii</i>	25.307 (1.553)		0.007 (0.025)	0.108 (0.090)	1.295 (0.903)			0.758 (0.150)	15.202 (2.338)			35.903 (1.187)		0.041 (0.052)
<i>P. avellanca</i>	25.970 (3.586)					0.052 (0.105)		0.867 (0.156)	14.050 (2.391)			38.807 (0.259)		
<i>P. azotofornans</i>	29.527 (3.209)					0.064 (0.114)		0.764 (0.348)	24.223 (10.450)	0.529 (0.441)		19.820 (3.507)	0.048 (0.055)	0.029 (0.049)
<i>P. balnearica</i>	21.870 (1.620)			0.223 (0.160)	5.997 (2.334)	2.190 (0.143)		0.327 (0.063)	12.971 (1.658)	4.647 (2.070)		19.820 (3.507)	0.013 (0.033)	
<i>P. bebeli</i>	4.863 (0.735)	1.265 (0.186)			0.378 (0.608)	2.217 (0.599)		0.463 (0.121)	27.407 (2.605)			10.695 (0.832)	1.875 (0.353)	
<i>P. borbori</i>	19.482 (0.354)		0.782 (0.243)		0.365 (0.382)	0.173 (0.040)	0.057 (0.091)	0.648 (0.138)	18.892 (1.779)			42.400 (1.558)	2.998 (0.400)	0.082 (0.065)
<i>P. boreopalis</i>	1.523 (0.416)	2.120 (0.328)			0.206 (0.544)	3.680 (0.652)	0.371 (0.079)	0.273 (0.044)	18.025 (1.195)			8.444 (2.67)	2.736 (0.562)	
<i>P. brassicacearum</i>	27.478 (2.275)			0.031 (0.066)	5.137 (2.719)			0.405 (0.125)	12.383 (3.088)		0.015 (0.047)	30.017 (2.987)		0.038 (0.064)
<i>P. bremeri</i>	26.409 (1.882)		0.010 (0.032)	0.401 (0.201)	4.152 (2.845)	0.033 (0.070)	0.076 (0.124)	0.622 (0.160)	11.391 (1.091)	0.030 (0.095)		32.277 (2.635)		
<i>P. cissicola</i>	3.871 (1.482)		0.013 (0.034)					0.946 (0.461)	15.054 (3.137)			36.879 (0.989)		
<i>P. cannabina</i>	26.859 (3.491)					0.050 (0.112)	0.030 (0.067)	0.673 (0.174)	32.751 (3.527)			17.450 (2.434)		0.010 (0.024)
<i>P. caricapapayae</i>	23.416 (1.325)			0.048 (0.107)	0.272 (0.321)	0.343 (0.037)	0.009 (0.028)	0.946 (0.138)	18.892 (1.779)			39.373 (0.541)		
<i>P. chloridismutans</i>	18.432 (0.833)		0.027 (0.047)		0.302 (0.144)	0.037 (0.076)		0.340 (0.276)	30.153 (2.303)	0.307 (0.132)	0.221 (0.046)	31.546 (0.746)	0.221 (0.046)	0.044 (0.039)
<i>P. dhonopaphis</i>	25.435 (4.298)	0.004 (0.024)			11.155 (6.692)			0.365 (0.236)	9.319 (2.719)	0.554 (0.586)	1.303 (1.136)	19.803 (8.612)	0.006 (0.023)	0.866 (0.787)
<i>P. eichorii</i>	25.207 (2.416)		0.009 (0.035)	0.138 (0.337)	0.817 (2.952)		0.013 (0.072)	1.063 (0.328)	17.338 (4.010)	0.008 (0.046)	0.375 (1.913)	35.072 (3.741)	0.002 (0.012)	0.077 (0.106)
<i>P. cissticola</i>	3.871 (1.482)	0.570 (0.509)				3.516 (2.572)	0.140 (0.263)	0.405 (0.125)	17.530 (32.372)		2.726 (5.157)	16.319 (6.203)	2.846 (1.860)	
<i>P. citrinellolis</i>	21.849 (0.690)				3.890 (1.688)		0.187 (0.089)	0.622 (0.160)	32.751 (3.527)	3.814 (2.080)		17.450 (2.434)		
<i>P. congelans</i>	24.618 (1.582)	0.107 (0.089)			0.578 (0.216)	0.292 (0.025)		0.673 (0.147)	13.285 (1.107)			39.373 (0.541)		
" <i>P. coronafaciens</i> "	24.585 (3.470)				5.530 (1.202)	0.175 (0.187)		0.648 (0.254)	12.994 (2.719)			39.355 (1.736)		
<i>P. corrugata</i>	24.771 (0.025)				4.773 (2.901)		0.010 (0.045)	0.544 (0.158)	15.716 (2.321)	0.415 (0.365)	0.514 (0.562)	28.281 (4.231)		0.108 (0.160)
<i>P. cosmanii</i>	30.472 (2.216)		0.075 (0.059)		6.430 (3.071)	0.072 (0.080)	0.023 (0.057)	0.775 (0.308)	20.852 (2.080)			27.513 (2.874)		0.055 (0.061)
<i>P. eremortolorata</i>	21.152 (1.304)				5.268 (1.705)	0.297 (0.064)		0.483 (0.194)	26.890 (2.803)	0.618 (0.236)		16.508 (2.037)		0.035 (0.086)
<i>P. extremorientalis</i>	30.763 (2.056)				15.662 (3.646)	0.022 (0.053)	0.456 (0.154)	0.690 (0.145)	12.540 (1.928)	0.792 (0.318)		21.534 (0.916)	0.110 (0.054)	0.013 (0.034)
<i>P. flavescens</i>	19.873 (0.825)	0.329 (0.289)						0.739 (0.339)	35.304 (3.389)			21.534 (0.916)		
<i>P. flavescens</i>	22.852 (16.018)		0.063 (0.101)		1.640 (1.505)	1.477 (2.288)		4.098 (6.299)	1.947 (1.543)	0.168 (0.415)	9.497 (7.530)	21.743 (16.908)	0.002 (0.017)	0.034 (0.113)
<i>P. fluorescens</i>	26.660 (3.510)		0.001 (0.009)		4.482 (4.396)	0.026 (0.077)	0.020 (0.064)	0.698 (0.339)	14.264 (3.386)	2.391 (2.065)	0.045 (0.161)	30.698 (5.567)		
<i>P. fragi</i>	26.719 (5.164)				17.756 (10.829)	0.207 (0.200)		0.618 (0.306)	7.317 (4.907)			18.485 (11.155)		
<i>P. frederiksbergensis</i>	23.935 (2.699)				4.507 (4.100)			0.362 (0.137)	7.415 (10.050)	0.037 (0.090)	0.687 (0.199)	25.338 (3.444)	0.038 (0.042)	1.307 (0.311)
<i>P. fulva</i>	32.757 (0.441)				3.303 (1.164)			0.480 (0.128)	19.093 (1.534)	0.167 (0.170)		24.497 (0.471)		
<i>P. fuscovaginatae</i>	26.334 (2.132)				3.009 (3.474)		0.005 (0.034)	0.704 (0.255)	14.483 (5.533)	0.060 (0.125)	0.058 (0.132)	32.453 (5.825)		0.132 (0.225)
<i>P. geniculata</i>	8.567 (0.730)	1.740 (0.232)			0.030 (0.073)	2.162 (0.297)	0.220 (0.049)	0.998 (2.445)	0.673 (0.154)			13.057 (6.628)		
<i>P. genomospecies</i>	24.464 (3.475)	0.017 (0.052)			0.051 (0.121)	0.174 (0.174)	0.026 (0.116)	1.423 (0.929)	16.782 (3.186)			36.888 (2.251)		0.002 (0.015)
<i>P. genomospecies</i>	23.654 (2.188)	0.028 (0.078)			1.823 (1.362)	0.155 (0.212)	0.018 (0.049)	0.836 (0.388)	16.914 (3.249)	0.057 (0.121)		35.324 (1.405)		0.022 (0.042)
<i>P. gessardii</i>	32.024 (2.869)		0.007 (0.022)		8.032 (5.583)		0.064 (0.106)	0.533 (0.182)	15.516 (1.194)			29.920 (5.057)		0.035 (0.054)
<i>P. graminis</i>	26.235 (2.184)	0.030 (0.099)			4.469 (0.201)	0.288 (0.057)		0.512 (0.114)	15.516 (1.194)		0.028 (0.069)	38.977 (0.925)	1.673 (0.280)	0.007 (0.024)
<i>P. hibiscicola</i>	5.810 (1.132)	1.247 (0.172)	0.057 (0.081)		0.420 (0.069)	2.660 (0.392)	0.157 (0.089)	0.093 (0.114)	9.722 (0.482)	0.162 (0.173)	0.056 (0.091)	9.772 (0.861)	1.850 (0.256)	0.160 (0.085)
<i>P. indica</i>	15.981 (1.236)	0.149 (0.099)	0.068 (0.073)		0.305 (0.115)	0.301 (0.079)	0.133 (0.132)	0.260 (0.060)	31.497 (3.603)			21.348 (1.533)	0.363 (0.054)	0.007 (0.024)
<i>P. jessenii</i>	28.857 (2.355)				1.257 (0.611)		0.060 (0.094)	1.098 (1.400)	9.993 (3.554)			34.792 (1.103)		0.160 (0.085)
<i>P. jinhuiensis</i>	19.495 (0.431)	0.038 (0.094)			6.230 (2.805)	0.025 (0.061)	1.283 (0.102)	0.472 (0.149)	25.458 (3.679)	3.058 (1.378)		21.278 (3.494)		0.065 (0.117)

Table A.5 continued.

Species name	C ₁₆₀	C ₁₆₀ iso	C _{16:1} w5c	C _{17:0}	C _{17:0} cyclo	C _{17:0} iso	C _{17:1} w5c	C _{18:0}	C _{18:1} w7c	C _{19:0} cyclo w5c	sum. feat. 2	sum. feat. 3	unknown 11.799	unknown 12.484
<i>P. kilonensis</i>	22.701 (2.456)			0.049 (0.080)	1.670 (3.362)		0.481 (0.157)	16.253 (1.880)				31.607 (1.790)	0.142 (0.063)	0.074 (0.069)
<i>P. laacknassii</i>	18.230 (2.457)			0.137 (0.113)	2.420 (0.710)		0.157 (0.150)	26.987 (4.212)		0.080 (0.139)		25.530 (1.591)	0.113 (0.103)	0.113 (0.103)
<i>P. borensis</i>	29.160 (0.880)		0.018 (0.045)	0.227 (0.120)	8.237 (7.465)		0.717 (0.160)	15.635 (0.762)		0.242 (0.385)		27.687 (7.883)		0.020 (0.049)
<i>P. libanensis</i>	29.103 (2.169)			0.150 (0.135)	14.417 (2.377)	0.075 (0.082)	0.379 (0.199)	8.020 (0.740)		1.915 (0.669)		20.822 (2.731)		0.226 (0.152)
<i>P. lini</i>	23.287 (1.593)			0.151 (0.162)	6.487 (3.868)	0.019 (0.060)	0.427 (0.153)	14.095 (1.802)		0.113 (0.172)	0.184 (0.132)	28.468 (4.293)	0.091 (0.064)	0.009 (0.026)
<i>P. landensis</i>	29.774 (3.570)		0.011 (0.043)	0.163 (0.150)	22.184 (6.528)	0.070 (0.095)	0.483 (0.232)	4.298 (3.766)		2.209 (0.849)		13.496 (6.647)	0.018 (0.048)	0.012 (0.029)
<i>P. larida</i>	23.785 (4.047)			0.126 (0.132)	8.683 (5.305)		0.374 (0.056)	15.036 (0.722)		0.208 (0.205)		24.487 (4.997)		0.032 (0.046)
<i>P. lutea</i>	23.614 (1.241)		0.089 (0.096)		0.954 (0.409)	0.140 (0.079)	0.299 (0.249)	39.344 (5.760)		0.504 (0.569)		37.298 (1.172)	0.038 (0.057)	
<i>P. itacola</i>	15.991 (2.077)				6.785 (1.413)		0.213 (0.109)	6.978 (0.575)				29.027 (1.567)		0.093 (0.050)
<i>P. mandeli</i>	28.988 (2.307)			0.058 (0.065)	6.798 (4.935)	0.006 (0.032)	0.613 (0.243)	11.895 (3.516)		0.179 (0.377)	0.009 (0.042)	28.504 (5.764)		0.021 (0.070)
<i>P. marghidis</i>	28.174 (3.181)			0.159 (0.167)	7.802 (2.316)	0.078 (0.101)	0.101 (0.191)	15.681 (1.313)		0.378 (0.186)		26.386 (2.690)	0.679 (1.383)	0.008 (0.025)
<i>P. mediterranea</i>	28.720 (1.596)		0.027 (0.064)	0.573 (0.324)	2.837 (5.657)	0.068 (0.137)	0.364 (0.132)	29.065 (11.071)		0.574 (1.642)	1.202 (3.511)	30.810 (4.589)		
<i>P. neodocina</i>	18.368 (5.779)			0.057 (0.098)	5.912 (3.102)		0.471 (0.160)	9.634 (1.717)		0.023 (0.075)	0.070 (0.124)	25.520 (3.258)		0.086 (0.087)
<i>P. nigulata</i>	28.209 (2.731)				4.730 (2.927)	0.052 (0.082)	0.782 (0.199)	20.498 (2.925)		0.280 (0.208)		18.362 (8.419)	0.032 (0.049)	
<i>P. montelli</i>	20.173 (1.170)			0.187 (0.092)	3.773 (2.333)	0.175 (0.089)	1.047 (0.179)	32.703 (1.908)		0.367 (0.352)		20.140 (2.928)		
<i>P. mozzelli</i>	18.413 (0.959)			0.157 (0.176)	1.102 (1.556)	0.035 (0.086)	2.365 (0.598)	18.173 (2.385)		1.296 (0.940)		30.625 (2.455)		0.061 (0.102)
<i>P. mucidolens</i>	25.818 (2.863)			0.124 (0.079)	3.306 (1.690)		0.407 (0.092)	30.325 (2.990)		0.129 (0.351)		17.120 (3.436)	0.242 (0.168)	0.014 (0.031)
<i>P. nitroreducens</i>	20.776 (2.119)			0.097 (0.162)	2.602 (3.700)	0.460 (0.333)	0.330 (0.107)	31.821 (10.254)		0.159 (0.364)		19.701 (1.456)	0.020 (0.057)	0.016 (0.031)
<i>P. alveolans</i>	16.972 (8.834)	0.019 (0.036)				0.118 (0.196)	0.801 (0.602)	39.284 (4.384)		0.699 (0.387)		19.626 (3.929)	4.249 (0.580)	0.007 (0.022)
<i>P. oryzahabitans</i>	19.295 (1.617)	0.005 (0.022)					0.700 (0.185)	14.763 (1.811)				18.562 (8.419)	0.113 (0.128)	0.021 (0.040)
<i>P. paleroniana</i>	28.998 (1.941)			0.423 (0.089)	13.182 (4.018)	0.821 (0.172)	0.041 (0.097)	0.564 (0.111)		3.725 (4.189)	0.028 (0.069)	6.896 (0.954)		0.282 (0.346)
<i>P. peli</i>	9.761 (1.707)		0.020 (0.057)	1.762 (0.663)	8.787 (9.824)	2.608 (2.386)	3.064 (0.894)	0.542 (0.321)				22.483 (3.465)		0.042 (0.102)
<i>P. peli</i>	8.547 (2.743)		0.540 (0.631)	3.950 (1.681)	8.077 (0.247)	0.973 (0.136)	4.878 (1.640)	0.126 (0.122)				32.312 (2.734)		0.014 (0.031)
<i>P. peritricingena</i>	2.019 (0.497)	13.071 (3.159)		0.010 (0.028)	8.777 (3.631)		0.378 (0.174)	39.502 (3.267)		0.179 (0.245)		23.082 (2.377)	0.208 (0.106)	
<i>P. plecoglossicida</i>	24.202 (2.384)			0.073 (0.114)	8.777 (3.631)		0.842 (0.200)	42.327 (4.287)				17.422 (1.811)		
<i>P. poue</i>	25.862 (2.470)			0.280 (0.151)	2.348 (2.249)	0.057 (0.088)	0.463 (0.247)	14.027 (3.327)		0.429 (0.631)	0.043 (0.127)	23.805 (7.545)	0.150 (0.078)	0.060 (0.127)
<i>P. proteolytica</i>	26.493 (1.689)		0.008 (0.035)	0.395 (0.106)	5.159 (3.335)	0.049 (0.084)	0.849 (0.695)	0.868 (1.642)		0.038 (0.094)		25.940 (2.871)		
<i>P. pseudocalcigenes</i>	13.287 (2.125)	0.064 (0.112)		0.533 (0.551)	1.177 (1.201)	1.050 (0.526)	0.717 (0.343)	21.677 (2.148)				37.755 (0.951)		
<i>P. psychotolerans</i>	21.463 (1.075)			0.121 (0.163)	9.510 (5.768)	0.007 (0.040)	0.012 (0.047)	0.868 (1.642)				27.073 (1.851)		0.015 (0.037)
<i>P. patida</i>	25.017 (3.720)	0.002 (0.022)	0.002 (0.015)	0.557 (0.032)	2.322 (1.993)	0.118 (0.099)	1.225 (0.105)	0.717 (0.343)				26.980 (2.621)		0.022 (0.048)
<i>P. resinovarus</i>	16.900 (1.648)		0.092 (0.104)	0.662 (0.096)	9.510 (5.768)	0.223 (0.182)	0.050 (0.080)	1.282 (1.435)				27.525 (3.704)	0.165 (0.071)	0.095 (0.054)
<i>P. rhodesiae</i>	20.257 (1.457)			0.095 (0.164)	5.748 (1.549)	0.032 (0.078)	0.126 (0.122)	15.345 (1.909)		0.827 (0.269)		18.362 (8.419)	0.170 (0.093)	0.017 (0.037)
<i>P. salomonii</i>	29.026 (1.593)			0.342 (0.148)	7.605 (2.614)	0.042 (0.089)	0.058 (0.100)	15.722 (1.588)		0.369 (0.173)		26.980 (2.621)		0.022 (0.048)
<i>P. straminea</i>	17.504 (1.692)	0.080 (0.095)		0.349 (0.118)	7.605 (2.614)	0.179 (0.034)	0.573 (0.237)	36.194 (2.958)		0.856 (0.876)	0.016 (0.077)	22.740 (1.384)	0.165 (0.071)	0.095 (0.054)
<i>P. stutzeri</i>	18.317 (2.466)			0.004 (0.024)	1.039 (1.831)	0.528 (0.196)	0.007 (0.045)	0.425 (0.263)				27.525 (3.704)	0.170 (0.093)	
<i>P. syntantha</i>	26.239 (0.900)			0.254 (0.178)	9.116 (3.965)		0.769 (0.085)	16.960 (1.796)		1.301 (0.657)		21.341 (4.026)	0.001 (0.009)	0.005 (0.023)
<i>P. syringae</i>	24.260 (2.912)	0.047 (0.135)		0.116 (0.283)	0.661 (1.179)	0.216 (0.227)	0.028 (0.109)	1.001 (0.700)		0.054 (0.260)		36.948 (3.233)		
<i>P. taetrolens</i>	29.632 (0.880)			0.172 (0.190)	8.513 (4.027)	0.115 (0.129)	0.028 (0.109)	7.158 (0.466)		0.522 (0.425)		30.360 (4.417)		0.013 (0.033)
<i>P. thermotolerans</i>	34.453 (1.035)	0.248 (0.035)	0.017 (0.041)	2.765 (0.159)	6.483 (0.307)	0.042 (0.089)	0.018 (0.045)	0.997 (0.158)		19.020 (1.856)		27.005 (2.561)	0.025 (0.039)	0.013 (0.033)
<i>P. thierivalensis</i>	29.944 (2.352)			0.030 (0.064)	9.015 (1.939)	0.362 (0.034)	0.667 (0.096)	7.177 (1.995)		0.430 (0.105)	0.248 (0.082)	23.140 (4.278)	0.165 (0.071)	0.284 (0.117)
<i>P. tolaasii</i>	32.415 (4.348)			0.434 (0.260)	12.489 (4.423)	0.018 (0.059)	0.387 (0.087)	9.520 (1.097)		0.719 (0.479)	0.007 (0.028)	32.684 (1.659)	0.024 (0.044)	0.008 (0.048)
<i>P. troneae</i>	26.964 (2.736)			0.365 (0.315)	0.950 (1.349)		1.247 (0.851)	12.997 (3.790)			5.469 (7.665)	33.120 (0.943)		0.009 (0.025)
<i>P. trivialis</i>	26.035 (1.489)		0.013 (0.033)	0.135 (0.109)	2.297 (0.948)		0.434 (0.171)	14.445 (2.511)				33.120 (0.943)		
<i>P. vancouverensis</i>	26.715 (1.453)			0.083 (0.129)	7.548 (4.510)		0.643 (0.134)	14.820 (1.103)		0.058 (0.093)	1.012 (0.260)	25.503 (5.680)	0.760 (0.246)	0.760 (0.246)
<i>P. vancouverensis</i>	28.413 (2.444)			0.175 (0.248)	13.793 (0.354)		0.510 (0.231)	10.498 (1.142)		0.532 (0.476)	0.127 (0.143)	22.095 (10.462)	0.150 (0.135)	0.150 (0.135)
<i>P. veronii</i>	25.449 (4.977)			0.377 (2.807)	6.377 (2.807)		0.804 (0.118)	13.150 (0.927)		0.613 (0.275)		26.289 (2.433)		
<i>P. viridiflava</i>	22.835 (2.178)			0.175 (0.248)	2.014 (2.155)	0.385 (0.309)	1.154 (0.318)	18.991 (2.333)		0.038 (0.106)		34.357 (2.641)		

A.3 PCA Biplots

The following figures show the biplots resulting from principal component analysis of the three genus data sets. Only biplots are given for the first two principal components.

Biplot of the genus <i>Bacillus</i>	p. 228
Biplot of the genus <i>Paenibacillus</i>	p. 229
Biplot of the genus <i>Pseudomonas</i>	p. 230

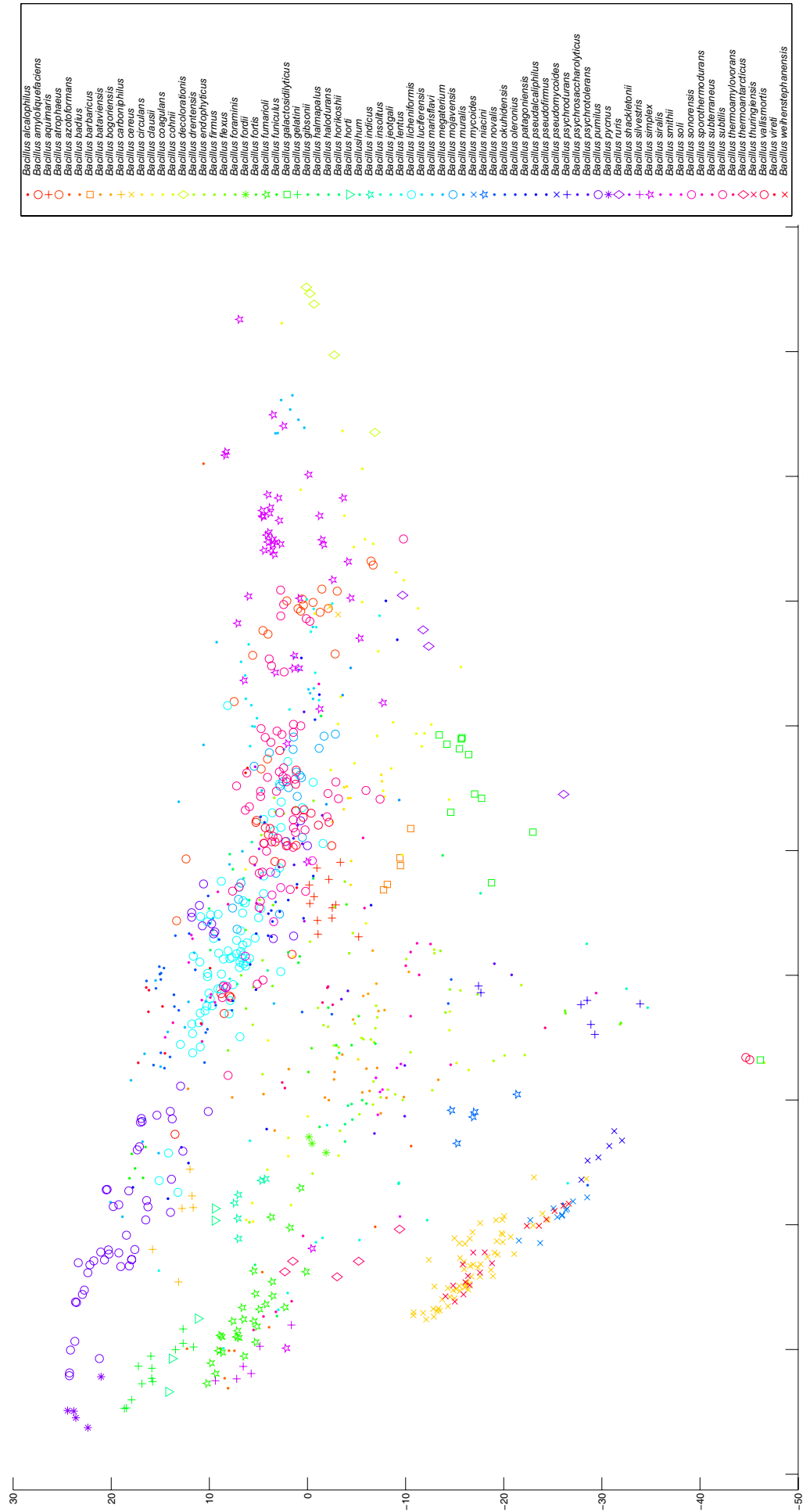


Figure A.1: Biplot of the first two principal components of PCA analysis of the *Bacillus* data set. Species denoted by the character 'x' belong to the *Bacillus cereus* species group, while species denoted by a circle belong to the *Bacillus subtilis* group. Besides those marks, single species that form a separable cluster are also marked, though by a different character.

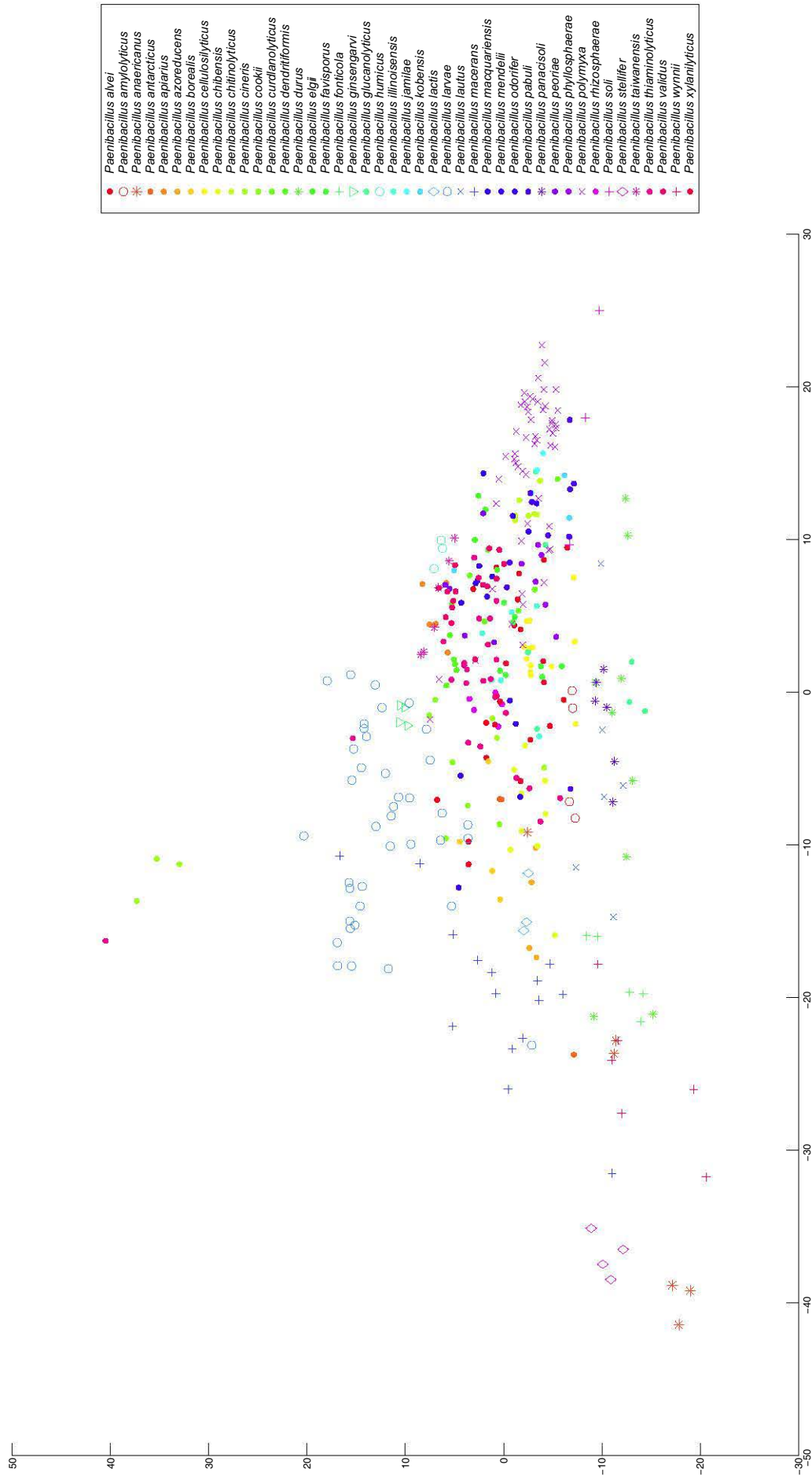


Figure A.2: Biplot of the first two principal components of PCA analysis of the *Paenibacillus* data set. Single species that cluster in a distinct group are marked by a different character for each species individually.

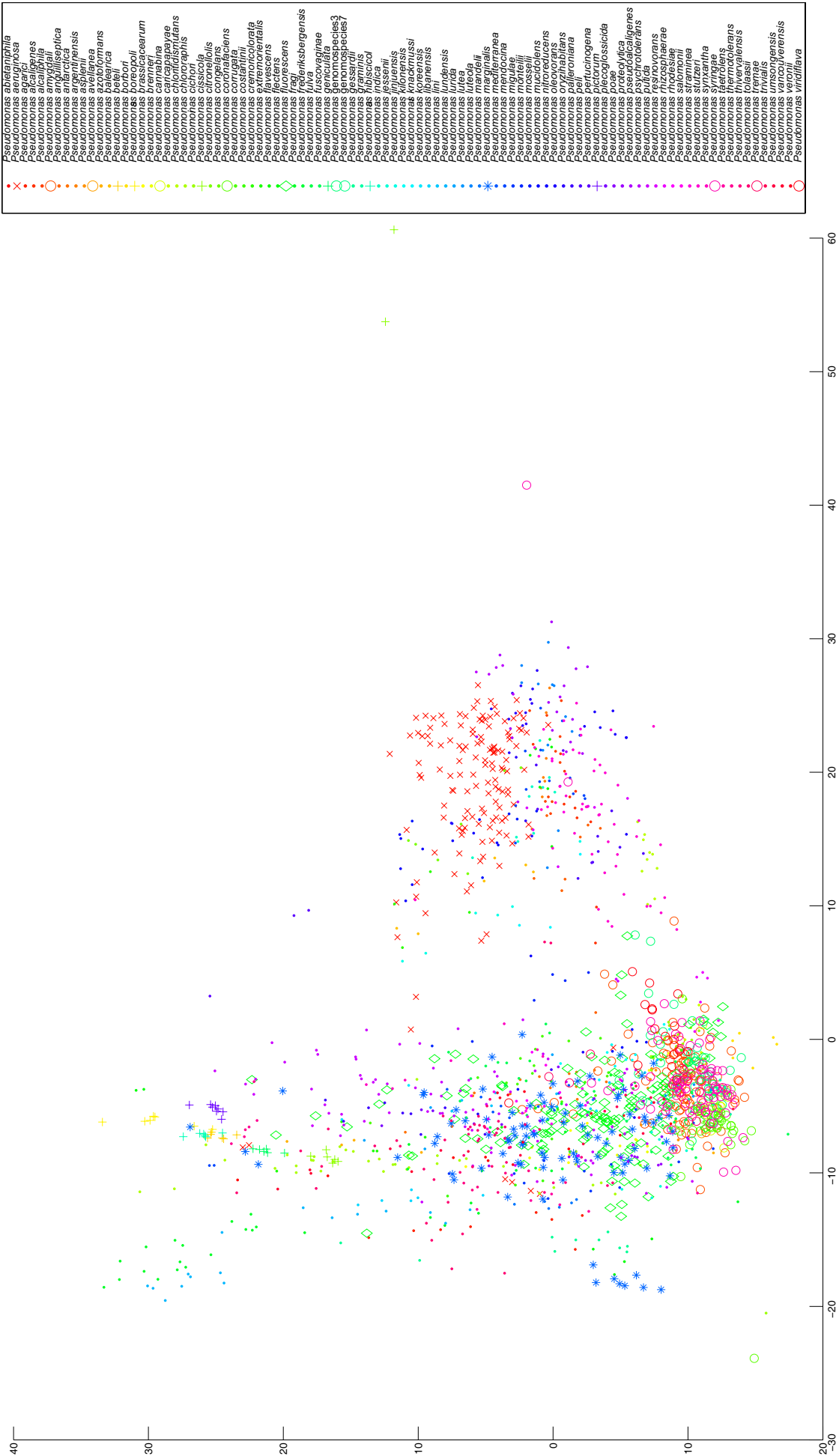


Figure A.3: Biplot of the first two principal components of *Pseudomonas* data set. Species denoted by the character ‘×’ belong to the *Pseudomonas aeruginosa* species group, while species denoted by a circle belong to the *Pseudomonas syringae* group and species denoted by a plus character belong to the *Pseudomonas beteli* outgroup. Besides those marks, single species that form a separable cluster are also marked, though by a different character.

APPENDIX B

The Sherlock MIS

B.1 TSBA50 Peak Naming Table

Peak Name	Nominal ECL	Summed Feature
9:0	9.000	
8:0 3OH	9.392	
unknown 9.531	9.531	
10:0 ISO	9.604	
10:0	10.000	
9:0 3OH	10.408	
11:0 ISO	10.606	
11:0 ANTEISO	10.695	
12:0 ALDE ?	10.914	2
unknown 10.928	10.928	2
11:0	11.000	
10:0 2OH	11.153	
10:0 3OH	11.422	
unknown 11.543	11.543	
12:0 ISO	11.609	
12:0 ANTEISO	11.699	
unknown 11.799	11.799	
12:1 AT 11-12	11.925	
12:0	12.000	
11:0 ISO 3OH	12.089	
11:0 2OH	12.16	
11:0 3OH	12.438	
unknown 12.484	12.484	
13:0 ISO	12.614	
13:0 ANTEISO	12.702	
13:1 AT 12-13	12.936	
13:0	13.000	
12:0 ISO 3OH	13.098	
12:0 2OH	13.177	
12:1 3OH	13.288	
14:1 ISO E	13.384	
12:0 3OH	13.454	
unknown 13.565	13.565	
14:0 ISO	13.619	
14:0 ANTEISO	13.707	
14:1 w5c	13.901	

Table B.1: The FAME peaks named by the Sherlock MIS TSBA50 peak naming table.

Table B.1 continued.

Peak Name	Nominal ECL	Summed Feature
unknown 13.957	13.957	
14:0	14.000	
13:0 ISO 3OH	14.109	
13:0 2OH	14.194	
unknown 14.263	14.263	
ISO 15:1 AT 5	14.389	
15:1 ISO F	14.415	
15:1 ISO G	14.44	
15:1 ISO H/13:0 3OH	14.46	1
13:0 3OH/15:1 i I/H	14.47	1
15:1 ISO I/13:0 3OH	14.478	1
unknown 14.502	14.502	
15:1 ANTEISO A	14.527	
15:0 ISO	14.623	
15:0 ANTEISO	14.713	
15:1 w8c	14.793	
15:1 w6c	14.856	
15:1 w5c	14.903	
unknown 14.959	14.959	
15:0	15.000	
14:0 ISO 3OH	15.119	
14:0 2OH	15.203	
16:1 w7c alcohol	15.387	
16:1 ISO G	15.442	
16:1 ISO H	15.461	
16:1 ISO I/14:0 3OH	15.48	2
14:0 3OH/16:1 ISO I	15.488	2
16:0 N alcohol	15.55	
16:0 ISO	15.627	
unknown 15.669	15.669	
16:0 ANTEISO	15.718	
16:1 w11c	15.757	
16:1 w9c	15.774	
16:1 w7c/15 iso 2OH	15.822	3
15:0 ISO 2OH/16:1w7c	15.852	3
16:1 w5c	15.909	
16:0	16.000	
15:0 ISO 3OH	16.134	
15:0 2OH	16.219	
ISO 17:1 w10c	16.388	
ISO 17:1 w9c	16.416	
16:0 10 methyl	16.432	
ISO 17:1 w5c	16.461	
17:1 ISO I/ANTEI B	16.476	4
17:1 ANTEISO B/i I	16.486	4
15:0 3OH	16.503	
ANTEISO 17:1 w9c	16.524	
17:1 ANTEISO A	16.54	
unknown 16.582	16.582	
17:0 ISO	16.63	

Table B.1 continued.

Peak Name	Nominal ECL	Summed Feature
17:0 ANTEISO	16.723	
17:1 w9c	16.772	
17:1 w8c	16.792	
17:1 w7c	16.818	
17:1 w6c	16.86	
17:0 CYCLO	16.888	
17:1 w5c	16.917	
17:0	17.000	
16:1 2OH	17.048	
16:0 ISO 3OH	17.15	
16:0 2OH	17.233	
17:0 10 methyl	17.409	
18:1 ISO H	17.464	
16:0 3OH	17.519	
18:3 w6c (6,9,12)	17.577	
18:0 ISO	17.632	
18:2 w6,9c/18:0 ANTE	17.72	5
18:0 ANTE/18:2 w6,9c	17.727	5
18:1 w9c	17.769	
18:1 w7c	17.823	
18:1 w6c	17.858	
18:1 w5c	17.919	
18:0	18.000	
11 methyl 18:1 w7c	18.081	
17:0 ISO 3OH	18.161	
17:0 2OH	18.254	
TBSA 10Me18:0	18.392	
19:1 ISO I	18.473	
17:0 3OH	18.536	
19:0 ISO	18.634	
19:0 ANTEISO	18.731	
19:1 w11c/19:1 w9c	18.756	6
19:1 w9c/19:1 w11c	18.768	6
unknown 18.814	18.814	
un 18.846/19:1 w6c	18.846	7
19:1 w6c/.846/19cy	18.858	7
19:0 CYCLO w10c/19w6	18.867	7
19:0 CYCLO w8c	18.902	
19:0	19.000	
18:1 2OH	19.089	
18:0 2OH	19.264	
19:0 10 methyl	19.368	
20:4 w6,9,12,15c	19.395	
18:0 3OH	19.55	
20:0 ISO	19.635	
20:2 w6,9c	19.732	
20:1 w9c	19.77	
20:1 w7c	19.831	
20:0	20.000	

B.2 TSBA50 Identification Library

The following table covers the entries of the TSBA50 identification library of Sherlock MIS corresponding to the genera *Bacillus*, *Paenibacillus* and *Pseudomonas*.

Entry	Valid species?
<i>Bacillus</i> GC group 22 (No 16S match to known species)	
<i>Bacillus agaradhaerens</i>	✓
<i>Bacillus alcalophilus</i> (some 48h)	✓
<i>Bacillus amyloliquefaciens</i> (<i>Bacillus subtilis</i>)	✓
<i>Bacillus atrophaeus</i>	✓
<i>Bacillus azotoformans</i>	✓
<i>Bacillus badius</i>	✓
<i>Bacillus cereus</i> GC subgroup A	✓
<i>Bacillus cereus</i> GC subgroup B	✓
<i>Bacillus circulans</i>	✓
<i>Bacillus cirroflagellosus</i> (48h)	✓
<i>Bacillus clausii</i>	✓
<i>Bacillus coagulans</i>	✓
<i>Bacillus cohnii</i>	✓
<i>Bacillus ehimensis</i>	<i>Paenibacillus ehimensis</i>
<i>Bacillus epiphytus</i>	
<i>Bacillus fastidiosus</i>	✓
<i>Bacillus filicolonicus</i>	
<i>Bacillus firmus</i>	✓
<i>Bacillus flexus</i>	✓
<i>Bacillus freudenreichii</i>	
<i>Bacillus fusiformis</i>	<i>Lysinibacillus fusiformis</i>
<i>Bacillus gibsonii</i>	✓
<i>Bacillus globisporus</i>	<i>Sporosarcina globispora</i>
<i>Bacillus halmapalus</i> (48h)	✓
<i>Bacillus halodenitrificans</i> (48h)	<i>Virgibacillus halodenitrificans</i>
<i>Bacillus horikoshii</i> (48h)	✓
<i>Bacillus insolitus</i>	✓
<i>Bacillus laevolacticus</i>	<i>Sporolactobacillus laevolacticus</i>
<i>Bacillus lentus</i>	✓
<i>Bacillus licheniformis</i> (<i>Bacillus subtilis</i> group)	✓
<i>Bacillus macroides</i>	✓
<i>Bacillus megaterium</i> GC subgroup A	✓
<i>Bacillus megaterium</i> GC subgroup B	✓
<i>Bacillus mycoides</i> GC subgroup A (<i>Bacillus cereus</i> group)	✓
<i>Bacillus mycoides</i> GC subgroup B (<i>Bacillus cereus</i> group)	✓
<i>Bacillus niacini</i>	✓
<i>Bacillus oleronius</i>	✓
<i>Bacillus psychrosaccharolyticus</i>	✓
<i>Bacillus pumilus</i> GC subgroup A	✓
<i>Bacillus pumilus</i> GC subgroup B	✓
<i>Bacillus racemilactis</i>	
<i>Bacillus simplex</i>	✓
<i>Bacillus smithii</i> (55°C)	✓

Table B.2: Entries of the Sherlock MIS TSBA50 identification library corresponding to the genus *Bacillus*, *Paenibacillus* and *Pseudomonas*. Entries validly described as a species according to the taxonomy of 03/2008 are marked, otherwise the correct species name is given (only if species was validly described). Subspecies or infrasubspecific ranks are not considered.

Table B.2 continued.

Entry	Valid species?
<i>Bacillus sphaericus</i>	<i>Lysinibacillus sphaericus</i>
<i>Bacillus sphaericus</i> GC subgroup A	<i>Lysinibacillus sphaericus</i>
<i>Bacillus sphaericus</i> GC subgroup B	<i>Lysinibacillus sphaericus</i>
<i>Bacillus sphaericus</i> GC subgroup C	<i>Lysinibacillus sphaericus</i>
<i>Bacillus sphaericus</i> GC subgroup D	<i>Lysinibacillus sphaericus</i>
<i>Bacillus sphaericus</i> GC subgroup E	<i>Lysinibacillus sphaericus</i>
<i>Bacillus sphaericus</i> GC subgroup F	<i>Lysinibacillus sphaericus</i>
<i>Bacillus subtilis</i>	✓
<i>Bacillus thuringiensis aizawai</i>	✓
<i>Bacillus thuringiensis canadensis</i>	✓
<i>Bacillus thuringiensis dendrolimus</i>	✓
<i>Bacillus thuringiensis entomocidus</i>	✓
<i>Bacillus thuringiensis gallieriae</i>	✓
<i>Bacillus thuringiensis israelensis</i>	✓
<i>Bacillus thuringiensis kurstakii</i>	✓
<i>Bacillus thuringiensis sotto</i>	✓
<i>Paenibacillus alginolyticus</i>	
<i>Paenibacillus alvei</i> GC subgroup A (<i>Bacillus</i>)	✓
<i>Paenibacillus alvei</i> GC subgroup B (<i>Bacillus</i>)	✓
<i>Paenibacillus amylolyticus</i>	✓
<i>Paenibacillus apiarius</i> (<i>Bacillus apiarius</i>)	✓
<i>Paenibacillus azotofixans</i> (<i>Bacillus azotofixans</i>)	<i>Paenibacillus durus</i>
<i>Paenibacillus chondroitinus</i>	✓
<i>Paenibacillus glucanolyticus</i> (<i>Bacillus</i>)	✓
<i>Paenibacillus larvae larvae</i> (72h, <i>Bacillus</i>)	✓
<i>Paenibacillus larvae larvae pulvifaciens</i> (48h, <i>Bacillus</i>)	✓
<i>Paenibacillus lautus</i>	✓
<i>Paenibacillus lentimorbus</i>	✓
<i>Paenibacillus macerans</i> (<i>Bacillus</i>)	✓
<i>Paenibacillus macquariensis</i>	✓
<i>Paenibacillus pabuli</i> (<i>Bacillus</i>)	✓
<i>Paenibacillus peoriae</i> (<i>Bacillus</i>)	✓
<i>Paenibacillus polymyxa</i> (<i>Bacillus</i>)	✓
<i>Paenibacillus popilliae</i> (<i>Bacillus</i>)	✓
<i>Paenibacillus thiaminolyticus</i> (<i>Bacillus</i>)	✓
<i>Paenibacillus validus</i> (<i>Bacillus gordonae</i>)	✓
<i>Pseudomonas aeruginosa</i>	✓
<i>Pseudomonas agarici</i>	✓
<i>Pseudomonas alcaligenes</i>	✓
<i>Pseudomonas amyloclavata</i> (not a valid name)	
<i>Pseudomonas balearica</i>	✓
<i>Pseudomonas chlororaphis/aureofaciens/aurantiaca</i>	✓
<i>Pseudomonas cichorii/viridiflava</i>	✓
<i>Pseudomonas corrugata</i>	✓
<i>Pseudomonas flectens</i>	✓
<i>Pseudomonas fluorescens biotype A</i>	✓
<i>Pseudomonas fluorescens biotype B</i>	✓
<i>Pseudomonas fluorescens biotype C/P. mandelii</i>	✓
<i>Pseudomonas fluorescens biotype F</i>	✓
<i>Pseudomonas fluorescens biotype G/taetrolens</i>	✓
<i>Pseudomonas fuscovaginae</i>	✓
<i>Pseudomonas huttiensis</i>	<i>Herbaspirillum huttiense</i>
<i>Pseudomonas lundensis</i>	✓
<i>Pseudomonas mendocina/straminea</i>	✓
<i>Pseudomonas mucidolens</i>	✓
<i>Pseudomonas oleovorans</i>	✓
<i>Pseudomonas pertucinogena</i>	✓

Table B.2 continued.

Entry	Valid species?
<i>Pseudomonas pseudoalcaligenes</i>	✓
<i>Pseudomonas putida</i> biotype A	✓
<i>Pseudomonas putida</i> biotype B	✓
<i>Pseudomonas resinovorans</i>	✓
<i>Pseudomonas savastanoi</i> fraxinus	✓
<i>Pseudomonas savastanoi</i> nerium	✓
<i>Pseudomonas savastanoi</i> oleae	✓
<i>Pseudomonas stutzeri</i> (<i>P. perfectomarina</i>)	✓
<i>Pseudomonas synxantha</i>	✓
<i>Pseudomonas syringae</i> atrofaciens	✓
<i>Pseudomonas syringae</i> coronafaciens (<i>P. coronafaciens</i>)	✓
<i>Pseudomonas syringae</i> glycinea	✓
<i>Pseudomonas syringae</i> lachrymans/pisi	✓
<i>Pseudomonas syringae</i> maculicola	✓
<i>Pseudomonas syringae</i> mori	✓
<i>Pseudomonas syringae</i> morsprunorum	✓
<i>Pseudomonas syringae</i> phaseolicola	✓
<i>Pseudomonas syringae</i> syringae	✓
<i>Pseudomonas syringae</i> tabaci	✓
<i>Pseudomonas syringae</i> tagetes	✓
<i>Pseudomonas syringae</i> tomato	✓
<i>Pseudomonas vancouverensis</i>	✓
<i>Pseudomonas veronii</i>	✓

APPENDIX C

16S rRNA Gene Sequences

Species name	Strain Number	Accession Number
<i>Bacillus alcalophilus</i>	DSM 485 ^T	X76436
<i>Bacillus amyloliquefaciens</i>	NBRC 15535 ^T	AB255669
<i>Bacillus aquimaris</i>	TF 12 ^T	AF483625
<i>Bacillus atrophaeus</i>	JCM9070 ^T	AB021181
<i>Bacillus azotoformans</i>	NBRC 15712 ^T	AB363732
<i>Bacillus badius</i>	NBRC 15713 ^T	AB271748
<i>Bacillus barbaricus</i>	DSM 14730 ^T	AJ422145
<i>Bacillus bataviensis</i>	LMG 21832 ^T	AJ542507
<i>Bacillus bogoriensis</i>	LMG 22234 ^T	AY376312
<i>Bacillus carboniphilus</i>	JCM 9731 ^T	AB021182
<i>Bacillus cereus</i>	CCM 2010 ^T	DQ207729
<i>Bacillus circulans</i>	ATCC 4513 ^T	AY043084
<i>Bacillus clausii</i>	DSM 8716 ^T	X76440
<i>Bacillus coagulans</i>	ATCC 7050 ^T	DQ297928
<i>Bacillus cohnii</i>	DSM 6307 ^T	X76437
<i>Bacillus decolorationis</i>	LMG 19507 ^T	AJ315075
<i>Bacillus drentensis</i>	LMG 21831 ^T	AJ542506
<i>Bacillus endophyticus</i>	CIP 106778 ^T	AF295302
<i>Bacillus firmus</i>	IAM 12464 ^T	D16268
<i>Bacillus flexus</i>	IFO15715 ^T	AB021185
<i>Bacillus foraminis</i>	LMG 23174 ^T	AJ717382
<i>Bacillus fordii</i>	LMG 22080 ^T	AY443039
<i>Bacillus fortis</i>	LMG 22079 ^T	AY443038
<i>Bacillus fumarioli</i>	LMG 17489 ^T	AJ250056
<i>Bacillus funiculus</i>	CIP 107128 ^T	AB049195
<i>Bacillus galactosidilyticus</i>	LMG 17892 ^T	AJ535638
<i>Bacillus gelatini</i>	LMG 21880 ^T	AJ551329
<i>Bacillus gibsonii</i>	DSM 8722 ^T	X76446
<i>Bacillus halmapalus</i>	DSM 8723 ^T	X76447

Table C.1: Strain list with according 16S rRNA gene accession numbers. List of the species included in the 2008 data set with the type strain number and the accession number according to the selected 16S rRNA gene sequence.

Table C.1 continued.

Species name	Strain Number	Accession Number
<i>Bacillus halodurans</i>	DSM 497 ^T	AJ302709
<i>Bacillus horikoshii</i>	DSM 8719 ^T	AB043865
<i>Bacillus horti</i>	JCM 9943 ^T	D87035
<i>Bacillus humi</i>	LMG 22167 ^T	AJ627210
<i>Bacillus indicus</i>	DSM 15820 ^T	AJ583158
<i>Bacillus insolitus</i>	DSM 5 ^T	AM980508
<i>Bacillus jeotgali</i>	JCM 10885 ^T	AF221061
<i>Bacillus lentus</i>	NCIMB 8773 ^T	AB021189
<i>Bacillus licheniformis</i>	DSM 13 ^T	X68416
<i>Bacillus luciferensis</i>	LMG 18422 ^T	AJ419629
<i>Bacillus marisflavi</i>	JCM 11544 ^T	AF483624
<i>Bacillus megaterium</i>	IAM 13418 ^T	D16273
<i>Bacillus mojavensis</i>	IFO 15718 ^T	AB021191
<i>Bacillus muralis</i>	LMG 20238 ^T	AJ628748
<i>Bacillus mycoides</i>	ATCC 6462 ^T	AB021192
<i>Bacillus niacini</i>	IFO 15566 ^T	AB021194
<i>Bacillus novalis</i>	LMG 21837 ^T	AJ542512
<i>Bacillus okuhidensis</i>	JCM 10945 ^T	AB047684
<i>Bacillus oleronius</i>	ATCC 700005 ^T	AY988598
<i>Bacillus patagoniensis</i>	DSM 16117 ^T	AY258614
<i>Bacillus pseudocaliphilus</i>	DSM 8725 ^T	X76449
<i>Bacillus pseudofirmus</i>	DSM 8715 ^T	X76439
<i>Bacillus pseudomycooides</i>	DSM 12442 ^T	AM747226
<i>Bacillus psychrodurans</i>	DSM 11713 ^T	AJ277984
<i>Bacillus psychrosaccharolyticus</i>	ATCC 23296 ^T	AB021195
<i>Bacillus psychrotolerans</i>	DSM 11706 ^T	AJ277983
<i>Bacillus pumilus</i>	DSM 27 ^T	AY456263
<i>Bacillus pycnus</i>	NBRC 101231 ^T	AB271739
<i>Bacillus ruris</i>	LMG 22866 ^T	AJ535639
<i>Bacillus shackletonii</i>	LMG 18435 ^T	AJ250318
<i>Bacillus silvestris</i>	DSM 12223 ^T	AJ006086
<i>Bacillus simplex</i>	DSM 1321 ^T	AJ439078
<i>Bacillus siralis</i>	CIP 106295 ^T	AF071856
<i>Bacillus smithii</i>	DSM 4216 ^T	Z26935
<i>Bacillus soli</i>	LMG 21838 ^T	AJ542513
<i>Bacillus sonorensis</i>	BCRC 17416 ^T	EF433411
<i>Bacillus sporothermodurans</i>	DSMZ 10599 ^T	U49078
<i>Bacillus subterraneus</i>	DSM 13966 ^T	AY672638
<i>Bacillus subtilis</i> subsp. <i>subtilis</i>	DSM 10 ^T	AJ276351
<i>Bacillus thermantarcticus</i>	DSM 9572 ^T	AJ493665
<i>Bacillus thermoamylovorans</i>	LMG 18084 ^T	L27478
<i>Bacillus thuringiensis</i>	ATCC 10792 ^T	AF290545
<i>Bacillus vallismortis</i>	DSM 11031 ^T	AB021198
<i>Bacillus vireti</i>	LMG 21834 ^T	AJ542509
<i>Bacillus weihenstephanensis</i>	DSM 11821 ^T	AB021199
<i>Paenibacillus alvei</i>	DSM 29 ^T	AJ320491
<i>Paenibacillus amylolyticus</i>	NRRL NRS-290 ^T	D85396
<i>Paenibacillus anaericanus</i>	DSM 15890 ^T	AJ318909
<i>Paenibacillus antarcticus</i>	LMG 22078 ^T	AJ605292

Table C.1 continued.

Species name	Strain Number	Accession Number
<i>Paenibacillus apiarius</i>	DSM 5581 ^T	AB073201
<i>Paenibacillus azoreducens</i>	DSM 13822 ^T	AJ272249
<i>Paenibacillus borealis</i>	CCUG 43137 ^T	AJ011322
<i>Paenibacillus cellulosityticus</i>	LMG 22232 ^T	DQ407282
<i>Paenibacillus chibensis</i>	JCM 9905 ^T	AB073194
<i>Paenibacillus chitinolyticus</i>	IFO 15660 ^T	AB021183
<i>Paenibacillus cineris</i>	LMG 18439 ^T	AJ575658
<i>Paenibacillus cookii</i>	LMG 18419 ^T	AJ250317
<i>Paenibacillus curdolanolyticus</i>	DSM 10247 ^T	AB073202
<i>Paenibacillus dendritiformis</i>	CIP 105967 ^T	AY359885
<i>Paenibacillus durus</i>	ATCC 35681 ^T	AJ251192
<i>Paenibacillus elgii</i>	KCTC 10016BP ^T	AY090110
<i>Paenibacillus favisporus</i>	LMG 20987 ^T	AY208751
<i>Paenibacillus fonticola</i>	LMG 23577 ^T	DQ453131
<i>Paenibacillus ginsengarvi</i>	DSM 18677 ^T	AB271057
<i>Paenibacillus glucanolyticus</i>	DSM 5162 ^T	AB073189
<i>Paenibacillus humicus</i>	LMG 23886 ^T	AM411528
<i>Paenibacillus illinoisensis</i>	JCM 9907 ^T	AB073192
<i>Paenibacillus jamilae</i>	CEC ^T 5266 ^T	AJ271157
<i>Paenibacillus kobensis</i>	DSM 10249 ^T	AB073363
<i>Paenibacillus lactis</i>	LMG 21940 ^T	AY257868
<i>Paenibacillus larvae</i> subsp. <i>larvae</i>	DSM 7030 ^T	AY530294
<i>Paenibacillus lautus</i>	JCM 9073 ^T	AB073188
<i>Paenibacillus macerans</i>	IAM 12467 ^T	AB073196
<i>Paenibacillus macquariensis</i>	DSM 2 ^T	AB073193
<i>Paenibacillus mendelii</i>	LMG 23002 ^T	AF537343
<i>Paenibacillus odorifer</i>	LMG 19079 ^T	AJ223990
<i>Paenibacillus pabuli</i>	NRRL NRS-924 ^T	AB045094
<i>Paenibacillus panacisoli</i>	LMG 23405 ^T	AB245384
<i>Paenibacillus peoriae</i>	DSM 8320 ^T	AJ320494
<i>Paenibacillus phyllosphaerae</i>	LMG 22192 ^T	AY598818
<i>Paenibacillus polymyxa</i>	DSM 36 ^T	AJ320493
<i>Paenibacillus rhizosphaerae</i>	LMG 21955 ^T	AY751754
<i>Paenibacillus soli</i>	LMG 23604 ^T	DQ309072
<i>Paenibacillus stellifer</i>	DSM 14472 ^T	AJ316013
<i>Paenibacillus taiwanensis</i>	BCRC 17411 ^T	DQ890521
<i>Paenibacillus thiaminolyticus</i>	IFO 15656 ^T	AB073197
<i>Paenibacillus validus</i>	JCM 9077 ^T	AB073203
<i>Paenibacillus wynnii</i>	LMG 22176 ^T	AJ633647
<i>Paenibacillus xylanilyticus</i>	LMG 21957 ^T	AY427832
<i>Pseudomonas abietaniphila</i>	ATCC 700689 ^T	AJ011504
<i>Pseudomonas aeruginosa</i>	ATCC 10145 ^T	AF094713
<i>Pseudomonas agarici</i>	LMG 2112 ^T	Z76652
<i>Pseudomonas alcaligenes</i>	LMG 1224 ^T	Z76653
<i>Pseudomonas alcaliphila</i>	JCM 10630 ^T	AB030583
<i>Pseudomonas amygdali</i>	LMG 2123 ^T	Z76654
<i>Pseudomonas anguilliseptica</i>	NCIMB 1949 ^T	X99540
<i>Pseudomonas antarctica</i>	DSM 1531 ^T	AJ537601
<i>Pseudomonas argentinensis</i>	LMG 22563 ^T	AY691188

Table C.1 continued.

Species name	Strain Number	Accession Number
<i>Pseudomonas asplenii</i>	ATCC 23835 ^T	AB021397
<i>Pseudomonas avellanea</i>	BPIC 714	AJ889838
<i>Pseudomonas azotoformans</i>	IAM 1603 ^T	D84009
<i>Pseudomonas balearica</i>	DSM 6083 ^T	U26418
<i>Pseudomonas beteli</i>	IAM 12423 ^T	AB294553
<i>Pseudomonas borbori</i>	LMG 23199 ^T	AM114527
<i>Pseudomonas boreopolis</i>	ATCC 33662 ^T	AB021391
<i>Pseudomonas brassicacearum</i>	CFBP 11706 ^T	AF100321
<i>Pseudomonas brenneri</i>	CIP 106646 ^T	AF268968
<i>Pseudomonas cannabina</i>	CFBP 2341 ^T	AJ492827
<i>Pseudomonas caricapapayae</i>	ATCC 33615 ^T	D84010
<i>Pseudomonas chloritidismutans</i>	AW 1 ^T	AY017341
<i>Pseudomonas chlororaphis</i>	DSM 6698 ^T	AY509898
<i>Pseudomonas cichorii</i>	LMG 2162 ^T	Z76658
<i>Pseudomonas cissicola</i>	ATCC 33616 ^T	AB021399
<i>Pseudomonas citronellolis</i>	DSM 50332 ^T	Z76659
<i>Pseudomonas congelans</i>	LMG 21466 ^T	AJ492828
“ <i>Pseudomonas coronafaciens</i> ”	LMG 13190 ^T	Z76660
<i>Pseudomonas corrugata</i>	ATCC 29736 ^T	D84012
<i>Pseudomonas costantinii</i>	CFBP 5705 ^T	AF374472
<i>Pseudomonas cremoricolorata</i>	IAM 1541 ^T	AB060137
<i>Pseudomonas extremorientalis</i>	LMG 19695 ^T	AF405328
<i>Pseudomonas flavescens</i>	NCPPB 3063 ^T	U01916
<i>Pseudomonas flectens</i>	ATCC 12775 ^T	AB021400
<i>Pseudomonas fluorescens</i>	IAM 12022 ^T	D84013
<i>Pseudomonas fragi</i>	ATCC 4973 ^T	AF094733
<i>Pseudomonas frederiksbergensis</i>	DSM 13022 ^T	AJ249382
<i>Pseudomonas fulva</i>	NRIC 0180 ^T	AB060136
<i>Pseudomonas fuscovaginae</i>	MAFF 301177 ^T	AB021381
<i>Pseudomonas geniculata</i>	ATCC 19374 ^T	AB021404
<i>Pseudomonas genomospecies3</i>	DC3000	AE016853
<i>Pseudomonas genomospecies7</i>	MAFF 302271	AB001449
<i>Pseudomonas gessardii</i>	CIP 105469 ^T	AF074384
<i>Pseudomonas graminis</i>	CIP 105469 ^T	Y11150
<i>Pseudomonas hibiscicola</i>	ATCC 19867 ^T	AB021405
<i>Pseudomonas indica</i>	MTCC 3713 ^T	AF302795
<i>Pseudomonas jessenii</i>	CIP 105274 ^T	AF068259
<i>Pseudomonas jinjuensis</i>	LMG 21316 ^T	AF468448
<i>Pseudomonas kilonensis</i>	DSM 13647 ^T	AJ292426
<i>Pseudomonas knackmussii</i>	LMG 23759 ^T	AF039489
<i>Pseudomonas koreensis</i>	LMG 21318 ^T	AF468452
<i>Pseudomonas libanensis</i>	CIP 105460 ^T	AF057645
<i>Pseudomonas lini</i>	CFBP 5737 ^T	AY035996
<i>Pseudomonas lundensis</i>	ATCC 49968 ^T	AB021395
<i>Pseudomonas lurida</i>	DSM 15835 ^T	AJ581999
<i>Pseudomonas lutea</i>	LMG 21974 ^T	AY364537
<i>Pseudomonas luteola</i>	IAM 13000 ^T	D84002
<i>Pseudomonas mandelii</i>	CIP 105273 ^T	AF058286
<i>Pseudomonas marginalis</i>	LMG 2210 ^T	Z76663

Table C.1 continued.

Species name	Strain Number	Accession Number
<i>Pseudomonas mediterranea</i>	CFBP 5447 ^T	AF386080
<i>Pseudomonas mendocina</i>	NCIB 10541 ^T	D84016
<i>Pseudomonas migulae</i>	CIP 105470 ^T	AF074383
<i>Pseudomonas monteilii</i>	CIP 104883 ^T	AF064458
<i>Pseudomonas mosselii</i>	CIP 105259 ^T	AF072688
<i>Pseudomonas mucidolens</i>	IAM 12406 ^T	D84017
<i>Pseudomonas nitroreducens</i>	IAM 1439 ^T	AM088473
<i>Pseudomonas oleovorans</i>	IAM 1508 ^T	D84018
<i>Pseudomonas oryzihabitans</i>	IAM 1568 ^T	D84004
<i>Pseudomonas palleroniana</i>	CFBP 4389 ^T	AY091527
<i>Pseudomonas peli</i>	LMG 23201 ^T	AM114534
<i>Pseudomonas pertucinogena</i>	IFO 14163 ^T	EF673695
<i>Pseudomonas pictorum</i>	ATCC 23328 ^T	AB021392
<i>Pseudomonas plecoglossicida</i>	ATCC 700383 ^T	AB009457
<i>Pseudomonas poae</i>	LMG 21465 ^T	AJ492829
<i>Pseudomonas proteolytica</i>	DSM 15321 ^T	AJ537603
<i>Pseudomonas pseudoalcaligenes</i>	LMG 1225 ^T	Z76666
<i>Pseudomonas psychrotolerans</i>	LMG 21977 ^T	AJ575816
<i>Pseudomonas putida</i>	IAM 1236 ^T	D84020
<i>Pseudomonas resinovorans</i>	LMG 2274 ^T	Z76668
<i>Pseudomonas rhizosphaerae</i>	LMG 21640 ^T	AY152673
<i>Pseudomonas rhodesiae</i>	CIP 104664 ^T	AF064459
<i>Pseudomonas salomonii</i>	CFBP 2022 ^T	AY091528
<i>Pseudomonas straminea</i>	IAM 1598 ^T	D84023
<i>Pseudomonas stutzeri</i>	ATCC 14405 ^T	U26420
<i>Pseudomonas synxantha</i>	IAM 12356 ^T	D84025
<i>Pseudomonas syringae</i>	ATCC 19310 ^T	AF094749
<i>Pseudomonas taetrolens</i>	IAM 1653 ^T	D84027
<i>Pseudomonas thermotolerans</i>	LMG 21284 ^T	AJ311980
<i>Pseudomonas thivervalensis</i>	CFBP 11261 ^T	AF100323
<i>Pseudomonas tolaasii</i>	LMG 2342 ^T	AF255336
<i>Pseudomonas tremae</i>	CFBP 6111 ^T	AJ492826
<i>Pseudomonas trivialis</i>	LMG 21464 ^T	AJ492831
<i>Pseudomonas umsongensis</i>	LMG 21317 ^T	AF468450
<i>Pseudomonas vancouverensis</i>	ATCC 700688 ^T	AJ011507
<i>Pseudomonas veronii</i>	CIP 104663 ^T	AF064460
<i>Pseudomonas viridiflava</i>	LMG 2352 ^T	Z76671

Bibliography

- Abel, K., Deschmertz, H., and Peterson, J. I.** (1963). Classification of microorganisms by analysis of chemical composition i. feasibility of utilizing gas chromatography. Journal of Bacteriology, 85(5):1039–1044.
- Aizerman, M., Braverman, E., and Rozonoer, L.** (1964). Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 25:821–837.
- Akashi, S. and Saito, K.** (1960). A branched saturated C15 acid (sarcinic acid) from *Sarcina* phospholipids and a similar acid from several microbial lipids. The Journal of Biochemistry, 47(2):222–229.
- Allwein, E., Schapire, R., and Singer, Y.** (2000). Reducing multi-class to binary: a unifying approach for margin classifiers. Journal of Machine Learning Research, 1:113–141.
- Amann, R. I., Ludwig, W., and Schleifer, K. H.** (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews, 59(1):143–169.
- Amaratunga, D., Cabrera, J., and Lee, Y.-S.** (2008). Enriched random forests. Bioinformatics, 24(18):2010–2014.
- Anzai, Y., Kim, H., Park, J.-Y., Wakabayashi, H., and Oyaizu, H.** (2000). Phylogenetic affiliation of the Pseudomonads based on 16S rRNA sequence. International Journal of Systematic and Evolutionary Microbiology, 50:1563–1589.
- Anzai, Y., Kudo, Y., and Oyaizu, H.** (1997). The phylogeny of the genera *Chryseomonas*, *Flavimonas*, and *Pseudomonas* supports synonymy of these three genera. International Journal of Systematic Bacteriology, 47(249-251).
- Ash, C., Farrow, J. A. E., Wallbanks, S., and Collins, M. D.** (1991). Phylogenetic heterogeneity of the genus *Bacillus* revealed by comparative analysis of small-subunit-ribosomal RNA sequences. Letters in Applied Microbiology, 13:202–206.
- Ash, C., Priest, F. G., and Collins, M. D.** (1993). Molecular identification of ribosomal-rna group 3 bacilli (Ash, Farrow, Wallbanks and Collins) using a PCR probe test - Proposal for the creation of a new genus *Paenibacillus*. Antonie van Leeuwenhoek International Journal of General and Molecular Microbiology, 64(3-4):253–260.
- Baida, N., Yazourh, A., Singer, E., and Izard, D.** (2001). *Pseudomonas brenneri* sp. nov., a new species isolated from natural mineral waters. Research in Microbiology, 152:493–502.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H.** (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, 16(5):412–424.
- Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G.** (2006). Hierarchical multi-label prediction of gene function. Bioinformatics, 22(7):830–836.
- Bavykin, S. G., Lysov, Y. P., Zakhariyev, V., Kelly, J. J., Jackman, J., Stahl, D. A., and Cherni, A.** (2004). Use of 16S rRNA, 23s rRNA, and *gyrB* gene sequence analysis to determine phylogenetic relationships of *Bacillus cereus* group organisms. Journal of Clinical Microbiology, 42(8):3711–3730.
- Behrendt, U., Ulrich, A., and Schumann, P.** (2003). Fluorescent pseudomonads associated with the phyllosphere of grasses; *Pseudomonas trivialis* sp. nov., *Pseudomonas poae* sp. nov. and *Pseudomonas congelans* sp. nov. International Journal of Systematic and Evolutionary Microbiology, 53:1461–1469.
- Behrendt, U., Ulrich, A., Schumann, P., Meyer, J.-M., and Spröer, C.** (2007). *Pseudomonas lurida* sp. nov., a fluorescent species associated with the phyllosphere of grasses. International Journal of Systematic and Evolutionary Microbiology, 57:979–985.

- Berkeley, R.** (2002). Whither *Bacillus*? In Berkeley, R., Heyndrickx, M., Logan, N., and De Vos, P., editors, Applications and Systematics of *Bacillus* and Relatives, pages 1–7. Blackwell Science Ltd., Malden, USA.
- Bertone, S., Giacomini, M., Ruggiero, C., Piccarolo, C., and Calegari, L.** (1996). Automated systems for identification of heterotrophic marine bacteria on the basis of their fatty acid composition. Applied and Environmental Microbiology, 62(6):2122–2132.
- Biau, G. e. r., Devroye, L., and Lugosi, G. a. b.** (2008). Consistency of Random Forests and other averaging classifiers. Journal of Machine Learning Research, 9:2015–2033.
- Bishop, C. M.** (1995). Neural Networks for Pattern Recognition. Oxford University Press, New York.
- Bishop, C. M.** (2006). Pattern Recognition and Machine Learning. Springer, New York, 1st edition.
- Blockeel, H., Bruynooghe, M., Dzeroski, S., Ramon, J., and Struyf, J.** (2002). Hierarchical multi-classification. In De Raedt, L., Dzeroski, S., and Wrobel, S., editors, KDD-2002 Workshop on Multi-Relational Data Mining, pages 21–35, Edmonton, Alberta, Canada.
- Bosch, A., Miñán, A., Vescina, C., Degrossi, J., Gatti, B., Montanaro, P., Messina, M., Franco, M., Vay, C., Schmitt, J., Naumann, D., and Yantorno, O.** (2008). Fourier transform infrared spectroscopy for rapid identification of nonfermenting Gram-negative bacteria isolated from sputum samples from cystic fibrosis patients. Journal of Clinical Microbiology, 46(8):2535–2546.
- Boser, B., Guyon, I., and Vapnik, V.** (1992). A training algorithm for optimal margin classifiers. In Haussler, D., editor, 5th Annual ACM Workshop on Computational Learning Theory (COLT), pages 144–152, Pittsburgh, PA, USA. ACM Press.
- Bradley, A. P.** (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7):1145–1159.
- Breiman, L.** (1996a). Bagging predictors. Machine Learning, 24:123–140.
- Breiman, L.** (1996b). Out-of-bag estimation. Technical report, University of California.
- Breiman, L.** (2001). Random forests. Machine Learning, 45(1):5–32.
- Breiman, L.** (2002). Wald lecture ii. looking inside the black box.
- Breiman, L.** (2004). Random Forests. URL: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.** (1984). Classification and regression trees. Chapman and Hall, New York, NY, USA.
- Breiman, L. and Spector, P.** (1992). Submodel selection and evaluation in regression: the X-random case. International Statistical Review, 60:291–319.
- Brenner, D. J., Krieg, N. R., and Staley, J. T.** (2005a). Bergey's Manual of Systematic Bacteriology Volume 2: The Proteobacteria Part A Introductory Assays, volume 2. Springer, New York, NY, USA.
- Brenner, D. J., Staley, J. T., and Krieg, N. R.** (2005b). Classification of procaryotic organisms and the concept of bacterial speciation. In Brenner, D. J., Krieg, N. R., and Staley, J. T., editors, Bergey's Manual of Systematic Bacteriology Volume 2: The Proteobacteria Part A Introductory Assays, volume 2, pages 27–38. Springer, New York, NY, USA.
- Bright, J. J., Claydon, M. A., Soufian, M., and Gordon, D. B.** (2002). Rapid typing of bacteria using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry and pattern recognition software. Journal of Microbiological Methods, 48:127–138.
- Burges, C. J. C.** (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):121–167.
- Butler, W. R., Thibert, L., and Kilburn, J. O.** (1992). Identification of *Mycobacterium avium* complex strains and some similar species by high-performance liquid chromatography. Journal of Clinical Microbiology, 30(10):2698–2704.
- Buyer, J. S.** (2002a). Identification of bacteria from single colonies by fatty acid analysis. Journal of Microbiological Methods, 48:259–265.
- Buyer, J. S.** (2002b). Rapid sampling processing and fast gas chromatography for identification of bacteria by fatty acid analysis. Journal of Microbiological Methods, 51:209–215.
- Buyer, J. S.** (2003). Improved fast gas chromatography for FAME analysis of bacteria. Journal of Microbiological

- Methods, 54:117–120.
- Buyer, J. S.** (2006). Rapid and sensitive FAME analysis of bacteria by cold trap injection gas chromatography. Journal of Microbiological Methods, 67:187–190.
- Bylander, T.** (2002). Estimating generalization error in two-class datasets using out-of-bag estimates. Machine Learning, 48:287–297.
- 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.
- CCUG** (2009). CCUG: Culture Collection, University of Göteborg, Sweden. URL: <http://www.ccug.se>.
- Cesa-Bianchi, N., Gentile, C., and Zaniboni, L.** (2006). Incremental algorithms for hierarchical classification. Journal of Machine Learning Research, 7:31–54.
- Chang, C. and Lin, C.** (2001). LIBSVM: a library for support vector machines. URL: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Chen, C., Liaw, A., and Breiman, L.** (2004). Using Random Forest to learn imbalanced data. Technical report, Department of Statistics, University of California.
- Cheong, S., Oh, S. H., and Lee, S.** (2004). Support vector machines with binary tree architecture for multi-class classification. Neural Information Processing - Letters and Reviews, 2(3):47 – 51.
- Chernick, M. R.** (2008). Bootstrap methods: a guide for practitioners and researchers. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, New Jersey, USA, 2nd edition.
- Chun, J., Atalan, E., Ward, A. C., and Goodfellow, M.** (1993). Artificial neural network analysis of pyrolysis mass spectrometric data in the identification of *Streptomyces* strains. FEMS Microbiology Letters, 107(2-3):321–325.
- Chun, J. and Sook Bae, K.** (2000). Phylogenetic analysis of *Bacillus subtilis* and related taxa based on partial *gyrA* gene sequences. Antonie van Leeuwenhoek, 78:123–127.
- Chung, A. P., Nunes, O. C., Tindall, B. J., and Milton, C. S.** (1993). The effect of the growth medium composition on the fatty acids of *Rhodothermus marinus* and ‘*Thermus thermophilus*’ HB-8. FEMS Microbiology Letters, 112(1):13–18.
- Cohan, F. M.** (2002). What are bacterial species? Annual Review of Microbiology, 56:457–487.
- Combet-Blanc, Y., Ollivier, B., Streicher, C., Patel, B. K. C., Dwivedi, P. P., Pot, B., Prensier, G., and Garcia, J.-L.** (1995). *Bacillus thermoamylovorans* sp. nov., a moderately thermophilic and amyolytic bacterium. International Journal of Systematic Bacteriology, 45(1):9–16.
- Coorevits, A., De Jonghe, V., Vandroemme, J., Reekmans, R., Heyrman, J., Messens, W., De Vos, P., and Heyndrickx, M.** (2008). Comparative analysis of the diversity of aerobic spore-forming bacteria in raw milk from organic and conventional dairy farms. Systematic and Applied Microbiology, 31:126–140.
- Coroler, L., Elomari, M., Hoste, B., Gillis, M., Izard, D., and Leclercq, J. P.** (1996). *Pseudomonas rhodesiae* sp. nov., a new species isolated from natural mineral waters. Systematic and Applied Microbiology, 19:600–607.
- Cortes, C. and Vapnik, V.** (1995). Support-vector networks. Machine Learning, 20:273–297.
- Crammer, K. and Singer, Y.** (2001). On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, 2:265–292.
- Crammer, K. and Singer, Y.** (2002). On the learnability and design of output codes for multiclass problems. Machine Learning, 47:201–233.
- Cristianini, N. and Shawe-Taylor, J.** (2000). An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press, Cambridge.
- Dawyndt, P.** (2004). Knowledge Accumulation of Microbial Data Aiming at a Dynamic Taxonomic Framework. PhD thesis, Ghent University, Belgium.
- Dawyndt, P.** (2009). StrainInfo.net. URL: <http://www.straininfo.net>.
- Dawyndt, P., Demeyer, H., and De Baets, B.** (2006). UPGMA clustering revisited: A weight-driven approach to transitive approximation. International Journal of Approximate Reasoning, 42(3):174–191.

- Dees, S. B. and Moss, C. W. (1975). Cellular fatty acids of *Alcaligenes* and *Pseudomonas* species isolated from clinical specimens. *Journal of Clinical Microbiology*, 1(5):414–419.
- Dees, S. B., Moss, C. W., Weaver, R. E., and Hollis, D. (1979). Cellular fatty acid composition of *Pseudomonas paucimobilis* and groups Ilk-2, Ve-1, and Ve-2. *Journal of Clinical Microbiology*, 10(2):206–209.
- Dekel, O., Keshet, J., and Singer, Y. (2004). Large margin hierarchical classification. In *21st International Conference on Machine Learning*, pages 1–8, Banff, Canada.
- Demuth, H., Beale, M., and Hagan, M. (2006). *Neural network toolbox user's guide*. Technical report, The MathWorks.
- Dietterich, T. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15.
- Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.
- Dietterich, T. G. and Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286.
- Dietterich, T. G. and Kong, E. B. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Department of Computer Science, Oregon State University.
- Donohue, J. P. and Welsh, W. J. (2004). Speciation of *Listeria* via pyrolysis/gas chromatography. *Journal of Analytical and Applied Pyrolysis*, 72(2):221–228.
- Doolittle, W. F. and Papke, R. T. (2006). Genomics and the bacterial species problem. *Genome Biology*, 7(9):116.1–116.7.
- Drobniewski, F. (1993). *Bacillus cereus* and related species. *Clinical Microbiology Reviews*, 6(4):324–338.
- Drucker, D. B. and Veazey, F. J. (1977). Fatty acid fingerprints of *Streptococcus mutans* NCTC 10832 grown at various temperatures. *Applied and Environmental Microbiology*, 33(2):221–226.
- DSMZ (2009). Bacterial Nomenclature Up-to-Date. URL: http://www.dsmz.de/microorganisms/bacterial_nomenclature.php.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley and Sons, New York.
- Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments *Statistica Sinica*, 12:111–139.
- Dumais, S. and Chen, H. (2000). Hierarchical classification of web content. In Belkin, N. J., Ingwersen, P., and Leong, M.-K., editors, *SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 256–263, Athens, Greece. ACM Press, New York, US.
- Dutta, R., Gardner, J. W., and Hines, E. L. (2004). Classification of ear, nose, and throat bacteria using a neural-network-based electronic nose. *MRS Bulletin*, 29(10):709–713.
- Dutta, R., Hines, E. L., Gardner, J. L., and Boilot, P. (2002). Bacteria classification using Cyranose 320 electronic nose. *Biomedical Engineering Online*, 1(4):1–7.
- Dziuba, B., Babuchowski, A., Dziuba, M., and Nałcz, D. (2007). Identification of lactic acid bacteria using FTIR spectroscopy and artificial neural networks. *Milchwissenschaft*, 62(1):28–32.
- Eerola, E. and Lehtonen, O. (1988). Optimal data processing procedure for automatic bacterial identification by gas-liquid chromatography of cellular fatty acids. *Journal of Clinical Microbiology*, 26(9):1745–1753.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton, Florida, USA.
- Elomari, M., Coroler, L., Hoste, B., Gillis, M., Izard, D., and Leclercq, J. P. (1996). DNA relatedness among *Pseudomonas* strains isolated from natural mineral waters and proposal of *Pseudomonas veronii* sp. nov. *International Journal of Systematic Bacteriology*, 46(4):1138–1144.
- Euzéby, J. P. (1997). List of Bacterial Names with Standing in Nomenclature: a folder available on the internet. *International Journal of Systematic Bacteriology*, 47:590–592.
- Fausett, L. (1994). *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*. Prentice-Hall, Englewood Cliffs, New Jersey.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874.
- Fei, B. and Liu, J. (2006). Binary tree of SVM: a new fast multiclass training and classification algorithm. *IEEE Transactions on Neural Networks*, 17(3):696–704.

- Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (version 3.2). *Cladistics*, 5:164–166.
- Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA, USA, 1st edition.
- Fensom, A. H. and Gray, G. W. (1969). The chemical composition of the lipopolysaccharide of *Pseudomonas aeruginosa*. *Biochemical Journal*, 114(2):185–196.
- Fichefet, J., Leclercq, J. P., Beyne, P., and Rousselet-Piette (1984). Microcomputer-assisted identification of bacteria and multicriteria decision models. *Computers and Operations Research*, 11(4):361–372.
- Fielding, A. H. and Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1):38–49.
- Fieldsend, J. E. and Everson, R. M. (2005). Formulation and comparison of multi-class ROC surfaces. In *Proceedings of the ICML 2005 workshop on ROC analysis in Machine Learning*, Bonn, Germany.
- Flach, P. (2004). The many faces of ROC analysis in machine learning.
- Freeman, R., Goodacre, R., Sisson, P. R., Magee, J. G., Ward, A. C., and Lightfoot, N. F. (1994). Rapid identification of species within the *Mycobacterium tuberculosis* complex by artificial neural network analysis of pyrolysis mass spectra. *Journal of Medical Microbiology*, 40(3):170–173.
- Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, pages 148–156, Bari, Italy.
- Fürnkranz, J. (2002). Round robin classification. *Journal of Machine Learning Research*, 2:723–747.
- 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.
- Garrity, G. M., Lilburn, T. G., Cole, J. R., Harrison, S. H., Euzéby, J., and Tindal, B. J. (2007). Taxonomic outline of the Bacteria and Archaea, Release 7.7. *Michigan State University, Board of Trustees*. URL: <http://www.taxonomicoutline.org>.
- Gatson, J. W., Benz, B. F., Chandrasekaran, C., Satomi, M., Venkateswaran, K., and Hart, M. E. (2006). *Bacillus tequilensis* sp. nov., isolated from a 2000-year-old Mexican shaft-tomb, is closely related to *Bacillus subtilis*. *International Journal of Systematic and Evolutionary Microbiology*, 56:1475–1484.
- Gaus, K., Rösch, P., Petry, R., Peschke, K.-D., Ronneberger, O., Burkhardt, H., Baumann, K., and Popp, J. (2006). Classification of lactic acid bacteria with UV-resonance Raman spectroscopy. *Biopolymers*, 82(4):286–290.
- Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F. L., and Swings, J. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, 3:733–739.
- Giacomini, M., Bertone, S., Soumetz, F. C., and Ruggiero, C. (2004). An advanced approach based on artificial neural networks to identify environmental bacteria. *International Journal of Computational Intelligence*, 1(2):96–103.
- Giacomini, M., Ruggiero, C., Calegari, F., and Bertone, S. (2000). Artificial neural network based identification of environmental bacteria by gas-chromatographic and electrophoretic data. *Journal of Microbiological Methods*, 43:45–54.
- Glucksman, A. M., Skipper, H. D., Brigmon, R. L., and Domingo, J. W. S. (2000). Use of the MIDI-FAME technique to characterize groundwater communities. *Journal of Applied Microbiology*, 88:711–719.
- Goodacre, R., Hiom, S. J., Cheeseman, S. L., Murdoch, D., Weightman, A. J., and Wade, W. G. (1996a). Identification and discrimination of oral asaccharolytic *Eubacterium* spp. by pyrolysis mass spectrometry and artificial neural networks. *Current Microbiology*, 32(2):77–84.
- Goodacre, R., Rooney, P. J., and Kell, D. B. (1998a). Rapid analysis of microbial systems using vibrational spectroscopy and supervised learning methods: Application to the discrimination between methicillin-resistant

- and methicillin-susceptible *Staphylococcus aureus*. In *SPIE - The International Society for Optical Engineering*, volume 3257, pages 220–229.
- Goodacre, R., Timmins, E. M., Burton, R., Kaderbhai, N., Woodward, A. M., Kell, D. B., and Rooney, P. J.** (1998b). Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. *Microbiology*, 144:1157–1170.
- Goodacre, R., Timmins, E. M., Rooney, P. J., Rowland, J. J., and Kell, D. B.** (1996b). Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks. *FEMS Microbiology Letters*, 140(2-3):233–239.
- Goris, J., Konstantinidis, K. T., Klappenback, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M.** (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57:81–91.
- Guermeur, Y.** (2007). VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8:2551–2594.
- Gupta, R. S.** (2005). Molecular sequences and the early history of life. In Sapp, J., editor, *Microbial Phylogeny and Evolution*, pages 160–183. Oxford University Press, New York, NY, USA.
- Haack, S. K., Garchow, H., Odelson, D. A., Forney, L. J., and Klug, M. J.** (1994). Accuracy, reproducibility, and interpretation of fatty acid methyl ester profiles of model bacterial communities. *Applied and Environmental Microbiology*, 60(7):2483–2493.
- Hancock, I. C., Humphreys, G. O., and Meadow, P. M.** (1970). Characterisation of the hydroxy acids of *Pseudomonas aeruginosa* 8602. *Biochimica et Biophysica Acta (BBA)*, 202:389–391.
- Hand, D. J. and Till, R. J.** (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45:171–186.
- Hanley, J. A. and McNeil, B. J.** (1983). A method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases. *Radiology*, 148(3):839–843.
- Hansen, B. M., Leser, T. D., and Hendriksen, N. B.** (2001). Polymerase chain reaction assay for the detection of *Bacillus cereus* group cells. *FEMS Microbiology Letters*, 202(2):209–213.
- Harrington, P. D. B., Voorhees, K. J., Basile, F., and Hendricker, A. D.** (2001). Validation using sensitivity and target transform factor analysis of neural network models for classifying bacteria from mass spectra. *Journal of the American Society for Mass Spectrometry*, 13:10–21.
- Hastie, T. and Tibshirani, R.** (1998). Classification by pairwise coupling. *The Annals of Statistics*, 26:451–471.
- Hastie, T., Tibshirani, R., and Friedman, J.** (2001). *The Elements of Statistical Learning*. Springer, New York, NY, USA.
- Hastie, T., Tibshirani, R., and Friedman, J.** (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer-Verlag, New York, USA, 2nd edition.
- Hettick, J. M., Kashon, M. L., Slaven, J. E., Ma, Y., Simpson, J. P., Siegel, P. D., Mazurek, G. N., and Weissman, D. N.** (2006). Discrimination of intact mycobacteria at the strain level: A combined MALDI-TOF MS and biostatistical analysis. *Proteomics*, 6(24):6416–6425.
- Heyndrickx, M., Vandemeulebroecke, K., Scheldeman, P., Kersters, K., De Vos, P., Logan, N. A., Aziz, A. M., Ali, N., and Berkeley, R. C. W.** (1996). A polyphasic reassessment of the genus *Paenibacillus*, reclassification of *Bacillus lautus* (Nakamura 1984) as *Paenibacillus lautus* comb. nov. and of *Bacillus peoriae* (Montefusco et al. 1993) as *Paenibacillus peoriae* comb. nov., and emended descriptions of *P. lautus* and of *P. peoriae*. *International Journal of Systematic Bacteriology*, 46(4):988–1003.
- Heyrman, J., Logan, N. A., Rodriguez-Diaz, M., Scheldeman, P., Lebbe, L., Swings, J., Heyndrickx, M., and De Vos, P.** (2005). Study of mural painting isolates, leading to the transfer of ‘*Bacillus maroccanus*’ and ‘*Bacillus carotarum*’ to *Bacillus simplex*, emended description of *Bacillus simplex*, re-examination of the strains previously attributed to ‘*Bacillus macroides*’ and description of *Bacillus muralis* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 55(1):119–131.
- Heyrman, J., Mergaert, J., Denys, R., and Swings, J.** (1999). The use of fatty methyl ester analysis (FAME) for the identification of heterotrophic bacteria present on three mural paintings showing severe damage by

- microorganisms. *FEMS Microbiology Letters*, 181:55–62.
- Heyrman, J., Vanparrys, B., Logan, N. A., Balcaen, A., Rodríguez-Díaz, M., Felske, A., and De Vos, P.** (2004). *Bacillus novalis* sp. nov., *Bacillus vireti* sp. nov., *Bacillus soli* sp. nov., *Bacillus bataviensis* sp. nov. and *Bacillus drentensis* sp. nov., from the Drentse A grasslands. *International Journal of Systematic and Evolutionary Microbiology*, 54(1):47–57.
- Higging, s. J. J.** (2004). *An Introduction to Modern Nonparametric Statistics*. Brooks/Cole - Thomson Learning, CA, USA.
- Hildebrand, D. C., Palleroni, N. J., Hendson, M., Toth, J., and Johnson, J. L.** (1994). *Pseudomonas flavescens* sp. nov., isolated from walnut blight cancers. *International Journal of Systematic Bacteriology*, 44(3):410–415.
- Hofmann, T., Cai, L., and Ciaramita, M.** (2003). Learning with taxonomies: classifying documents and words. In *Workshop on syntax, semantics and statistics of the 17th annual conference on neural information processing systems*, Whistler, British Columbia, Canada.
- Höfte, M., and De Vos, P.** (2007). Plant pathogenic *Pseudomonas* species. In Gnanamanickam S. S., editor, *Plant-associated bacteria*. Springer, Dordrecht, The Netherlands.
- Hsu, C. and Lin, C.** (2002). A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2).
- 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(1):228–230.
- Hutsebaut, D., Vandroemme, J., Heyrman, J., Dawyndt, P., Vandenabeele, P., Moens, L., and De Vos, P.** (2006). Raman microspectrometry as an identification tool within the phylogenetically homogeneous ‘*Bacillus subtilis*’-group. *Systematic and Applied Microbiology*, 29:650–660.
- Huys, G., Kämpfer, P., Vancanneyt, M., Coopman, R., Janssen, P., and Kersters, K.** (1997). Effect of the growth medium on the cellular fatty acid composition of aeromonads: consequences for the chemotaxonomic differentiation of DNA hybridization groups in the genus *Aeromonas*. *Journal of Microbiological Methods*, 28:89–97.
- IJSEM** (2009). *International Journal of Systematic and Evolutionary Microbiology*. URL: <http://ijs.sgmjournals.org/>.
- Ikemoto, S., Kuraishi, H., Komagata, K., Azuma, R., Suto, T., and Murooka, H.** (1978). Cellular fatty acid composition in *Pseudomonas* species. *Journal of General and Applied Microbiology*, 24:199–213.
- INSDC** (2009). International nucleotide sequence database collaboration. URL: <http://www.insdc.org/>.
- Ivanova, E. P., Gorshkova, N. M., Sawabe, T., Hayashi, K., Kalinovskaya, N. I., Lysenko, A. M., Zhukova, N. V., Nicolau, D. V., Kuznetsova, T. A., Mikhailov, V. V., and Christen, R.** (2002). *Pseudomonas extremorientalis* sp. nov., isolated from a drinking water reservoir. *International Journal of Systematic and Evolutionary Microbiology*, 52(2113–2120).
- Iversen, C., Lancashire, L., Waddington, M., Forsythe, S., and Ball, G.** (2006). Identification of *Enterobacter sakazakii* from closely related species: the use of artificial neural networks in the analysis of biochemical and 16s rDNA data. *BMC Microbiology*, 6(28):1–8.
- Janse, J. D., Rossi, P., Angelucci, L., Scortichini, M., Derks, J. H. J., Akermans, A. D. L., De Vrijer, R., and Psallidas, P. G.** (1996). Reclassification of *Pseudomonas syringae* pv. *avellanae* as *Pseudomonas avellanae* spec. nov., the bacterium causing cancer of hazelnut (*Corylus avellanae* L.). *Systematic and Applied Microbiology*, 19:589–595.
- James, A. T. and Martin, A. J. P.** (1952). Gas-liquid partition chromatography: the separation and micro-estimation of volatile fatty acids from formic acid to dodecanoic acid. *Biochemical Journal*, 50(5):679–690.
- Japkowicz, N. and Stephen, S.** (2002). The class imbalance problem: a systematic study. *Intelligent Data Analysis*, 6:429–449.
- Juneja, V. K. and Davidson, P. M.** (1993). Influence of temperature on the fatty acid profile of *Listeria monocytogenes*. *Journal of Rapid Methods & Automation in Microbiology*, 2(1):73–81.
- Kämpfer, P.** (1994). Limits and possibilities of total fatty acid analysis for classification and identification of *Bacillus* species. *Systematic and Applied Microbiology*, 17:86–98.

- Kämpfer, P.** (2002). Whole-cell fatty acid analysis in the systematics of *Bacillus* and related genera. In Berkeley, R., Heyndrickx, M., Logan, N., and De Vos, P., editors, Applications and Systematics of Bacillus and Relatives, pages 271–299. Wiley-Blackwell, Oxford, UK.
- Kämpfer, P. and Kroppenstedt, R. M.** (1996). Numerical analysis of fatty acid patterns of coryneform bacteria and related taxa. Canadian Journal of Microbiology, 42:989–1005.
- Kaneda, T.** (1963a). Biosynthesis of branched chain fatty acids. I. isolation and identification of fatty acids from *Bacillus subtilis* (ATCC 7059). Journal of Biological Chemistry, 238(4):1222–1228.
- Kaneda, T.** (1963b). Biosynthesis of branched-chain fatty acids II. microbial synthesis of branched long chain fatty acids from certain short chain fatty acid substrates. The Journal of Biological Chemistry, 238(4):1229–1235.
- Kaneda, T.** (1966a). Biosynthesis of branched-chain fatty acids. IV. factors affecting relative abundance of fatty acids by *Bacillus subtilis*. Canadian Journal of Microbiology, 12:501–514.
- Kaneda, T.** (1966b). Biosynthesis of branched-chain fatty acids V. microbial stereospecific syntheses of D-12-methyltetradecanoic and D-14-methylhexadecanoic acids. Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism, 125(1):43–54.
- Kaneda, T.** (1967). Fatty acids in the genus *Bacillus* I. iso- and anteiso-fatty acids as characteristic constituents of lipids in 10 species. Journal of Bacteriology, 93(3):894–903.
- Kaneda, T.** (1968). Fatty acids in the genus *Bacillus* II. similarity in the fatty acid compositions of *Bacillus thuringiensis*, *Bacillus anthracis*, and *Bacillus cereus*. Journal of Bacteriology, 95(6):2210–2216.
- Kaneda, T.** (1971). Factors affecting the relative ratio of fatty acids in *Bacillus cereus*. Canadian Journal of Microbiology, 17(2):269–275.
- Kaneda, T.** (1972). Positional preference of fatty acids in phospholipids of *Bacillus cereus* and its relation to growth temperature. Biochimica et Biophysica Acta (BBA), 280:297–305.
- Kaneda, T.** (1977). Fatty acids of the genus *Bacillus*: an example of branched-chain preference. Bacteriological reviews, 41(2):391–418.
- Kaneda, T.** (1991). Iso- and anteiso-fatty acids in bacteria: biosynthesis, function and taxonomic significance. Microbiological Reviews, 55(2):288–302.
- Kerstens, 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. In Fourteenth International Joint Conference on Artificial Intelligence, volume 2, pages 1137–1145, MontréAl, Canada. Morgan Kaufmann.
- Koller, D. and Sahami, M.** (1997). Hierarchically classifying documents using very few words. In 14th International Conference on Machine learning, pages 170–187, Nashville, Tennessee, USA.
- Konstantinidis, K. T., Ramette, A., and Tiedje, J. M.** (2006a). The bacterial species definition in the genomic era. Philosophical Transactions of the Royal Society B, 361:1929–1940.
- Konstantinidis, K. T., Ramette, A., and Tiedje, J. M.** (2006b). Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. Applied and Environmental Microbiology, 72(11):7286–7293.
- Konstantinidis, K. T. and Tiedje, J. M.** (2005a). Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences of the USA, 102(7):2567–2572.
- Konstantinidis, K. T. and Tiedje, J. M.** (2005b). Towards a genome-based taxonomy for prokaryotes. Journal of Bacteriology, 187(18):6258–6264.
- Konstantinidis, K. T. and Tiedje, J. M.** (2007). Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Current Opinion in Microbiology, 10:504–509.
- Kotilainen, P., Huovinen, P., and Eerola, E.** (1991). Application of gas-liquid chromatographic analysis of cellular fatty acids for species identification and typing of coagulase-negative Staphylococci. Journal of Clinical Microbiology, 29(2):315–322.
- Kriegel, H.-P., Kröger, P., Pryakhin, A., and Schubert, M.** (2004). Using support vector machines for classifying large sets of multi-represented objects. In 4th SIAM International Conference on Data Mining, pages 102–144, Lake Buena Vista, Florida, USA.

- Kunitsky, C., Osterhout, G., and Sasser, M.** (2006). Identification of microorganisms using fatty acid methyl ester (FAME) analysis and the MIDI Sherlock Microbial Identification System. In Miller, M., editor, Encyclopedia of Rapid Microbiological Methods, volume 3, pages 1–18. PDA, Bethesda.
- Kwok, S. W. and Carter, C.** (1990). Multiple decision trees.
- Lambert, M. A. and Moss, C. W.** (1983). Comparison of the effect of acid and base hydrolysis on hydroxy and cyclopropane fatty acids in bacteria. Journal of Clinical Microbiology, 18:1370–1377.
- Lancashire, L., Schmid, O., Shah, H., and Ball, G.** (2005). Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis. Bioinformatics, 21(10):2191–2199.
- Langkriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S.** (2004). A statistical framework for genomic data fusion. Bioinformatics, 20(16):2626–2635.
- Lang, E., Griese, B., Spröer, C., Schumann, P., Steffen, M., and Verbarq, S.** (2007). Characterization of ‘*Pseudomonas azelaica*’ DSM 9128, leading to emended descriptions of *Pseudomonas citronellosis* Seubert 1960 (Approved Lists 1980) and *Pseudomonas nitroreducens* Iizuka and Komagata 1964 (Approved Lists 1980), including *Pseudomonas multiresinivorans* as its later heterotypic synonym. International Journal of Systematic and Evolutionary Microbiology, 57:878–882.
- Lapage, S. P., Sneath, P. H. A., Lessel, E. F., Skerman, V. B. D., Seeliger, H. P. R., and Clark, W. A.** (1992). International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision. ASM Press, Washington D.C., USA.
- Lasko, T. A., Bhagwat, J. G., Zou, K. H., and Ohno-Machado, L.** (2005). The use of receiver operating characteristic curves in biomedical informatics. Journal of Biomedical Informatics, 28:404–415.
- Lee, J.-S. and Oh, I.-S.** (2003). Binary classification trees for multi-class classification problems. In Seventh International Conference on Document Analysis and Recognition, pages 1–5, Edinburgh, Scotland.
- Leslie, C. S., Eskin, E., Cohen, A., Weston, J., and Noble, W. S.** (2004). Mismatch string kernels for discriminative protein classification. Bioinformatics, 20(4):467–476.
- Letunic, I. and Bork, P.** (2007). Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics, 23(1):127–128.
- Li, W.-H.** (1997). Molecular Evolution. Sinauer Associates, Inc., Sunderland, Massachusetts, USA.
- Liaw, A.** (2009). randomForest: Breiman and Cutler’s random forests for classification and regression. URL: <http://cran.r-project.org/web/packages/randomForest/index.html>.
- Liaw, A. and Wiener, M.** (2002). Classification and regression by randomForest. R News, 2(3):18–22.
- Logan, N. A. and De Vos, P.** (2009). Genus I. *Bacillus Cohn 1872, 174^{AL}*. In De Vos, P., Garrity, G. M., Jones, D., Krieg, N. R., Ludwig, W., Rainey, F. A., Schleifer, K.-H., and Whitman, W. B., editors, Bergey’s Manual of Systematic Bacteriology Volume 3: The Firmicutes, volume 3, page 1330. Springer, New York, NY, USA, 2nd edition.
- Lu, Z., Lin, F., and Ying, H.** (2007). Design of decision tree via kernelized hierarchical clustering for multiclass support vector machines. Cybernetics and Systems, 38:187–202.
- Ludwig, W. and Klenk, H.-P.** (2001). Overview: A phylogenetic backbone and taxonomic framework for prokaryotic systematics. In Brenner, D. J., Krieg, N. R., and Staley, J. T., editors, Bergey’s Manual of Systematic Bacteriology Volume Two: The Proteobacteria Part A Introductory Essays, volume 2nd, pages 49–65. Springer, New York, NY, USA.
- Madigan, M. T., Martinko, J. M., Dunlap, P. V., and Clarck, D. P.** (2009). Brock Biology of Microorganisms. Pearson Education Inc., San Francisco, 12th edition.
- Marr, A. G. and Ingraham, J. L.** (1962). Effect of the temperature on the composition of fatty acids in *Escherichia coli*. Journal of Bacteriology, 84(6):1260–1267.
- Mccallum, A., Rosenfeld, R., Mitchell, T., and Ng, A.** (1998). Improving text classification by shrinkage in a hierarchy of classes. In 15th International Conference of Machine Learning, pages 359–367, Madison, Wisconsin, USA.
- McMurry, J. and Castellion, M. E.** (2002). Fundamentals of General, Organic and Biological Chemistry. Prentice

- Hall, Upper Saddle River, NJ, USA, 4th edition.
- MIBBI** (2009). MIBBI: Minimum Information for Biological and Biomedical Investigations. URL: <http://www.mibbi.org>.
- MIDI** (1990). Fatty acid profiling by gas chromatography for the Sherlock® MIS. Technical report, MIDI Inc. Year unknown.
- MIDI** (2003). Aerobic Bacteria Library (TSBA50/CLIN50). Technical report, MIDI Inc.
- MIDI** (2005a). Aerobic Bacteria Library (TSBA6/RTSBA6 - CLIN6/RCLIN6). Technical report, MIDI Inc.
- MIDI** (2005b). Sherlock® Microbial Identification System Version 6.0 - MIS Operating Manual.
- MIDI** (2005c). TSBA6/RTSBA6 Removed Species. Technical report, MIDI Inc.
- MIDI** (2009a). Instant FAME™. Zero In, Lock On, Identify. The New Sherlock® Microbial Identification System. Technical report, MIDI Inc.
- MIDI** (2009b). Website. URL: <http://www.midi-inc.com>.
- MIDI** (2009c). What is it? Sherlock® knows! Microbial Identification System. Technical report, MIDI Inc.
- Miller, L. T.** (1982). Single derivatization method for routine analysis of bacterial whole-cell fatty acid methyl esters, including hydroxy acids. *Journal of Clinical Microbiology*, 16(3):584–586.
- Minsky, M. and Papert, S.** (1969). *Perceptrons*. MIT Press, Cambridge, MA, USA.
- Mirkin, B.** (2005). *Clustering for data mining: a data recovery approach*. Computer Science and Data Analysis Series. Chapman and Hall/CRC, London, UK.
- Mitchell, T. M.** (1997). *Machine Learning*. McGraw-Hill, Boston.
- Miyajima, K., Tanii, A., and Akita, T.** (1983). *Pseudomonas fuscovaginae* sp. nov., nom. rev.. *International Journal of Systematic Bacteriology*, 33:656–657.
- Moens, M., Smet, A., Naudts, B., Verhoeven, J., Ieven, M., Jorens, P., Geise, H. J., and Blockhuys, F.** (2006). Fast identification of ten clinically important micro-organisms using an electronic nose. *Letters in Applied Microbiology*, 42(2):121–126.
- Mohn, W. W., Wilson, A. E., Bicho, P., and Moore, E. R.** (1999). Physiological and phylogenetic diversity of bacteria growing on resin acids. *Systematic and Applied Microbiology*, 22(1):68–78.
- Molin, K.** (2004). Analysis of cellular fatty acids for identification of microorganisms. Technical report, CCUG, Sweden.
- Molin, K.** (2008). CFA-FAME. Technical report, CCUG, Sweden.
- 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 relationships. *Systematic and Applied Microbiology*, 19:478–492.
- Moschetti, G., Blaiotta, G., Villani, F., Coppola, S., and Parente, E.** (2001). Comparison of statistical methods for identification of *Streptococcus thermophilus*, *Enterococcus faecalis*, and *Enterococcus faecium* from randomly amplified polymorphic DNA patterns. *Applied and Environmental Microbiology*, 67(5):2156–2166.
- Moss, C. W.** (1975). Gas-liquid chromatography as an analytical tool in microbiology. *Journal of Chromatography*, 203:337–347.
- Moss, C. W.** (1981). 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 obtained from clinical specimens. *Journal of Clinical Microbiology*, 4(6):492–502.
- Moss, C. W., Dees, S. B., and Guerrant, G. O.** (1980). Gas-liquid chromatography of bacterial fatty acids with a fused-silica capillary column. *Journal of Clinical Microbiology*, 12(1):127–130.
- Moss, C. W., Lambert, M. A., and Merwin, W. H.** (1974). Comparison of rapid methods for analysis of bacterial fatty acids. *Applied Microbiology*, 28(1):80–85.
- Moss, C. W. and Samuels, S. B.** (1974). Short-chain acids of *Pseudomonas* species encountered in clinical

- specimens. *Applied Microbiology*, 27(3):570–574.
- Moss, C. W., Samuels, S. B., and Weaver, R. E. (1972). Cellular fatty acid composition of selected *Pseudomonas* species. *Applied Microbiology*, 24(4):596–598.
- Moura, H., Woolfitt, A. R., Carvalho, M. G., Pavlopoulos, A., Teixeira, L. M., Satten, G. A., and Barr, J. R. (2008). MALDI-TOF mass spectrometry as a tool for differentiation of invasive and noninvasive *Streptococcus pyogenes* isolates. *FEMS Immunology and Medical Microbiology*, 53(3):333–342.
- Mouwen, D. J. M., Capita, R., Alonso-Calleja, C., Prieto-Gómez, J., and Prieto, M. (2006). Artificial neural network based identification of *Campylobacter* species by Fourier transform infrared spectroscopy. *Journal of Microbiological Methods*, 67(1):131–140.
- Mukwaya, G. M. and Welch, D. F. (1989). Subgrouping of *Pseudomonas cepacia* by cellular fatty acid composition. *Journal of Clinical Microbiology*, 27(12):2640–2646.
- Munsch, P., Alatosava, T., Marttinen, N., Meyer, J.-M., Christen, R., and Gardan, L. (2002). *Pseudomonas constantinii* 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.
- NCBI (2009a). Complete Microbial Genomes. URL: <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>.
- NCBI (2009b). Taxonomy browser. URL: <http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi>.
- Nielsen, P., Fritze, D., and Priest, F. G. (1995). Phenetic diversity of alkaliphilic *Bacillus* strains: proposal for nine new species. *Microbiology*, 141(7):1745–1761.
- Nishimori, E., Kita-Tsukamoto, K., and Wakabayashi, H. (2000). *Pseudomonas plecoglossicida* sp. nov., the causative agent of bacterial haemorrhagic ascites of ayu, *Plecoglossus altivelis*. *International Journal of Systematic and Evolutionary Microbiology*, 50:83–89.
- O'Donnell, A. G., Nahaie, M. R., Goodfellow, M., Minnikin, D. E., and Hájek, V. (1985). Numerical analysis of fatty acid profiles in the identification of Staphylococci. *Journal of General Microbiology*, 131:2023–2033.
- O'Leary, W. M. (1962). The fatty acids of bacteria. *Bacteriological reviews*, 26(4):421–447.
- 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(9):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(1):17–40.
- Page, R. D. M. and Holmes, E. C. (1998). *Molecular Evolution. A Phylogenetic Approach*. Blackwell Science Ltd., Oxford, UK.
- Palleroni, N. J. (1984). Genus I. *Pseudomonas* Migula 1984. In Krieg, N. R. and Holt, J. G., editors, *Bergey's Manual of Systematic Bacteriology Volume 1*, volume 1, pages 141–199. Williams and Wilkins, Baltimore, USA.
- Palleroni, N. J. (2005). Genus I. *Pseudomonas* Migula 1894, 237^{AL}. In Brenner, D. J., Krieg, N. R., and Staley, J. T., editors, *Bergey's Manual of Systematic Bacteriology Volume 2: The Proteobacteria Part B The Gammaproteobacteria*, volume 2, pages 323–379. Springer, New York, NY, USA.
- Palleroni, N. J. (2008). The road to the taxonomy of *Pseudomonas*. In Cornelis, P., editor, *Pseudomonas: Genomics and Molecular Biology*, pages 1–18. Caister Academic Press, Norfolk.
- Palleroni, N. J., Kunisawa, R., Contopoulou, R., and Doudoroff, M. (1973). Nucleic acid homologies in the genus *Pseudomonas*. *Systematic and Applied Microbiology*, 23(4):333–339.
- Parker, J. P., Günter, S., and Bedo, J. (2007). Stratification bias in low signal microarray studies. *BMC Bioinformatics*, 8(326):1–16.
- Peix, A., Berge, O., Rivas, R. u. I., Abril, A., and Velázquez, E. (2005). *Pseudomonas argentinensis* sp. nov., a novel yellow pigment-producing bacterial species, isolated from rhizospheric soil in Córdoba, Argentina. *International Journal of Systematic and Evolutionary Microbiology*, 55:1107–1112.
- Pineiro-Vidal, M., Pazos, F., and Santos, Y. (2008). Fatty acid analysis as a chemotaxonomic tool for taxonomic and epidemiological characterization of four fish pathogenic *Tenacibaculum* species. *Letters in Applied*

- Microbiology*, 46(5):548–554.
- Piraino, P., Ricciardi, A., Salzano, G., Zotta, T., and Parente, E.** (2006). Use of unsupervised and supervised artificial neural networks for the identification of lactic acid bacteria on the basis of SDS-PAGE patterns of whole cell proteins. *Journal of Microbiological Methods*, 66(2):336–346.
- Platt, J. C., Cristianini, N., and Shawe-Taylor, J.** (2000). Large margin DAGs for multiclass classification. *Advances in Neural Information Processing Systems 12*, pages 547–553.
- Priest, F. G.** (2009). Genus XX. *Paenibacillus* Ash, Priest and Collins 1994, 852^{VP}. In De Vos, P., Garrity, G. M., Jones, D., Krieg, N. R., Ludwig, W., Rainey, F. A., Schleifer, K.-H., and Whitman, W. B., editors, *Bergey's Manual of Systematic Bacteriology Volume 3: The Firmicutes*, volume 3, page 1330. Springer, New York, 2nd edition.
- Priest, F. G., Barker, M., Baillie, L. W. J., Holmes, E. C., and Maiden, M. C. J.** (2004). Population structure and evolution of the *Bacillus cereus* group. *Journal of Bacteriology*, 186(23):7959–7990.
- Provost, F. and Domingos, P.** (2001). Well-trained PETs: improving probability estimation trees. Technical report, Stern School of Business, New York University.
- Provost, F. and Fawcett, T.** (2001). Robust classification for imprecise environments. *Machine Learning*, 42:203–231.
- Provost, F., Fawcett, T., and Kohavi, R.** (1998). The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 445–453, Madison, USA. Morgan-Kaufmann.
- 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 compatible with ARB. *Nucleic Acids Research*, 35(21):7188–7196.
- Quezada, M., Buitron, G., Moreno-Andrade, I., Moreno, G., and Lopez-Marin, L. M.** (2007). The use of fatty acid methyl esters as biomarkers to determine aerobic, facultatively aerobic and anaerobic communities in wastewater treatment systems. *FEMS Microbiology Letters*, 266(1):75–82.
- Quinlan, J. R.** (1993). *C4.5: Programs for machine learning*. Morgan Kaufman series in machine learning. Morgan Kaufman Publishers, Inc., San Mateo, California, USA.
- Quinteiro Rodríguez, M. P.** (2000). Fourier transform infrared (FTIR) technology for the identification of organisms. *Clinical Microbiology Newsletter*, 22(8):57–61.
- Rasko, D. A., Altherr, M. R., Han, C. S., and Ravel, J.** (2005). Genomics of the *Bacillus cereus* group of organisms. *FEMS Microbiology Reviews*, 29:303–329.
- Raviv, Y. and Intrator, N.** (1996). Bootstrapping with noise: an effective regularization technique. *Connection Science*, 8(3–4):355–372.
- Rebuffo, C. A., Schmitt, J., Wenning, M., von Stetten, F., and Scherer, S.** (2006). Reliable and rapid identification of *Listeria monocytogenes* and *Listeria* species by artificial neural network-based Fourier transform infrared spectroscopy. *Applied and Environmental Microbiology*, 72(2):994–1000.
- Rebuffo-Scheer, C. A., Schmitt, J., and Scherer, S.** (2007). Differentiation of *Listeria monocytogenes* serovars by using artificial neural network analysis of fourier-transformed infrared spectra. *Applied and Environmental Microbiology*, 73(3):1036–1040.
- Reddy, G. S. N., Matsumoto, G. I., Schumann, P., Stackebrandt, E., and Shivaji, S.** (2004). Psychrophilic pseudomonads from Antarctica: *Pseudomonas antarctica* sp. nov., *Pseudomonas meridiana* sp. nov. and *Pseudomonas proteolytica* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 54:713–719.
- Riedmiller, M. and Braun, H.** (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, San Francisco, USA.
- Rifkin, R. and Klautau, A.** (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141.
- Rilfors, L., Wieslander, A., and Ståhl, S.** (1978). Lipid and protein composition of membranes of *Bacillus megaterium* variants in the temperature range 5 to 70°C. *Journal of Bacteriology*, 135(3):1043–1052.
- Rivas, R. u. I., Mateos, P. F., Martínez-Molina, E., and Velázquez, E.** (2005). *Paenibacillus xylanilyticus* sp.

- nov., an airborne xylanolytic bacterium. International Journal of Systematic and Evolutionary Microbiology, 55(1):405–408.
- Rock, C. O. and Jackowski, S.** (2002). Forty years of bacterial fatty acid synthesis. Biochemical and Biophysical Research Communications, 292:1155–1166.
- Rösch, P., Harz, M., Schmitt, M., Peschke, K.-D., Ronneberger, O., Burkhardt, H., Motzkus, H.-W., Lankers, M., Hofer, S., Thiele, H., and Popp, J.** (2005). Chemotaxonomic identification of single bacteria by micro-Raman spectroscopy: application to clean-room-relevant biological contaminations. Applied and Environmental Microbiology, 71(3):1626–1637.
- Rosselló-Mora, R. and Amann, R.** (2001). The species concept for prokaryotes. FEMS Microbiology Reviews, 25:39–67.
- Rousu, J., Saunders, C., Szedmak, S., and Shawe-Taylor, J.** (2006). Kernel-based learning of hierarchical multi-label classification models. Journal of Machine Learning Research, 7:1601–1626.
- Ruggiero, C., Giacomini, M., Calegari, F., Berti, R., Bertone, S., and Casareto, L.** (1993). Interpretation of gaschromatographic data via artificial neural networks for the classification of marine bacteria. Cytotechnology, 11:S83–85.
- Saito, K.** (1960a). Chromatographic studies on bacterial fatty acids. The Journal of Biochemistry, 47:699)709.
- Saito, K.** (1960b). Studies on bacterial fatty acids. Structure of subtilopentadecanoic and subtiloheptadecanoic acids. The Journal of Biochemistry, 47(6):710–719.
- Saitou, N. and Nei, M.** (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4:406–425.
- Sapp, J.** (2005). The bacterium's place in nature. In Sapp, J., editor, Microbial Phylogeny and Evolution, pages 3–52. Oxford University Press, New York, NY, USA.
- Sasser, M.** (1990). Bacterial Identification by Gas Chromatographic Analysis of Fatty Acids Methyl Esters (GC-FAME) - MIDI Technical Note #101. Technical report, MIDI, Inc.
- Satten, G. A., Datta, S., Moura, H., Woolfitt, A. R., Carvalho, M. G., Carlone, G. M., De, B. K., Pavlopoulos, A., and Barr, J. R.** (2004). Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. Bioinformatics, 20(17):3128–3136.
- Schapire, R., Freund, Y., Bartlett, P., and Lee, W. S.** (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. The Annals of Statistics, 26(5):1651–1686.
- Schmid, O., Ball, G., Lancashire, L., Culak, R., and Shah, H.** (2005). New approaches to identification of bacterial pathogens by surface enhanced laser desorption/ionization time of flight mass spectrometry in concert with artificial neural networks, with special reference to *Neisseria gonorrhoeae*. Journal of Medical Microbiology, 54(12):1205–1211.
- Schölkopf, B. and Smola, A.** (2002). Learning with Kernels. MIT Press, Cambridge, MA, USA.
- Schwenker, F. and Palm, G. u. n.** (2001). Tree-structured support vector machines for multi-class pattern recognition. Lecture Notes in Computer Science, 2096:409–417.
- Seurinck, S., Deschepper, E., Deboch, B., Verstraete, W., and Siciliano, S.** (2005). Characterization of *Escherichia coli* isolates from different fecal sources by means of classification tree analysis of fatty acid methyl ester (FAME) profiles. Environmental Monitoring and Assessment, 114(1-3):433–445.
- Shaffer, J. P.** (1995). Multiple hypothesis testing. Annual Review of Psychology, 46:561–584.
- Shawe-Taylor, J. and Cristianini, N.** (2004). Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, UK.
- Shen, P. Y., Coles, E., Foote, J. L., and Stenesh, J.** (1970). Fatty acid distribution in mesophilic and thermophilic strains of the genus *Bacillus*. Journal of Bacteriology, 103(2):479–481.
- Sheskin, D. J.** (2004). Handbook of Parametric and Nonparametric Statistical Procedures. Chapman and Hall/CRC, Florida, USA, 3rd edition.
- Shivaji, S., Chaturvedi, P., Suresh, K., Reddy, G. S. N., Dutt, C. B. S., Wainwright, M., Narlikar, J. V., and Bhargava, P. M.** (2006). *Bacillus aerius* sp. nov., *Bacillus aerophilus* sp. nov., *Bacillus stratosphericus* sp. nov. and *Bacillus altitudinis* sp. nov., isolated from cryogenic tubes used for collecting air samples from high

- altitudes. International Journal of Systematic and Evolutionary Microbiology, 56:1465–1473.
- Sikorski, J. and Nevo, E.** (2005). Adaptation and incipient sympatric speciation of *Bacillus simplex* under microclimatic contrast at "Evolution Canyons" I and II, Israel. Proceedings of the National Academy of Sciences of the USA, 102(44):15924–15929.
- Sikorski, J. and Nevo, E.** (2007). Patterns of thermal adaptation of *Bacillus simplex* to the microclimatically contrasting slopes of 'Evolution Canyons' I and II, Israel. Environmental Microbiology, 9(3):716–726.
- Smith, I. M., Dunez, J., Philips, D. H., Lelliot, R. A., and Archer, S. A.** (1988). In European Handbook of Plant Diseases. Blackwell Scientific Publications, Oxford, UK.
- Sneath, P. H. A. and Brenner, D. J.** (1992). "official" nomenclature lists. ASM News, 58:175.
- Sokal, R. R. and Michener, C. D.** (1958). A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, 38:1409–1438.
- Song, Y., Yang, R., Guo, Z., Zhang, M., Wang, X., and Zhou, F.** (2000). Distinctness of spore and vegetative cellular fatty acid profiles of some aerobic endospore-forming bacilli. Journal of Microbiological Methods, 39:225–241.
- Stackebrandt, E. and Goebel, B. M.** (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. International Journal of Systematic Bacteriology, 44(4):846–849.
- Stackebrandt, E. and Swiderski, J.** (2002). From phylogeny to systematics: the dissection of the genus *Bacillus*. In Berkeley, R., Heyndrickx, M., Logan, N., and De Vos, P., editors, Applications and Systematics of Bacillus and Relatives, pages 8–22. Blackwell Science Ltd., Malden, USA.
- Stamatakis, A.** (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analysis with thousands of taxa and mixed models. Bioinformatics, 22(21):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(2):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.
- Steele, M., McNab, W. B., Read, S., Poppe, C., Harris, L., Lammerding, A. M., and Odumeru, J. A.** (1997). Analysis of whole-cell fatty acid profiles of verotoxigenic *Escherichia coli* and *Salmonella enteritidis* with the Microbial Identification System. Applied and Environmental Microbiology, 63(2):757–760.
- Stolz, A., Busse, H.-J., and Kämpfer, P.** (2007). *Pseudomonas knackmussii* sp. nov. International Journal of Systematic and Evolutionary Microbiology, 57:572–576.
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A.** (2008). Conditional variable importance for random forests. BMC Bioinformatics, 9(1):307.
- 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.
- Tibshirani, R.** (1996). Bias, variance and prediction error for classification rules. Technical report, University of Toronto.
- Tijms, H.** (2004). Understanding Probability. Chance Rules in Everyday Life. Cambridge University Press, Cambridge, UK.
- Topić, G.** (2004). PARF - Parallel RF Algorithm. URL: <http://www.irb.hr/en/research/projects/it/2004/2004-111/>.
- Tourasse, N. J., Helgason, E., Økstad, O. A., Hegna, I. K., and Kolstø, A.-B.** (2006). The *Bacillus cereus* group: novel aspects of population structure and genome dynamics. Journal of Applied Microbiology, 101:579–593.
- Tuang, F. N., Rademaker, J. L. W., Alocilja, E. C., Louws, F. J., and De Bruijn, F. J.** (1999). Identification of bacterial rep-pcr genomic fingerprints using a backpropagation neural network. FEMS Microbiology Letters, 177(2):249–256.
- Tvzrová, L., Schumann, P., Spröer, C., Sedláček, I., Páčová, Z., Šedo, O., Zdráhal, Z., Steffen, M., and Lang, E.** (2006). *Pseudomonas moraviensis* sp. nov. and *Pseudomonas vranovensis* sp. nov., soil bacteria

- isolated on nitroaromatic compounds, and emended description of *Pseudomonas asplenii*. International Journal of Systematic and Evolutionary Microbiology, 56:2657–2663.
- Čechová, L., Durnová, E., Šikutová, S., Halouzka, J., and Němec, M. (2004). Characterization of spirochetal isolates from arthropods collected in South Moravia, Czech Republic, using fatty acid methyl ester analysis. Journal of Chromatography B, 808:249–254.
- Udelhoven, T., Naumann, D., and Schmitt, J. (2000). Development of a hierarchical classification system with artificial neural networks and FT-IR spectra for the identification of bacteria. Applied Spectroscopy, 54(10):1471–1479.
- Vaerewijck, M. J. M., De Vos, P., Lebbe, L., Scheldeman, P., Hoste, B., and Heyndrickx, M. (2001). Occurrence of *Bacillus sporothermodurans* and other aerobic spore-forming species in feed concentrate for dairy cattle. Journal of Applied Microbiology, 91:1074–1084.
- 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(4):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.
- Vandamme, P., Pot, B., Gillis, M., De Vos, P., Kersters, K., and Swings, J. (1996). Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiological Reviews, 60(2):407–438.
- Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY, USA.
- Vapnik, V. (1998). Statistical Learning Theory. Wiley, New York, NY, USA.
- Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7(1):91.
- Vens, C., Struyf, J., Schietgat, L., Dzeroski, S., and Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. Machine Learning, 73:185–214.
- Vert, J.-P. (2002). A tree kernel to analyse phylogenetic profiles. Bioinformatics, 18(suppl1):S276–284.
- Vilas-Bôas, G. T., Peruca, A. P. S., and Arantes, O. M. N. (2007). Biology and taxonomy of *Bacillus cereus*, *Bacillus anthracis*, and *Bacillus thuringiensis*. Canadian Journal of Microbiology, 53:673–687.
- Voisin, S., Terreux, R., Renaud, F. N. R., Freney, J., Domard, M., and Deruaz, D. (2004). Pyrolysis patterns of 5 close *Corynebacterium* species analyzed by artificial neural networks. Antonie van Leeuwenhoek International Journal of General and Molecular Microbiology, 85(4):287–296.
- Vural, V. and Dy, J. G. (2004). A hierarchical method for multi-class support vector machines. In 21st International Conference on Machine Learning, pages 1–8, Banff, Canada.
- Wang, L.-T., Lee, F.-L., Tai, C.-J., and Kasai, H. (2007a). Comparison of *gyrB* gene sequences, 16S rRNA gene sequences and DNA-DNA hybridization in the *Bacillus subtilis* group. International Journal of Systematic and Evolutionary Microbiology, 57(8):1846–1850.
- Wang, L.-T., Lee, F.-L., Tai, C.-J., and Kuo, H.-P. (2008). *Bacillus velezensis* is a later heterotypic synonym of *Bacillus amyloliquefaciens*. International Journal of Systematic and Evolutionary Microbiology, 58(3):671–675.
- Wang, L.-T., Lee, F.-L., Tai, C.-J., Yokota, A., and Kuo, H.-P. (2007b). Reclassification of *Bacillus axarquiensis* Ruiz-Garcia et al. 2005 and *Bacillus malacitensis* Ruiz-Garcia et al. 2005 as later heterotypic synonyms of *Bacillus Mojavensis* Roberts et al. 1994. International Journal of Systematic and Evolutionary Microbiology, 57(7):1663–1667.
- Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E., Stackebrandt, E., Starr, M. P., and Trüper, H. G. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. International Journal of Systematic Bacteriology, 37(4):463–464.
- Weiss, G. M. and Provost, F. (2003). Learning when training data are costly: the effect of class distribution on tree induction. Journal of Artificial Intelligence Research, 19:315–354.

- Welch, D. F.** (1991). Applications of cellular fatty acid analysis. *Clinical Microbiology Reviews*, 4(4):422–438.
- Weston, J. and Watkins, C.** (1998). Multi-class support vector machines. Technical report, Department of Computer Science.
- WFCC** (2009). World Federation of Culture Collections. URL: <http://www.wfcc.info/>.
- Whitman, W. B.** (2009). The modern concept of the prokaryote. *Journal of Bacteriology*, 191(7):2000–2005.
- Williamson, Y. M., Moura, H., Woolfitt, A. R., Pirkle, J. L., Barr, J. R., Carvalho, M. D. G., Ades, E. P., Carlone, G. M., and Sampson, J. S.** (2008). Differentiation of *Streptococcus pneumoniae* conjunctivitis outbreak isolates by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Applied and Environmental Microbiology*, 74(19):5891–5897.
- Witten, I. H. and Frank, E.** (2005). *Data Mining Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.
- Witten, I. H. and Frank, E.** (2009). WEKA data mining software. URL: <http://www.cs.waikato.ac.nz/ml/weka/>.
- Woese, C. R. and Fox, G. E.** (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the USA*, 74(11):5088–5090.
- Wolpert, D. H. and MacReady, W. G.** (1999). An efficient method to estimate bagging's generalization error. *Machine Learning*, 35:41–55.
- Wu, F., Zhang, J., and Honavar, V.** (2005). Learning classifiers with hierarchically structured class taxonomies. *Lecture Notes in Computer Science*, 3607:313–320.
- Xia, S., Li, J., Xia, L., and Ju, C.** (2007). Tree-structured support vector machines for multi-class classification. *Lecture Notes in Computer Science*, 4493:392–398.
- Xu, M., Voorhees, K. J., and Hadfield, T. L.** (2003). Repeatability and pattern recognition of bacterial fatty acid profiles generated by direct mass spectrometric analysis of in situ thermal hydrolysis/methylation of whole cells. *Talanta*, 59:577–589.
- Yang, Y.-C., Yu, H., Xiao, D.-W., Liu, H., Hu, Q., Huang, B., Liao, W.-J., and Huang, W.-F.** (2009). Rapid identification of *Staphylococcus aureus* by surface enhanced laser desorption and ionization time of flight mass spectrometry. *Journal of Microbiological Methods*, 77(2):202–206.
- Yarza, P., Richter, M., Peplies, J., Euzéby, J. P., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F. O. and Rosselló-Mora, R.** (2009). The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Systematic and Applied Microbiology*, 31:241–250.
- Yong, C. K., Lim, C. M., Plumbley, M. D., Beighton, D., and Davidson, R.** (2002). Identification of dental bacteria using statistical and neural approaches. In *9th International Conference on Neural Information Processing*, pages 606–610, Singapore.

Summary

The bacterial landscape is a continuously changing medium. Almost every day, a new bacterial species is described. In view of this evolution, it is critical to keep identification methods up-to-date with the current bacterial taxonomy. Generally, bacterial identification can be based on the genotype as well as on the phenotype. In other words, it can rely on the genetic composition as well as on the observable characteristics. Bacterial species identification is mainly performed by genotypic analysis, though the generally accepted species definition states that the phenotype should confirm the findings based on the genotype. It is clear that the phenotype was, is and will remain a key player in bacterial taxonomy.

Phenotypic methods are routinely used to achieve a fast bacterial identification. Most methods allow to discriminate between bacteria at the genus level and the species level, though also methods exist that allow for an identification at strain level. Besides the analysis of proteins, enzymes, metabolism or morphology, gas chromatographic analysis of bacterial fatty acids is a routinely applied method used in many laboratory and institutes. The main reasons are the low cost and rapid analysis of each run, and the possibilities for automation and high-throughput analysis. For gas chromatographic analysis, fatty acids are methylated and esterified (i.e. fatty acid methyl ester or FAME), which make the fatty acid more volatile. At present, bacterial FAME analysis is commercially exploited by the company MIDI, Inc. (Newark, Delaware, USA). Their identification system Sherlock MIS allows bacterial species identification in many environmental, clinical and industrial niches. However, a major problem with phenotypic methods is that they rely on identification libraries that mostly are not up-to-date with the current bacterial taxonomy. This is also the case for FAME analysis. Every month, several new and re-evaluated bacterial species are validly published. This requires a possible update of the identification libraries every month. Besides this, identification libraries are often constructed for the identification of important bacterial groups. For the reasons above, phenotypic methods are typically used as first-line identification methods. When phenotypic analysis is compared to genotypic analysis, it is also important to mention that different environmental conditions may influence the phenotype, such as temperature, pH, atmosphere, etc. In order to use FAME data for identification, for comparative analysis or, simply, for data sharing, it is critical to focus only on data that is resulting from bacteria grown and cultured under standard growth and culture conditions. Only under these conditions, stable fatty acid patterns are obtained. In this study, we focused on the protocol as described by the peak naming table TSBA50 of the Sherlock MIS system.

In this dissertation, we studied bacterial species identification based on the data available in the joint FAME database of the Laboratory of Microbiology (Ghent University, Belgium) and

the BCCMTM/LMG Bacteria Collection (Belgium). Since 1989, both groups have performed bacterial FAME analysis for identification, quality control, comparative studies and numerical analysis. Twenty years of fatty acid research has resulted in a FAME database that currently contains more than 71,000 FAME profiles. This database is an ideal source for research on data mining and knowledge discovery. With this dissertation, we aimed at an improved FAME-based bacterial species identification by using machine learning techniques. Three techniques were used: artificial neural networks, support vector machines and random forests. In a supervised setting, the goal was to distinguish between the different bacterial species. We chose to model the species of the genera *Bacillus*, *Paenibacillus* and *Pseudomonas* because of the large amount of data available in the database and the expertise at the Laboratory of Microbiology. Starting from the FAME database, standard FAME profiles were exported for all validly described species belonging to the respective genera. Based on those profiles, different data sets were created: three genus data sets, a genera data set and two data sets regarding plant pathogenicity. In the latter case, the first data set consists of plant-pathogenic and non-plant-pathogenic *Pseudomonas* species, while the second data set only comprises plant-pathogenic *Pseudomonas* species.

For the different data sets, a standard data analysis was performed. Similar to several studies handling numerical analysis of different bacterial genera, we calculated average FAME profiles and peak distributions. By these analyses, the fatty acid composition was evaluated and core-genus, species-specific and strain-specific peaks were observed. A second experiment dealt with peak clustering with and without an additional clustering of the different species. From this analysis, it was clear that many species had similar FAME profiles. These findings were supported by a TaxonGap analysis, that clearly indicated that FAME data cannot be used as a taxonomic marker at species level. Finally, we performed a principal component analysis from which it could be concluded that different fatty acids are correlated. From the biplots of the first two principal components and for the three genera considered, we observed overlapping FAME data for a majority of the species. This was especially true for the species of the genus *Pseudomonas*.

Following this data analysis, a first machine learning experiment was performed to investigate how well artificial neural networks were able to distinguish between the different *Bacillus* species. Different parameter settings were investigated and statistically analyzed. This concerns the use of imbalanced or balanced data sets, the use of simple validation or cross-validation, and which neural network activation function should be chosen. Finally, the best *Bacillus* species identification was achieved by an imbalanced data set and by validating the models by stratified cross-validation. Statistical analysis also showed that selection of a good activation function required an additional optimization step. In a second phase, we expanded our machine learning research towards three bacterial genera. Two identification strategies were investigated: a stratified approach where profiles were identified from genus level to species level and a straight identification approach by which all species of the three genera were distinguished from each other. In the first approach, two types of data sets were used: a genera data set with profiles annotated by genus name and three genus data sets with profiles annotated by genus and species name. In a second approach, only one data set was used that comprised all profiles of all species

of the three genera, consisting of profiles annotated by genus and species name. Three machine learning techniques were evaluated in both approaches: artificial neural networks, support vector machines and random forests. A better performance was obtained with random forests in the stratified approach. A good species identification was obtained for the species of the genus *Paenibacillus*, a relatively good identification within the genus *Bacillus* and a low to moderate identification within the genus *Pseudomonas*. All results were discussed with respect to the findings in the data analysis experiments and to literature. Identification performance was also compared to the identification reported by the commercial system Sherlock MIS, though only for the species included in both systems. A clear improvement of the identification was obtained for the considered species. Furthermore, when compared to the commercial system, the machine learning approach has a distinct advantage. With the monthly changing bacterial taxonomy, it is critical that identification systems are kept up-to-date. Machine learning techniques easily handle this problem by a fast update of the corresponding data sets and a fast retraining of the respective models. Finally, the machine learning models were also tested on independent test sets which also resulted in an improved identification.

From the data analysis and the machine learning experiments, it was clear that the resolution of FAME analysis for bacterial species identification is limited. The models developed in the previous experiments did, however, not allow to determine the ability of FAME data to distinguish between the different bacterial species. Hence, a next step was set in our machine learning research in which we investigated how taxonomic and phylogenetic knowledge could be integrated in the classification models. A straightforward possibility was to construct a taxonomic or phylogenetic tree and to use a binary tree classifier to train a model on each node of the tree. Herein, each model tries to distinguish between the species and/or species groups corresponding to the two branches splitting from the corresponding node. Two strategies were evaluated. First, we focused on the algorithm of the binary tree classifier in which a tree was inferred from the respective data, that was subsequently used as a hierarchical classification system. Thus, in this experiment, we investigated the construction of a FAME tree. By a proof-of-concept experiment with 15 *Bacillus* species, a FAME tree was constructed by divisive clustering of the identification scores obtained by the model. This tree was subsequently used as a binary tree classifier for classification of the different bacterial species. A good performance was obtained. However, extending this experiment to all 74 *Bacillus* species was computationally infeasible. In a second strategy, a phylogenetic tree was inferred from 16S rRNA gene sequences. Sequences were selected in the manually controlled sequence database SILVA and the two methods of neighbour-joining and UPGMA were used for tree inference. The two resulting phylogenetic trees were subsequently also used as binary tree classifiers. In this setting, two types of data were used. The phylogenetic tree inferred from the 16S rRNA gene sequences determined the different binary classification tasks which were performed using FAME data. This classification approach has not been described before. Because this classification was based on FAME data and the models integrated data describing evolutionary relationships, we called this method phylogenetic learning. Compared to flat multi-class classification, a lower classification performance was obtained using this hierarchical framework. However, in this framework it is possible to analyze, evaluate, exploit and visualize the resolution of FAME data for bacterial

species discrimination. Herein, a possible post-processing approach was to develop a pruning method in order to obtain an optimized classification model. Also another important finding was obtained: species wrongly identified by a flat multi-class classification model were better identified with phylogenetic learning. Moreover, when looking at misclassified profiles, most misclassifications occurred in the first parent nodes of the different leaves. With this analysis, the limited resolution of FAME data became also immediately clear, which supports again the approach of integrating this resolution in classification models. In an initial experiment on this topic, the resolution was visualized by statistical analysis of the identification results of the different binary tree classifiers. The species and species groups that were hard to distinguish became clearly visualized.

In summary, with the goal of improving bacterial species identification by FAME data, we constructed different models for a computational identification by machine learning techniques. Different data analysis experiments were performed to analyze the data and different machine learning techniques were evaluated for a genus-wide identification of the species of three genera. Compared to the identifications obtained by the commercial system Sherlock MIS, an improved performance was obtained for the species included in both systems. Different identifications were performed on independent test sets. By the first machine learning experiments, different advantages became obvious: fast model construction, improved identification and, importantly, the possibility of keeping the models up-to-date with the current bacterial taxonomy. This dissertation is a good base for further extending the computational research towards more bacterial genera. It is, however, important to mention that not all bacteria grow under standard growth and culture conditions, and that many bacteria are even unculturable. In the former case, clear agreements should be made for an objective and reliable research.

By a hierarchical classification approach, we investigated the possibility of knowledge integration into classification models. Two strategies were evaluated and the combination of FAME data with 16S rRNA gene sequence data was promising. By this method, the resolution of FAME data can rapidly be evaluated by the integration of data that allows to discriminate the considered species. This approach was investigated by a statistical analysis of the identification results of each binary tree classifier. A better optimization of this approach requires, however, further research.

As a final research topic, we investigated how the restrictions imposed by the in-house FAME database could be solved. From our research, it was immediately clear that the experiments were restricted by the limited amount of data for each bacterial species. This does not only correspond to the number of profiles per species, but also to the number of strains per species and the number of species per genus. Therefore, we launched a public FAME database, FAME-bank.net, for the online sharing and querying of bacterial FAME profiles. In this way, we provide a possible solution for further extending the performed research by a larger number of genera, species, strains and FAME profiles.

Samenvatting

Het bacteriële landschap is een sterk evoluerend medium. Zo wordt quasi iedere dag een nieuwe bacteriële soort beschreven. Gelet op deze evolutie is het uitermate van belang dat identificatiemethoden up-to-date gehouden worden met de huidige bacteriële taxonomie. In het algemeen kan identificatie van bacteriën gebeuren op basis van zowel het genotype als het fenotype. Of, anders geformuleerd, identificatie kan zowel gericht zijn op de genetische samenstelling als op alle waarneembare karakteristieken van een bacterie. De klemtoon voor de identificatie van een bacteriële soort ligt op het genotype, maar de algemeen erkende soortdefinitie stelt heel duidelijk dat het fenotype de genotypische identificatie moet bevestigen. Het fenotype was, is en blijft dus een zeer belangrijke speler in bacteriële taxonomie.

Fenotypische methoden worden heel vaak routinematig aangewend om tot een snelle identificatie te komen. De meeste methoden laten toe om bacteriën te onderscheiden tot op het genus- en soortniveau, al hebben sommige technieken ook de mogelijkheid om tot een identificatie op stamniveau te komen. Naast het analyseren van bijvoorbeeld enzymen, eiwitten, metabolische reacties of morfologie, is ook een gaschromatografische analyse van bacteriële vetzuren een wijdverspreide methode die routinematig aangewend wordt in heel veel laboratoria en instituten. De hoofdredenen hiertoe omvatten de lage kostprijs en de hoge snelheid per analyse, samen met de mogelijkheid tot automatisering en de daaraan gekoppelde hoge verwerkingscapaciteit. Om een snelle en eenvoudige gaschromatografische analyse toe te laten, worden alle geëxtraheerde vetzuren gemethyleerd en veresterd. In hun methylester vorm (in Engels: fatty acid methyl ester of FAME) zijn vetzuren meer volatiel en dus meer geschikt voor gaschromatografische analyse. Op dit moment wordt bacteriële FAME-analyse commercieel geëxploiteerd door de firma MIDI, Inc. (Newark, Delaware, USA) die met hun identificatiesysteem Sherlock MIS bacteriële soortidentificatie toelaten in een groot aantal niches met een potentieel belang in het milieu, de klinische praktijk en de industrie. Het grote probleem van fenotypische methoden, en dus ook van FAME analyse, is echter dat identificatie aan de hand van deze methoden vaak gebaseerd is op bibliotheken die niet up-to-date zijn met de huidige bacteriële taxonomie. Maandelijks worden nieuwe updates gepubliceerd met beschrijvingen van nieuwe soorten als ook herbeschrijvingen van reeds beschreven soorten. Het vraagt dus veel werk om deze identificatiebibliotheken gesynchroniseerd te houden met de huidige taxonomie. Daarnaast zijn fenotypische identificatiebibliotheken er ook vaak op gericht om interessante bacteriële groepen te identificeren. Om deze redenen wordt fenotypische analyse dan ook vaak aangewend als een eerstelijns identificatiemethode. Wanneer fenotypische analyse tegenover genotypische analyse wordt gesteld, dan is het ook zeer belangrijk om aan te geven dat het fenotype (sterk) onderhevig is aan omgevingsfactoren zoals temperatuur, pH, atmosfeer, etc. Om fenotypische data te kun-

nen gebruiken voor identificatie, voor vergelijkende studies of, simpelweg, om uit te wisselen, dan is het strikt noodzakelijk dat de gegevens afkomstig zijn van analyse onder gestandaardiseerde condities. Zo is, in het geval van gaschromatografische vetzuuranalyse, het uitermate van belang dat bacteriële stammen opgegroeid worden onder standaard groei- en cultuurcondities. Enkel onder deze omstandigheden is het mogelijk om stabiele vetzuurprofielen te bekomen. In deze studie hebben we gewerkt met de condities zoals beschreven door de Sherlock MIS voor de TSBA50 piekentabel.

In deze studie richten we ons op FAME analyse voor bacteriële identificatie en, meer specifiek, op de gezamenlijke FAME databank van het Laboratorium voor Microbiologie (Universiteit Gent, België) en de BCCMTM/LMG Bacterie Collectie (België). Sinds 1989 hebben beide groepen zich gericht op bacteriële vetzuuranalyse vanuit het oogpunt van identificatie, kwaliteitscontrole, vergelijkende studies en numerieke analyse. Twintig jaar vetzuuranalyse heeft uiteindelijk geleid tot een sterke groei van de FAME databank die, op dit moment, meer dan 71 000 bacteriële vetzuurprofielen bevat. Een databank van deze omvang vormt dan ook een ideaal startpunt voor een uitgebreide data mining en kennisextractie. In deze thesis werken we toe naar een verbeterd identificatiesysteem voor bacteriële soorten door gebruik te maken van intelligente leersystemen of machine learning modellen. Er werd gebruik gemaakt van drie verschillende machine learning modellen: artificiële neurale netwerken, support vector machines en random forests. In een gesuperviseerde setting willen we de soorten van verschillende genera van elkaar onderscheiden. Omwille van de beschikbaarheid van grote hoeveelheid data en uitgebreide expertkennis over de betreffende genera, hebben we gekozen om te werken met de drie genera *Bacillus*, *Paenibacillus* en *Pseudomonas*. Alle standaard vetzuurprofielen van geldig beschreven soorten van deze drie genera werden uit de databank geëxporteerd volgens een manuele procedure. Op basis van deze profielen werden verschillende genus data sets, een genera data set en twee data sets betreffende plantpathogeniciteit samengesteld. In dit laatste geval werd zowel een data set ontworpen met niet-plantpathogene *Pseudomonas* soorten ten opzichte van plantpathogene *Pseudomonas* soorten, als een data set met louter plant pathogene *Pseudomonas* soorten.

Op basis van de samengestelde data sets, werd een typische data analyse uitgevoerd. Gelijkaardig aan reeds uitgevoerde numerieke analyse studies in verschillende genera, werden gemiddelde vetzuurprofielen berekend. Met deze analyse konden genus-gerelateerde, soort-specifieke en stam-specifieke pieken onthuld worden en kon de samenstelling van pieken geëvalueerd worden aan de hand van piekdistributies. Daarnaast werd ook een clustering uitgevoerd van de pieken met en zonder soortclustering. Uit deze analyse kwam heel duidelijk tot uiting dat veel soorten gelijkaardige vetzuurprofielen hebben. Dit werd bevestigd aan de hand van een TaxonGap analyse die duidelijk aangaf dat FAME geen goede taxonomische merker is voor het soortniveau. Finaal werd nog een principale componenten analyse uitgevoerd die aantoonde dat de verschillende vetzuren gecorreleerd zijn. Verder maakten de biplots van de eerste twee principale componenten data duidelijk dat voor elk van de drie genera geldt dat de vetzuurdata overlapt voor een meerderheid van de soorten. Dit was specifiek heel duidelijk waarneembaar voor de verschillende *Pseudomonas* soorten.

Na deze data analyse werd een eerste machine learning experiment opgezet waarbij nage-

gaan werd hoe goed artificiële neurale netwerken in staat zijn om de verschillende *Bacillus* soorten te onderscheiden. Verschillende parameters werden onderzocht en statistisch geanalyseerd. Zo werd onderzocht welke strategie gevolgd moest worden i.v.m. het opstellen van de data set: ongebalanceerd of gebalanceerd. Er werd nagegaan welke validatiestrategie tot de beste resultaten leidt: het gebruik van een aparte validatieset of cross-validatie. Finaal werd ook onderzocht welke neurale netwerk activatiefuncties gekozen moesten worden. Uiteindelijk werd de beste *Bacillus* soortidentificatie bekomen door gebruik te maken van een ongebalanceerde data set en door modellen te valideren met een gestratificeerde cross-validatie. Voor de keuze van een goede activatiefunctie bleek een bijkomende optimalisatie noodzakelijk. In een tweede setting werd het machine learning onderzoek opgeschaald naar drie genera. Twee identificatie strategieën werden onderzocht: een gestratificeerde aanpak waarbij profielen eerst op genus niveau geïdentificeerd worden en vervolgens op soortniveau, en een gegroepeerde aanpak waarbij getracht werd alle soorten van de drie genera uit elkaar te houden. In de eerste aanpak werd gebruik gemaakt van twee type data sets: een genera data set met profielen die enkel geannoteerd werden via de corresponderende genus naam en drie genus data sets waarbij, voor elk genus, de profielen van elke soort geannoteerd werden met genus- en soortnaam. In de tweede aanpak werd slechts één data set gebruikt waarbij alle profielen van alle soorten van de drie genera ingesloten werden en geannoteerd werden met genus- en soortnaam. De drie machine learning technieken artificiële neurale netwerken, support vector machines en random forests werden geëvalueerd in deze twee strategieën. Een betere performantie werd bekomen aan de hand van de random forests techniek via een gestratificeerde aanpak. Een goede soortidentificatie werd bekomen binnen het genus *Paenibacillus*, een relatief goede identificatie binnen het genus *Bacillus* en een minder goede identificatie binnen het genus *Pseudomonas*. Deze bekomen resultaten werden ook getoetst aan de verschillende resultaten en conclusies van de data analyse experimenten en de literatuur. De performantie van de machine learning modellen werd vergeleken met de identificatie bekomen met het commercieel identificatiesysteem Sherlock MIS en voor de soorten aanwezig in beide systemen werd een duidelijk verbeterde identificatie vastgesteld. In vergelijking met het commercieel systeem is het belangrijk te vermelden dat het gebruik van machine learning modellen voor vetzuur-gebaseerde bacteriële soortidentificatie een aantal duidelijke voordelen heeft. Met de maandelijks veranderende bacteriële taxonomie is het noodzakelijk dat identificatiesystemen up-to-date gehouden worden. Machine learning modellen lenen zich daar zeer goed toe door een snelle aanpassing van de desbetreffende data sets en een snelle hertraining van de corresponderende modellen. Daarnaast werden de identificatiemodellen ook getest op onafhankelijke test sets en ook hier werd een verbeterde identificatie bekomen.

Uit de data analyse en de machine learning experimenten bleek duidelijk dat vetzuuranalyse beperkt is in de mogelijkheid tot classificatie tot op soortniveau. Via de ontwikkelde modellen kan echter niet nagegaan worden hoe sterk FAME data soorten van elkaar kan onderscheiden. Om dit verder te analyseren, hebben we een stap verder gezet in het machine learning onderzoek waarin getracht werd om taxonomische en phylogenetische kennis in te bouwen in de classificatiemodellen. Een voor de hand liggende piste hierin is het construeren van een taxonomische of fylogenetische boom en gebruik te maken van binary tree classificatiemodellen. Op

elke node van de boom wordt een binary tree model getraind met als doel de soorten en soortgroepen corresponderend met de twee onderliggende takken van elkaar te onderscheiden. Twee strategieën werden geëvalueerd. In een eerste experiment werd gefocust op het algoritme van de binary tree waarbij aan de hand van de te classificeren data een boom werd opgesteld die vervolgens gebruikt werd als classificatiemethode. In dit experiment werd dus nagegaan hoe een boom geconstrueerd kon gemaakt worden op basis van de FAME data. In een proof-of-concept experiment op basis van 15 *Bacillus* soorten werd aan de hand van divisive clustering van de identificatiescores een vetzuurboom opgesteld. Vervolgens werd deze boom als een binary tree model aangewend voor hiërarchische classificatie van de verschillende bacteriële soorten. Een goede performantie werd hierbij bekomen. Echter, uitbreiding van dit clustering experiment naar de 74 aanwezige soorten was computationeel onhaalbaar. In een tweede strategie werd een fylogenetische boom opgesteld aan de hand van het 16S rRNA gen. Sequenties werden geselecteerd in de gecureerde sequentiedatabank SILVA en de twee algoritmes neighbour-joining en UPGMA werden aangewend voor de constructie van twee fylogenetische bomen. Aan de hand van deze bomen werd het binary tree algoritme opnieuw aangewend als classificatieschema. Classificatie in deze strategie was dus gebaseerd op twee data types. Via de geconstrueerde boom bepalen de 16S rRNA sequenties de verschillende binaire classificatietaken die uitgevoerd worden aan de hand van de FAME data. Aangezien dergelijke methode voorheen nog nergens beschreven en toegepast werd, kan deze methode als vernieuwend beschouwd worden. Omwille van het classificeren van bacteriële soorten op basis van FAME data en gebruik makend van data die evolutionaire verwantschappen beschrijft, hebben we deze techniek 'fylogenetisch leren' genoemd. Vergeleken met een gewone multi-klasse classificatiesysteem, is de classificatie bekomen met deze strategie minder goed. Echter, in deze structuur kunnen we duidelijk het discriminerend vermogen van vetzuurdata analyseren, evalueren, exploiteren en visualiseren. Zo kan, bijvoorbeeld, via een snoeistrategie een geoptimaliseerde classificatiemodel bekomen worden dat aangepast is aan het vermogen van de vetzuurdata om bacteriële soorten te onderscheiden. Verder werd nog een andere belangrijke bevinding bekomen. Soorten die slecht geclassificeerd werden door een gewoon multi-klasse classificatiesysteem werden hoofdzakelijk beter geclassificeerd binnen het onderzochte fylogenetische classificatiesysteem. Daarnaast werd een analyse uitgevoerd van het misclassificatiepad, wat aantoonde dat misclassificatie hoofdzakelijk gebeurde in de eerste ouder-nodes boven de soorten. Hieruit werd onmiddellijk het beperkte vermogen van vetzuurdata om verschillende soorten te onderscheiden opnieuw duidelijk. Deze bevinding ondersteunt nogmaals de noodzaak om dit vermogen in de classificatiemodellen te integreren. Finaal werd getracht om het discriminerend vermogen van vetzuurdata te visualiseren op de geconstrueerde bomen door een statistische analyse van de identificatieresultaten van de verschillende classificatiemodellen getraind op alle nodes van de bomen. Met deze analyse werd onmiddellijk duidelijk welke soorten of soortgroepen niet van elkaar te onderscheiden zijn.

Kort samengevat, met als doel vetzuur-gebaseerde bacteriële soortidentificatie te verbeteren werden in deze studie verschillende modellen voor een computationele identificatie ontwikkeld aan de hand van intelligente leertechnieken. Verschillende data analyse experimenten werden uitgevoerd om een beeld te krijgen over de patronen in de data. Vervolgens werden ver-

schillende machine learning technieken geëvalueerd voor een genus-wijde identificatie van de soorten van drie genera en gebruik makend van verschillende identificatiestrategieën. In vergelijking met de identificatie bekomen met het commercieel systeem Sherlock MIS werd een verbeterde identificatie bekomen voor soorten ingesloten in beide systemen. Ook identificatie van onafhankelijke test sets werd onderzocht. Met de eerste machine learning experimenten kwamen verschillende voordelen duidelijk naar voren: snelle modelconstructie, verbeterde identificatie en, belangrijk, de mogelijkheid tot het up-to-date houden met de huidige bacteriële taxonomie van het ontwikkelde identificatieschema. Een goede basis werd gelegd voor verdere uitbreiding naar meerdere genera. Het is echter belangrijk op te merken dat niet alle bacteriën groeien onder standaardcondities en dat zelfs bepaalde bacteriën helemaal niet cultiveerbaar zijn. In het eerste geval zijn duidelijke afspraken noodzakelijk voor een objectief en betrouwbaar onderzoek. Via een hiërarchische classificatie werd getracht om het discriminerend vermogen van vetzuurdata voor bacteriële soorten te integreren in de computationele modellen. Twee strategieën werden onderzocht en de combinatie van vetzuurdata met 16S rRNA sequenties is hierbij veelbelovend. Met deze methode is het mogelijk om het discriminerend vermogen van vetzuurdata snel te evalueren aan de hand van data met een discriminerend vermogen voor de verschillende bacteriële soorten. Dit werd initieel reeds onderzocht via een statistische analyse van de identificatieresultaten corresponderend met elke node. Verder onderzoek is echter aangewezen om een geoptimaliseerd hiërarchisch classificatiesysteem op te zetten.

Als finaal onderzoeksthema werd de mogelijkheid onderzocht om een oplossing te bieden voor de beperkingen opgelegd door de FAME databank. In het onderzoek werd onmiddellijk duidelijk dat de machine learning experimenten gelimiteerd waren door de afwezigheid van een voldoende hoeveelheid data voor elke bacteriële soort. Dit betreft niet enkel het aantal profielen per soort maar ook het aantal stammen per soort, en het aantal soorten per genus. Daarom werd deze thesis gefinaliseerd met het oprichten van een publieke vetzuurdatabank FAME-bank.net voor het online delen en opzoeken van bacteriële vetzuurdata. Op deze manier werd een verlengstuk gecreëerd dat toelaat om het uitgevoerde onderzoek in de toekomst uit te breiden naar meerdere bacteriële genera, soorten, stammen en vetzuurprofielen.

