Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Microbiological Methods 66 (2006) 410–433
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

Mining fatty acid databases for detection of novel compounds in aerobic bacteria

Peter Dawyndt [a,b,*], Marc Vancanneyt [c], Cindy Snauwaert [c], Bernard De Baets [b], Hans De Meyer [d], Jean Swings [b]

[a] Laboratory of Microbiology, Ghent University, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium
[b] Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Gent, Belgium
[c] BCCM™/LMG Bacteria Collection, Ghent University, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium
[d] Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 (S9), B-9000 Gent, Belgium

Received 18 November 2005; received in revised form 11 January 2006; accepted 23 January 2006
Available online 7 March 2006

Abstract

This study examines how the discriminatory power of an automated bacterial whole-cell fatty acid identification system can be significantly enhanced by exploring the vast amounts of information accumulated during 15 years of routine gas chromatographic analysis of the fatty acid content of aerobic bacteria. Construction of a global peak occurrence histogram based upon a large fatty acid database is shown to serve as a highly informative tool for assessing the delineation of the naming windows used during the automatic recognition of fatty acid compounds. Along the lines of this data mining application, it is suggested that several naming windows of the Sherlock MIS TSBA50 peak naming method may need to be re-evaluated in order to fit more closely with the bulk of observed fatty acid profiles. At the same time, the global peak occurrence histogram has put forward the delineation of 32 new peak naming windows, accounting for a 26% increase in the total number of fatty acid features taken into account for bacterial identification. By scrutinizing the relationships between the newly delineated naming windows and the many taxonomic units covered within a proprietary fatty acid database, all new naming windows were proven to correspond with stable features of some specific groups of microorganisms. This latter analysis clearly underscores the impact of incorporating the new fatty acid compounds for improving the resolution of the bacterial identification system and endorses the applicability of knowledge discovery in databases within the field of microbiology.

Keywords: Bacterial identification; Data mining; Fatty acid analysis; Feature extraction; Knowledge discovery in databases

1. Introduction

Bacterial fatty acids, unlike many other phenotypic characteristics, are genetically highly conserved, owing their essential role in cell structure and function. Ever since the introduction of gas chromatographic analysis of cellular fatty acids by Abel et al. (Abel et al., 1963), this technique has been frequently applied in various taxonomic studies (Vauterin et al., 1996). Numerous

* Corresponding author. Laboratory of Microbiology, Department of Biochemistry, Physiology and Microbiology, Ghent University, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium. Tel.: +32 9 264 5132; fax: +32 9 264 5092.
E-mail addresses: Peter.Dawyndt@UGent.be (P. Dawyndt), Marc.Vancanneyt@UGent.be (M. Vancanneyt), Cindy.Snauwaert@UGent.be (C. Snauwaert), Bernard.DeBaets@UGent.be (B. De Baets), Hans.DeMeyer@UGent.be (H. De Meyer), Jean.Swings@UGent.be (J. Swings).

0167-7012/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2006.01.008

scientific papers have used the fatty acids between 9 and 20 carbons in length to characterize genera and species of bacteria, e.g. nonfermentative Gram-negative microorganisms (Sasser, 1990). With the many improvements in automated calibration and interpretation of the chromatographic profiles, reproducible fatty acid profiles nowadays can be generated rapidly, provided that strains are grown under specified standardized conditions (Osterhout et al., 1998), and the identification of microorganisms on the basis of their cellular fatty acid composition has become a standard of polyphasic taxonomy for over a decade (Vandamme et al., 1996).

Apart from commercial identification libraries that offer automated taxonomic positioning based on fatty acid profile screening of vast amounts of well-characterized microorganisms that cover a broad range of the bacterial diversity, several microbiological research institutes are compiling vast databases on the fatty acid content of freshly isolated organisms. In the long term, these in-house data repositories turn out to become valuable identification assets as they encompass the specialized niches of the research centers. This is exemplified by the whole-cell fatty acid methyl ester (FAME) database that is collectively established by the Laboratory of Microbiology at the Ghent University and the BCCM™/LMG Bacteria Collection using the Sherlock Microbial Identification System (MIS; Microbial ID, Inc. (MIDI), Newark, Delaware, USA). We will refer to this proprietary FAME database as the BAME@LMG database throughout the rest of this paper, where BAME is used as acronym for the bacterial fatty acid methyl ester profiles captured within the database. For the construction of the BAME@LMG database the Sherlock MIS system was chosen merely because of its library generation capability, the large number of known compounds recognized by the system's peak naming tables, the ability to compare a large number of strains over a period of time, and the routine use of this system for bacterial analysis in many laboratories (Tighe et al., 2000). The chromatographic fatty acid peaks are automatically named and quantified by the system, and the wealth of information contained in these compounds can be used for bacterial identification by considering not only the presence or absence of each fatty acid, but also by using the data in a quantitative fashion (Sasser, 1990).

The majority of taxa covered within the BAME@LMG database reflects the general interest of both laboratories in the taxonomy of aerobic bacteria, primarily isolated from environmental as well as some clinical samples. Most strains were subjected to fatty acid analysis in the framework of taxonomic research projects, during the initial screening stages for polyphasic classification (Vandamme et al., 1996) of large sets of samples from the genera *Aeromonas* (Huys et al., 1994, 1995, 1996), *Arcobacter* (Vandamme et al., 1992), *Bordetella* (Vancanneyt et al., 1995), *Flavobacterium* (Bernardet et al., 1996), *Pseudomonas* (Vancanneyt et al., 1996), *Rhizobium* (Tighe et al., 2000), *Streptomyces* (Mergaert et al., 1994), and *Xanthomonas* (Vauterin et al., 1991, 1992, 1996; Yang et al., 1993), among many others. Chromatographic analysis of the fatty acid composition of a large group of strains isolated from Arctic and Antarctic waters revealed members of new and old taxa related to the genera *Alteromonas, Cytophaga, Glaciecola, Halomonas, Pseudoalteromonas, Rhizobium, Rhodococcus, Shewanella* and *Sulfitobacter* (Mergaert et al., 2001; Van Trappen et al., 2002). It was found that in order to maintain fluidity of the membranes under low temperature conditions, polar isolates are characterized by high amounts of unsaturated fatty acids (Mergaert et al., 2001). Microbial identification based on fatty acid analysis was also implemented for investigating the biodiversity of heterotrophic bacteria colonizing mural paintings that showed visual deterioration by microorganisms, uncovering the dominant presence of *Arthrobacter, Bacillus, Paenibacillus, Micrococcus* and *Staphylococcus* species, but also nocardioform actinomycetes and Gram-negative bacteria (Heyrman et al., 1999). Apart from its use as a fast screening technique, knowledge of the cellular fatty acid composition also plays an important role in the description of new bacterial taxa in many scientific publications. As an illustration, we refer to the descriptions of *Arcobacter butzleri* and *A. skirrowii* (Vandamme et al., 1992), *Brachybacterium fresconis* and *B. sacelli* (Heyrman et al., 2002), *Flavobacterium hydatis* (Bernardet et al., 1996) and *Leeuwenhoekiella aequorea* (Nedashkovskaya et al., 2004). Given the relatively low operating cost, high throughput capabilities and long-term reproducibility of gas chromatographic analysis, the BCCM™/LMG Bacteria Collection also produces fatty acid profiles as part of its implementation of a total data quality management system to safeguard the authenticity of the strains in its holdings.

In the present paper we intend to explore how the information content of the large BAME@LMG database can be fully exploited to improve to overall accuracy of bacterial identification on the basis of fatty acid analysis, thereby taking into account both the chromatographic peaks that were named as known fatty acids by the Sherlock MIS system, as well as those that remained unidentified by the system. This exploratory data analysis process will be gradually built up in a number of consecutive steps, each one of them delving into the problem with more detail. We start with an investigation of some of the qualitative properties related to the distribution of fatty acid compounds in