Applied and Environmental Microbiology, Feb. 2011, p. 896–905 Vol. 77, No. 3
0099-2240/11/$12.00 doi:10.1128/AEM.01686-10

Determination of Serotypes of Shiga Toxin-Producing *Escherichia coli* Isolates by Intact Cell Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry▽

Axel Karger,[1] Mario Ziller,[3] Barbara Bettin,[1] Birgit Mintel,[2] Susann Schares,[2] and Lutz Geue[2]*

Institute of Molecular Biology, Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Südufer 10, D-17493 Greifswald-Insel Riems, Germany[1]; Institute of Epidemiology, Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Seestrasse 55, D-16868 Wusterhausen, Germany[2]; and Working Group Biomathematics, Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Südufer 10, D-17493 Greifswald-Insel Riems, Germany[3]

Received 14 July 2010/Accepted 17 November 2010

Shiga toxin-producing *Escherichia coli* (STEC) isolates representing the serotypes O165:H25, O26:H11/H32, and O156:H25 were analyzed by matrix-assisted laser desorption/ionization (MALDI) mass spectra of whole cells, a procedure also known as intact cell mass spectrometry (ICMS or IC-MALDI MS) or MALDI-typing. We demonstrate that within the given species the three serotypes can be well discriminated by ICMS. Conditions for the construction of serotype-specific prototypic mass spectra were systematically optimized by filtering out masses that do not contribute to the discrimination of the serotypes. Binary distances between prototypic spectra and sample spectra were used to determine serotypes of unknown samples. With parameters optimized, only 0.7% of the assignments were incorrect compared to 31% when distances were calculated from alignments of unfiltered mass spectra. Within the different serotypes, clusters of genetically related *E. coli* most probably originating from single clones could be distinguished by restriction fragment length polymorphism analysis. Since ICMS did not reproduce these clusters, we conclude that the power of ICMS is just sufficient to discriminate *E. coli* serotypes under certain conditions but fails for the differentiation of *E. coli* below this level.

Shiga toxin-producing *Escherichia coli* (STEC) strains comprise a group of zoonotic enteric pathogens (20). In humans, infections with some STEC serotypes may result in hemorrhagic or nonhemorrhagic diarrhea, which can be complicated by the hemolytic-uremic syndrome (HUS) (12). These STEC strains are also designated enterohemorrhagic *Escherichia coli* (EHEC). Consequently, EHEC represent a subgroup of STEC with a high pathogenic potential for humans. Although *E. coli* O157:H7 is the most common EHEC serotype implicated in HUS, other serotypes can also cause this complication. Non-O157:H7 EHEC are present in ca. 50% of stool cultures from German HUS patients (1, 19). However, STEC strains that cause human infection belong to a large number of *E. coli* serotypes, including serotypes O26:H11, O156:H25, and O165:H25. Certainly, human disease is far more frequently caused by STEC O26:H11 than by STEC of serotypes O156:H25 or O165:H25. Ruminants, especially cattle, are considered the primary reservoir for human infection with EHEC. Therefore, the epidemiological situation in beef herds regarding isolates of these serotypes was examined by restriction fragment length polymorphism (RFLP) analysis (5, 6, 8). Due to the high resolution of RFLP-based cluster analysis, clusters of isolates within the three serotypes had been defined in the course of these studies that had most probably resulted from the expansion of single clones.

The introduction of mass spectrometric (MS) identification and classification of bacteria (11, 21, 25) has revolutionized the species determination of bacteria (27; see reference 9 for a recent review). The technique has also been expanded to other microorganisms such as fungi (17, 18) and protozoans (30) and to tissue cultures (13). Most publications emphasize the speed, high-throughput capabilities, and cost efficiency of the procedure (27). In some cases, intact cell mass spectrometry (ICMS) showed better correlation to genetic markers than conventional morphological classification (17). The technique or variants, including a digestion step ("shotgun mass mapping") (26), have been successfully used for subspecies-level classification but also for the source tracking of bacteria in environmental samples (28). Commercial solutions for the identification of bacteria by matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) MS are available. The characteristic features of ICMS are the minimalistic sample preparation procedures of whole cells, spectrum acquisition in the low mass range, and analysis based on comparison of sample spectra with reference spectra from authentic samples without the need for other *a priori* information about the sample. By statistical approaches, the similarities between mass spectra can be exploited for the identification of microorganisms but also for phylogenetic analysis.

The aim of the present study was to explore the potential of ICMS for the characterization of EHEC strains below the species level and to compare its power to genetic typing by

* Corresponding author. Mailing address: Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Institute of Epidemiology, Seestrasse 55, D-16868 Wusterhausen, Germany. Phone: 49 33979 80189. Fax: 49 33979 80222. E-mail: Lutz.Geue@fli.bund.de.

▽ Published ahead of print on 29 November 2010.

RFLP analysis. To this end, both techniques were applied to a set of 94 *E. coli* serotype O26:H11/H32, O156:H25, and O165:25 isolates. The construction of phylogenetic trees based on raw sample spectra suggested that the serotypes are indeed determinants for ICMS-based characterization of *E. coli* isolates, since three major groups evolved which, with a number of exceptions, represented the three serotypes. Therefore, a strategy to improve the discrimination power of ICMS for the phylogenetic analysis of closely related organisms was developed and a procedure for the determination of the serotypes of unknown samples was devised and evaluated.

MATERIALS AND METHODS

***E. coli* strains.** A total of 94 *E. coli* isolates of three O serogroups (O165:H25, $n = 30$; O26:H11, $n = 45$; O26:H32, $n = 1$; O156:H25, $n = 18$) were included in the present study (Table 1). All strains were isolated during a monitoring program in four German cattle farms (7) and were characterized as EHEC in previous studies (5, 6, 8). Serotypes were determined by the National Reference Centre for Salmonella and other Enterics (Robert-Koch-Institut Wernigerode, Germany) as published elsewhere (5, 6, 8). In addition, O serogroups were characterized by MboII restriction endonuclease digests of amplified O-antigen gene clusters (*rfb*-RFLP) as described by Coimbra et al. (2).

ICMS. Samples were essentially prepared as described previously (25). Specimens from single colonies were suspended in 300 μl of water, precipitated by addition of 900 μl of ethanol (98% [vol/vol]) and sedimented for 5 min at 10,000 × *g*. The supernatant was carefully removed, and the sediment resuspended in 50 μl of 70% (vol/vol) formic acid. After mixing with 50 μl of acetonitrile, the suspension was centrifuged as described above, and the supernatant transferred to a fresh tube. A 1.5-μl portion of the extract was spotted to a steel MALDI target plate and allowed to dry at ambient temperature. Finally, the dried extract was overlaid with 1.5 μl of a saturated solution of α-cyano-4-hydroxycinnamic acid in 50% acetonitrile–2.5% trifluoroacetic acid as matrix and was again allowed to dry. A custom-made database of reference spectra was constructed by using BioTyper software (version 1.1; Bruker Daltonics, Bremen, Germany) according to the guidelines of the manufacturer. Each sample was spotted onto six target spots of a steel MALDI target. Spectra were acquired with an Ultraflex I instrument (Bruker) in the linear positive mode in the range of 2,000 to 15,000 Da. Acceleration voltage was 25 kV, and the instrument was calibrated in the range of 4,364 to 10,299 Da with reference masses of an extract of a *E. coli* DH5-α strain prepared as described previously (25). Four single spectra with 500 shots each were acquired from each spot, and a reference spectrum was calculated from the 24 single spectra. Reference spectra contained the usual parameters of mass spectra (peak mass and intensity) and additional information on the reproducibility of the mass peaks, i.e., the frequency of occurrence of every peak in the underlying 24 single spectra. Reference spectra were generated within the mass range of 2,500 to 15,000 Da using the following default parameter settings in the BioTyper software: compression of the spectrum data by a factor of 10, baseline smoothing by the Savitsky-Golay algorithm (25-Da frame size), baseline correction by two runs of the multipolygon algorithm, and a peak search by spectrum differentiation. The number of peaks was limited to 100 per reference spectrum, and all peaks of a reference spectrum were normalized to the most intense peak with an intensity of 1.0. The minimum frequency of occurrence within the 24 single spectra was set to 50% for every mass. Peak lists of reference spectra were exported for further evaluation.

Transfer of data into statistical programming language R. Peak lists of the reference spectra generated by the BioTyper software were converted into a format compatible with the R-package caMassClass which was originally developed for the classification of surface-enhanced laser desorption/ionization (SELDI) mass spectra (29) for further statistical analysis in R version 2.10.1. (22) available at the R-project homepage (http://www.r-project.org). Peak lists were aligned by the msc.peaks.align command of caMassClass and transformed into a binary mass table where rows represented all unique masses of the aligned spectra set and every column represented the spectrum of one sample. The size of the mass ranges defining a unique peak in the alignment, designated as bin size, was restricted to a minimum of 500 ppm and a maximum of 2,000 ppm. Among other features, the algorithm of the msc.peaks.align command minimizes the bin size in the given range, maximizes the space between bins, and ensures that no two peaks of the same spectrum are in the same bin. The presence of the respective mass in the spectrum of a sample was marked with "1", and absence was marked with "0", i.e., all mass intensities were removed. This table was the basis for the calculation of binary distances (R-routine "dist", parameter "binary" for the distance measure) which were used for the construction of phylogenetic trees, Sammon maps (24), and k-means cluster analysis using the R-routines "hclust" (parameter "ward" for the agglomeration method [4]), "sammon" (used with default settings), and "kmeans" (three initial cluster centers, maximum of 100 iterations, Hartigan-Wong algorithm [10]).

Calculation of serospecific prototype spectra and assignment of serotypes. For the calculation of serotype-specific prototype spectra and the assignment of serotypes to unknown samples, a program in R was developed. Details concerning the optimization of the parameters are given in the results section. For the Fisher test, the R-routine "fisher.test" was used.

RFLP analysis by PFGE. Genomic DNA for contour-clamped homogeneous electric field pulsed-field gel electrophoresis (CHEF-PFGE) was prepared as previously described (5, 8, 16). Slices of DNA agarose plugs were equilibrated in the respective restriction endonuclease buffer and then digested for 4 h with XbaI, NotI, BlnI (AvrII), or SpeI (New England Biolabs, GmbH, Frankfurt/M, Germany). The resulting fragments were separated in 1.0% agarose gels (Biozyme Gold agarose; Biozyme GmbH, Germany) in 0.5× Tris-borate-EDTA at 10°C in a CHEF Mapper XA system (Bio-Rad Laboratories GmbH, Munich, Germany). The pulse times for XbaI and NotI digests were increased from 5 to 50 s (gradient, 6 V/cm) for 25 h at a constant angel of 120°. The switch time values for BlnI and SpeI were set by using the Auto Algorithm function of the CHEF Mapper XA to separate fragments in the range of 50 to 450 kb (BlnI) or 30 to 350 kb (SpeI), respectively. After electrophoresis, the gels were stained with 500 ml of ethidium bromide solution (50 μg/ml), and the banding patterns recorded under UV illumination. Interpretation of the RFLP patterns was performed by visual inspection and computer analysis (BioNumerics software v6.0; Applied Maths NV, Sint-Martens-Latem, Belgium). All fragments larger than 45 kb were included in the clonal analysis of the isolates. Dendrograms based on RFLP patterns of all four restriction endonucleases were constructed on the basis of a binary table in which the presence or absence (1/0) of restriction fragments was recorded. The "Ward" algorithm of the R-routine "hclust" was used.

RESULTS

Cluster analysis based on ICMS. Initially, spectra from 94 samples were analyzed. The resulting dendrogram showed three major branches, two of which, representing the O156:H25 and O165:H25 serotypes, were free of other serotypes, whereas eight O165:H25 serotype strains clustered with the O26:H11/H32 samples (Fig. 1, top panel). Therefore, the *E. coli* O serogroups of all isolates were reanalyzed by MboII restriction endonuclease digests of amplified O-antigen gene clusters (*rfb*-RFLP) as described by Coimbra et al. (2). The patterns of six questionable samples, namely, WH-2/24/007-6, WH-02/09/010-1, WH-02/09/010-2, WH-02/18/011-1 (all presumably serotype O165:H25), and WH-01/27/017-5, and WH-01/26/001-2 (both presumably serotype O26:H11) differed markedly from the typical O165 and O26 patterns described in the *rfb*-RFLP database (2) so that a classification of these was equivocal, most probably as a result of recombinational events in the *rfb* regions of STEC that have been reported previously (15). After removal of these six questionable samples from the data set, the separation of the three serotypes into three branches of the hierarchical dendrogram (Fig. 1, bottom panel) was mutually exclusive. The distance matrix of the remaining 88 isolates that was calculated from the alignment of the mass spectra was used for a k-means cluster analysis. The three resulting clusters were almost congruent with the serotypes of the isolates with the exception of WH-01/29/002-4 (O26:H11) (Fig. 2A, arrow), indicating that the serotypes are significant determinants of the ICMS mass patterns. Although the reconstruction of serotype groups by ICMS was successful, the analysis of the binary mass table suggested that careful selection of masses used for further calculations might improve

TABLE 1. *E. coli* isolates and some EHEC-associated virulence factors

Isolate no.	Isolate designation[a]	Host	Farm/group	Serotype	*stx* subtype	*eae* subtype	Presence of:		
							EHEC *hlyA*	*katP* gene	*espP* gene
1	WH-01/02/003-1	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
2	WH-01/02/003-2	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
3	WH-01/02/003-5	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
4	WH-01/02/003-6	Cattle	Farm A	O26:H11	stx_1	β	–	+	+
5	WH-01/02/003-7	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
6	WH-01/02/003-8	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
7	WH-01/02/003-9	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
8	WH-01/02/003-10	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
9	WH-01/06/002-1	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
10	WH-01/06/002-2	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
11	WH-01/06/002-3	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
12	WH-01/08/002-2	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
13	WH-01/09/016-2	Cattle	Farm A	O26:H32	stx_1/stx_2	–	–	–	–
14	WH-01/26/001-2	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
15	WH-01/26/001-5	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
16	WH-01/26/001-6	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
17	WH-01/26/002-5	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
18	WH-01/26/002-8	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
19	WH-01/26/002-9	Cattle	Farm A	O26:H11	stx_1	β	+	–	–
20	WH-01/26/002-10	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
21	WH-01/27/005-6	Cattle	Farm A	O26:H11	stx_1	β	–	–	–
22	WH-01/27/009-1	Cattle	Farm A	O26:H11	stx_1/stx_2	β	+	+	+
23	WH-01/27/009-2	Cattle	Farm A	O26:H11	stx_1/stx_2	β	+	+	+
24	WH-01/27/009-9	Cattle	Farm A	O26:H11	stx_1/stx_2	β	+	+	+
25	WH-01/27/014-3	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
26	WH-01/27/014-4	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
27	WH-01/27/014-5	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
28	WH-01/27/017-1	Cattle	Farm A	O26:H11	stx_1/stx_2	β	+	+	+
29	WH-01/27/017-5	Cattle	Farm A	O26:H11	stx_1/stx_2	β	+	+	+
30	WH-01/27/017-6	Cattle	Farm A	O26:H11	stx_1/stx_2	β	+	+	+
31	WH-01/27/017-7	Cattle	Farm A	O26:H11	stx_1/stx_2	β	+	+	+
32	WH-01/27/017-10	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
33	WH-01/29/002-1	Cattle	Farm A	O26:H11	stx_1	β	+	–	–
34	WH-01/29/002-3	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
35	WH-01/29/002-4	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
36	WH-01/29/002-5	Cattle	Farm A	O26:H11	stx_1	β	–	–	+
37	WH-01/29/010-1	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
38	WH-01/29/010-2	Cattle	Farm A	O26:H11	stx_1	β	+	–	–
39	WH-01/29/010-3	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
40	WH-01/29/013-4	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
41	WH-01/29/013-7	Cattle	Farm A	O26:H11	stx_1	β	+	+	+
42	WH-03/14/004-8	Cattle	Farm D, group 1	O26:H11	stx_1	β	–	+	+
43	WH-04/07/001-6	Cattle	Farm C	O26:H11	stx_1/stx_2	β	+	+	+
44	WH-04/22/001-1	Cattle	Farm C	O26:H11	stx_1	β	+	+	+
45	WH-04/22/001-4	Cattle	Farm C	O26:H11	stx_1	β	+	+	+
46	WH-04/22/001-5	Cattle	Farm C	O26:H11	stx_1	β	+	+	+
47	WH-02/04/017-9	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
48	WH-02/23/021-2	Cattle	Farm B	O156:H25	stx_1	ζ	+	+	+
49	WH-02/23/021-3	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
50	WH-02/23/021-5	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
51	WH-02/23/021-9	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
52	WH-02/23/021-10	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
53	WH-02/25/010-10	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
54	WH-02/25/019-1	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
55	WH-02/25/024-3	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
56	WH-02/25/024-5	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
57	WH-02/26/008-10	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
58	WH-02/28//018-1	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
59	WH-02/28//018-3	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+
60	WH-02/28//018-4	Cattle	Farm B	O156:H25	stx_1	ζ	+	–	+

Continued on following page

TABLE 1—*Continued*

Isolate no.	Isolate designation[a]	Host	Farm/group	Serotype	*stx* subtype	*eae* subtype	Presence of:		
							EHEC *hlyA*	*katP* gene	*espP* gene
61	WH-03/12/016-2	Cattle	Farm D, group 1	O156:H25	stx_1	ζ	+	–	+
62	WH-03/12/016-8	Cattle	Farm D, group 1	O156:H25	stx_1	ζ	+	–	+
63	WH-04/25/005-1	Cattle	Farm C	O156:H25	stx_1	ζ	+	–	+
64	WH-05/25/004-1	Cattle	Farm D, group 2	O156:H25	stx_1	ζ	+	–	+
65	WH-02/09/010-1	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
66	WH-02/09/010-2	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
67	WH-02/17/009-9	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
68	WH-02/18/011-1	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
69	WH-02/18/011-6	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
70	WH-02/19/013-8	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
71	WH-02/19/013-10	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
72	WH-02/24/007-3	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
73	WH-02/24/007-4	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
74	WH-02/24/007-5	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
75	WH-02/24/007-6	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
76	WH-02/24/007-7	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
77	WH-02/24/007-10	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
78	WH-02/25/007-1	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
79	WH-02/25/007-4	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
80	WH-02/25/007-5	Cattle	Farm B	O165:H25	stx_2	ε	+	–	–
81	WH-02/25/007-6	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
82	WH-02/25/007-8	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
83	WH-02/25/007-9	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
84	WH-02/25/007-10	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
85	WH-02/25/008-1	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
86	WH-02/25/008-3	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
87	WH-02/26/006-1	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
88	WH-02/26/006-2	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
89	WH-02/26/006-3	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
90	WH-02/26/006-4	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
91	WH-02/26/006-6	Cattle	Farm B	O165:H25	stx_2	ε	+	+	–
92	WH-02/26/006-7	Cattle	Farm B	O165:H25	stx_2	ε	+	–	–
93	WH-02/26/006-8	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+
94	WH-02/26/006-10	Cattle	Farm B	O165:H25	stx_2	ε	+	+	+

[a] Sample designations were constructed according to the following pattern: laboratory name (WH for all samples)-number of farm/number of animal/sampling day-isolate number.

the discriminatory power and would allow to assign serotypes to unknown samples. Since all samples represented the same species, *E. coli*, mass patterns of all spectra were very similar. In an alignment of all 276 masses that were present in the spectra of the 88 samples, 17 were present in all samples; the frequencies of 27 masses were greater than 90% and those of 46 masses were greater than 50%. Therefore, parameters for the selection of serotype-specific masses from the alignment were developed and optimized. A procedure was devised to assign serotypes to unknown samples. The percentage of correct assignments served as target parameter for the optimization of the parameters for mass selection.

Determination of serotypes of unknown samples. The strategy for the optimization of the parameters for the determination of serotypes of unknown samples followed five steps. (i) Sample spectra from every serotype group were randomly divided into a "query" group containing one, two, and four representatives of the O156, O165, and O26 serotypes, respectively, and the "reference" group with the rest of the samples. A peak alignment was calculated from the reference samples, and a binary mass table was calculated. (ii) For every mass, the frequency of its occurrence in each of the three groups was calculated, the variation of the frequency between the groups was tested by the Fisher test, and the *P* values were calculated. Prototype spectra were deduced from this mass table with different parameters sets defining a maximum *P* value and a minimum frequency of occurrence for every mass of the alignment. Masses qualified for the prototype spectrum of a serotype group, if its frequency in the respective group exceeded the required minimum frequency and its *P* value was equal or fell below the specified maximum *P* value. All masses that were not present at least in one of the prototype spectra were removed from the alignment table. (iii) The "query" sample spectra, which were not part of the reference spectra used for the construction of the alignment, were added to the mass table using the mass ranges (bins) defined for every mass of the alignment. (iv) A distance matrix was calculated for the prototype and the "query" spectra. (v) Every sample of the "query" set was assigned to the serotype of the prototype spectrum with the closest distance in the distance matrix. To

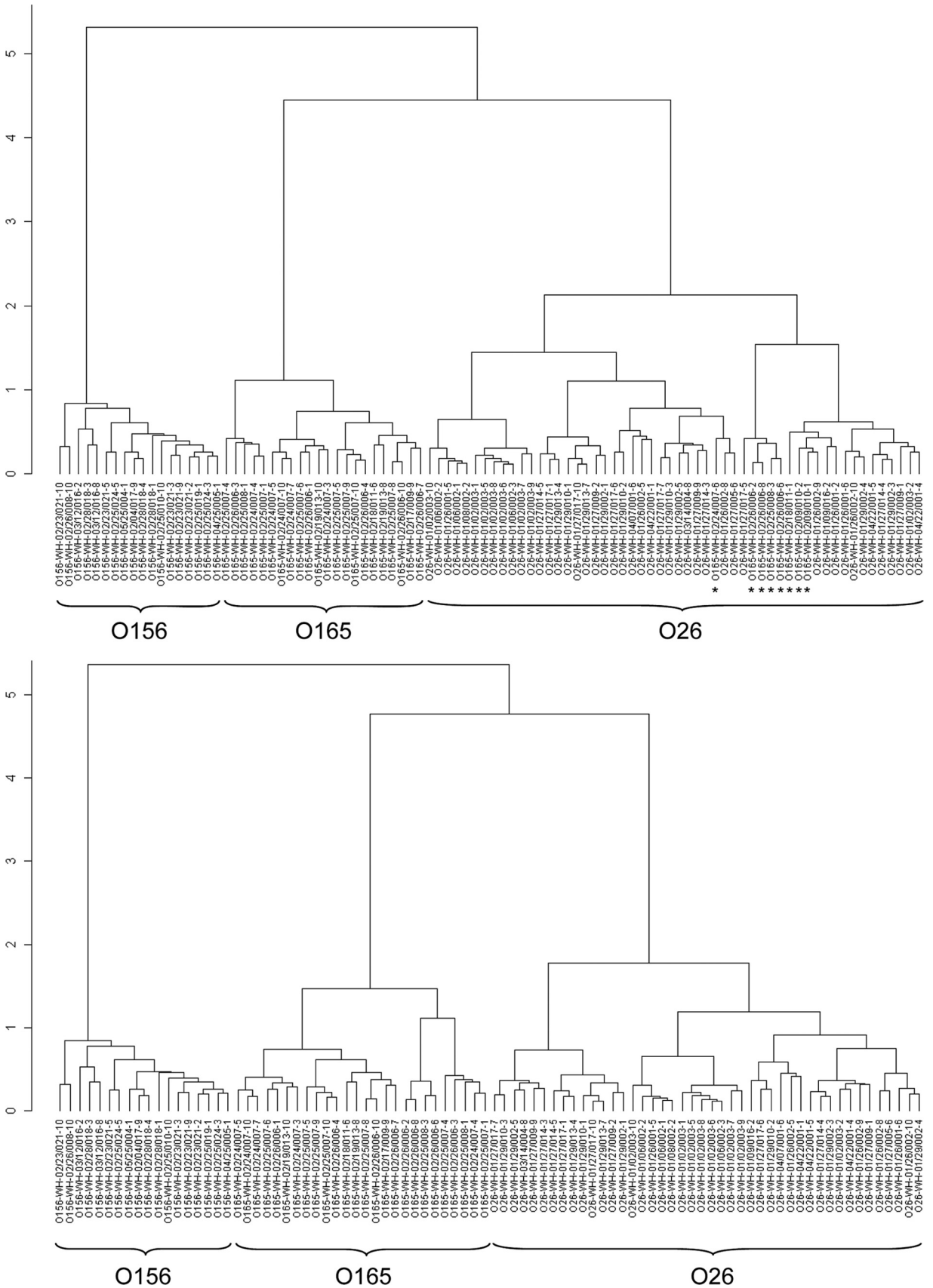

FIG. 1. Phylogenetic trees were constructed from mass spectra of 94 *E. coli* isolates representing three different serotypes (top panel). Of the eight samples that were found not to cluster with their own serotype (asterisks), six were found to be mistyped and were excluded from further analysis. Construction of the dendrogram from the spectra of the remaining 88 samples (bottom panel) resulted in a tree with three major branches representing the three different serotypes.

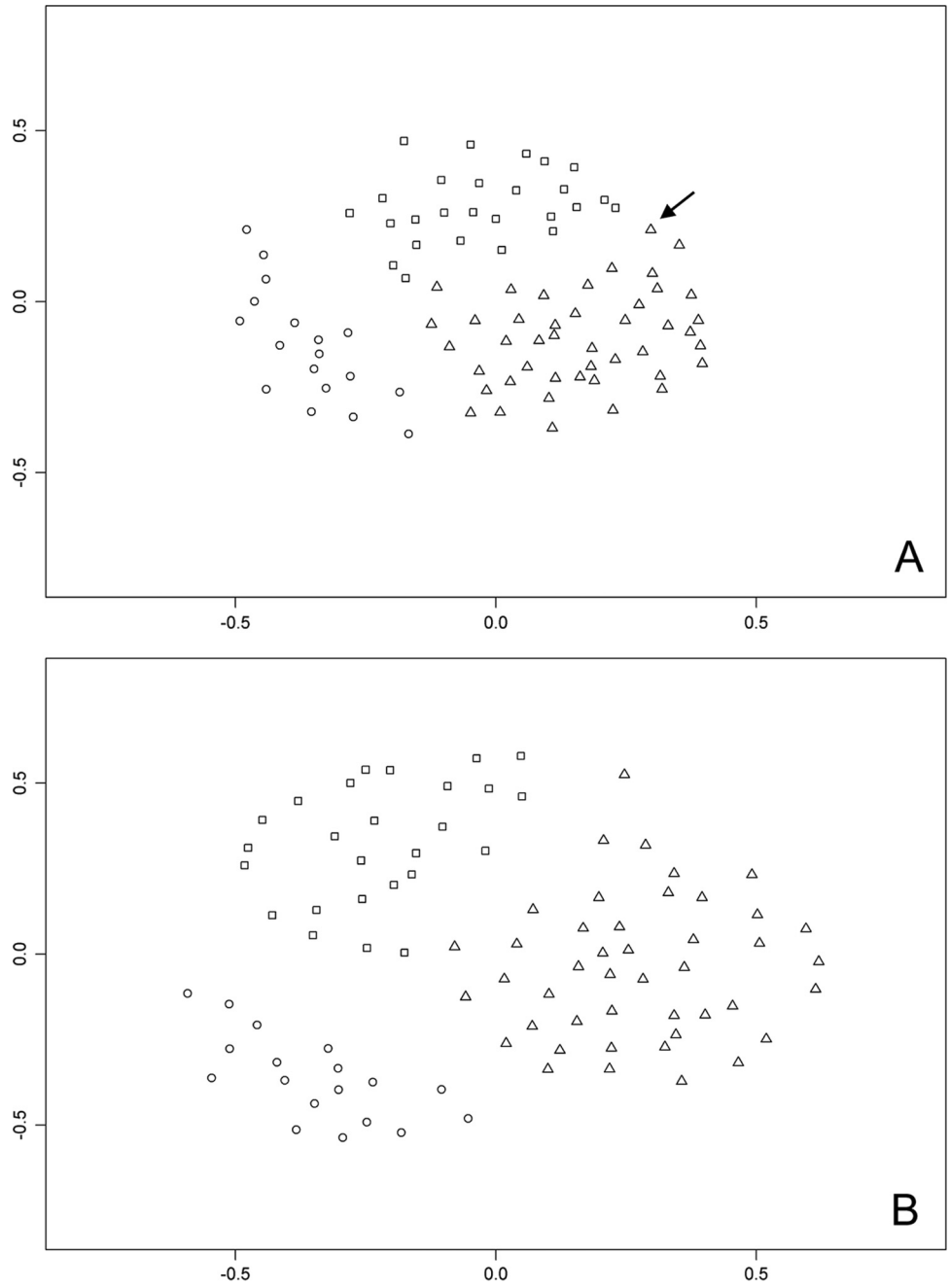

FIG. 2. Sammon representation of the binary distance between the samples. k-means cluster analysis resulted in three clusters which, with one exception (arrow), were congruent with the three serotypes (○, O156:H25; △, O26:H11/H32; □, O165:H25). The plot in panel A was calculated from the complete mass table, whereas the plot in panel B was calculated from the condensed mass table with the parameters for optimal serotype assignment ($P \leq 0.1$, frequencies > 0.3).

ensure that every sample was part of the "query" set for a number of times, the procedure was repeated 50 times. Combinations of maximum P values and minimum frequencies were systematically varied within the desired limits and the number of misidentifications recorded (Table 2).

Under optimized conditions ($P \leq 0.1$, minimal frequency $>$ 0.3), incorrect serotype assignments were obtained only for two isolates: WH-02/26/006-2 (O165:H25), which was assigned to the O26 serotype in one of two tests, and WH-01/27/005-6 (O26:H11), which was classified as O165 in two of three tests.

For the unchanged mass table ($P \leq 1$; frequency > 0), the optimized parameter set and a parameter set with even more stringent conditions ($P \leq 0.1$, frequency > 0.6) quality curves for the assignment of serotypes were calculated (Fig. 3). Separation of O156:H25 from the other two serotypes was still improved under the most stringent conditions (Fig. 3, right

TABLE 2. Incorrect serotype assignments for different parameter sets[a]

P	% incorrect serotype assignments with a minimal frequency of:									
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	31	8.7	3.3	1	1.7	3.3	2.7	6.7	5.3	3
0.5	12.7	5.7	2.3	1	2	3.3	3	5.7	5.3	3.7
0.4	13	5	2.3	0.7	2	3.3	2.7	5.7	6	3
0.3	14	5	2.3	0.7	2	3.3	2.7	6.3	5.3	3
0.2	4.3	5	2.3	0.7	2	3.3	2.7	6.3	4.7	3.3
0.1	5	4.3	2	0.7	2	3.3	3	6.3	4.7	3.7
0.05	5	4.3	1.3	1.3	2.3	3.3	3	6	4.7	3.7
0.02	5.3	3	0.7	1	2	3.3	2.7	4.7	3.7	2.7
0.01	5.3	3.3	3.0	0.7	1.7	2.7	3.3	4	3.7	2.7
0.001	6	4.3	3.0	0.7	2.3	3.7	3	3.7	4.3	3.7

[a] The minimum frequency of occurrence and the maximum *P* value of the Fisher test were used as criteria for the selection of masses that were used for the calculation of prototype spectra. Serotypes were then assigned by comparison of sample spectra with the prototype spectra. The table lists the percentages of incorrect serotype assignments under the conditions applied.

panels), albeit at the cost of more misidentifications for the other serotypes. The Sammon representation of all 88 samples (Fig. 2B) under the optimized conditions ($P \leq 0.1$, minimal frequency > 0.3) showed a slightly enhanced separation of the three serotype groups which was also reflected in an increased quotient of the distances of the group means and the mean distances of the group members to their group mean (Table 3) in the distance matrix.

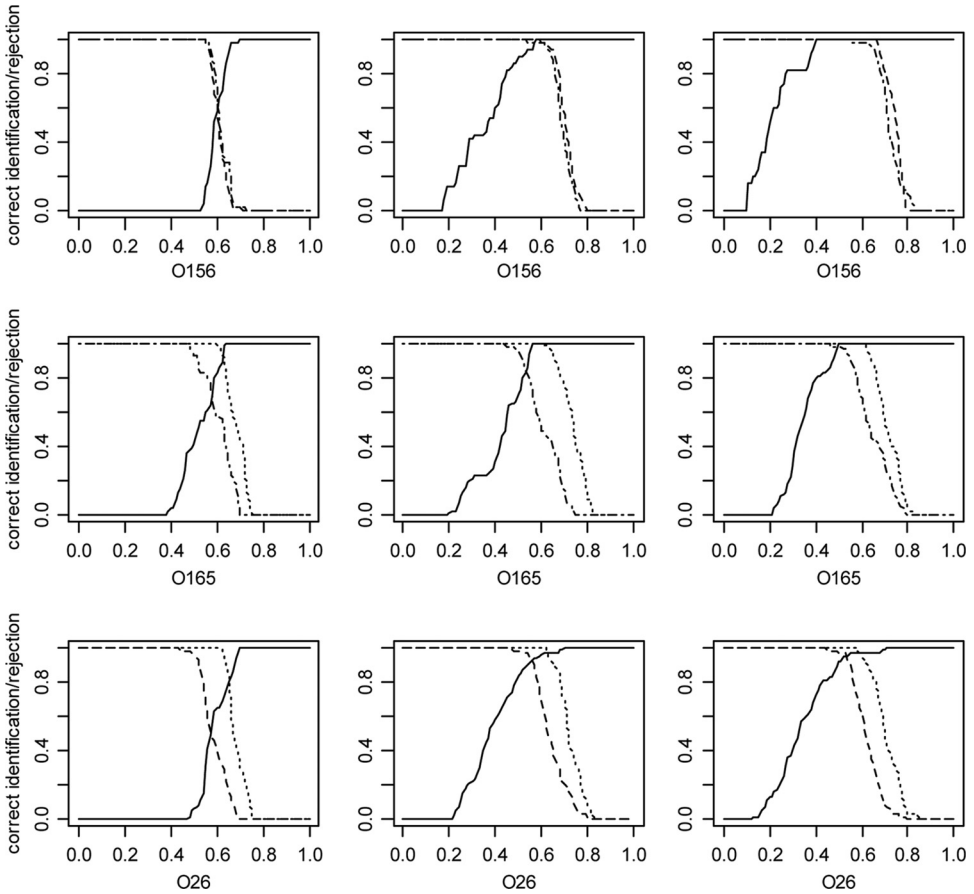

FIG. 3. Quality functions for the assignment of serotypes. The results for three parameter sets are shown: $P \leq 1$, minimum frequency > 0, corresponding to the complete complement of masses from the alignment (left panels); $P \leq 0.1$, minimum frequency > 0.3, corresponding to the optimized conditions for the assignment of all serotypes (middle panels); and *P* values ≤ 0.1, minimum frequency > 0.6 (right panels). The *x* axes represent binary distances of the tested spectra to the prototype spectra. Heavy lines indicate the percentage of isolates of the serotype indicated under the plot that have a distance smaller than the distance represented by the *x* axis. Dotted, dashed, and dot-dashed lines indicate the percentages of isolates of the serotypes not indicated under the plot that have a distance greater than the distance indicated by the *x* axis. Areas where the bold lines and the other lines reach 1.0 represent complete unequivocal discrimination of the respective groups. Dotted lines, O156:H25; dashed lines, O165:H25; dot-dashed lines, O26:H11/H32.

TABLE 3. Comparison of group distances and distances within groups[a]

pVal[b]	minFreq[c]	Group 1	Group 2	Group distance[d]	Distance within groups[e]	Quotient[f]
1	0	O156	O165	0.58	0.40	1.45
	0	O156	O26	0.56	0.38	1.48
	0	O165	O26	0.50	0.38	1.31
0.1	0.3	O156	O165	0.74	0.45	1.63
	0.3	O156	O26	0.73	0.45	1.60
	0.3	O165	O26	0.68	0.48	1.41
	0.6	O156	O165	0.69	0.37	1.87
	0.6	O156	O26	0.67	0.37	1.81
	0.6	O165	O26	0.63	0.41	1.54

[a] As a measure for the discrimination power of the serotype assignment under different conditions, the distance between the group means, the weighted mean distance of group members to their group mean and the quotient of these values were calculated. Higher qoutients indicate a better discrimination between the groups.
[b] pVal, maximum *P* value of the Fisher test accepted for every mass.
[c] minFreq, minimum frequency accepted for every mass.
[d] That is, the distance between the group means in the distance matrix.
[e] Weighted mean distances of all group members to their group means
[f] Calculated as the group distance/distance within groups.

Comparison of ICMS and RFLP analysis. Phylogenetic trees were created for each serotype based on ICMS results and compared to the phylogenetic trees based on the RFLP analysis.

In previous studies, the spatial and temporal relations of the STEC O165:H25, STEC O26:H11/H32, and STEC O156:H25 isolates in the beef herds had been analyzed during a monitoring program (5, 6, 8). Different STEC O26:H11/H32 clusters defined by RFLP analysis were found in three different farms. The STEC O26:H11 isolates of each farm represented independent clusters (Fig. 4, cluster 1, cluster 5). In contrast, the STEC O156:H25 strains did not represent independent farm-specific clusters, as members of the same clusters were found in three geographically distant farms. The genetic clusters also differed in their temporal distribution. Members of several clusters of STEC O26:H11 and STEC O156:H25 were detected over long periods (10 and 8 months, respectively), whereas members of other clusters were only detected on single occasions. These results suggested that some of the STEC O26:H11 and STEC O156:H25 strains had the potential for a longer persistence in the host population while others had not. Therefore, the rapid and cost-effective determination of serotypes and, if possible, the genetically defined clusters within the serotypes by ICMS would have been desirable to track the dissemination of EHEC isolates.

Although ICMS generally failed to reproduce the spatial or temporal relations between different *E. coli* isolates that had been identified by RFLP, rudimental clustering of ICMS based dendrograms according to the genetically determined clusters was occasionally observed. As an example, phylogenetic trees of O26:H11/H32 serotype samples constructed from RFLP and ICMS data are compared in Fig. 4. Whereas the branch of the ICMS based dendrogram highlighted by an asterisk represents 12 of the 15 members of cluster 7 of the RFLP based dendrogram, representatives of the other RFLP clusters were randomly distributed over the ICMS based dendrogram.

DISCUSSION

Species identification of bacterial samples by ICMS has become routine in many laboratories in recent years, but ICMS and related techniques have also been used for research applications such as subspecies level characterization of bacteria. As shown in Fig. 1, serotypes strongly influence the clustering of mass spectra taken from *E. coli* samples. However, a main obstacle for the ICMS-based determination of serotypes was the high similarity of the spectra from all serotypes. This problem has been addressed for the characterization of closely related species or for subspecies characterization of bacterial isolates by different degrees of data reduction leading to one or a few species-identifying biomarker ions (SIBI) or, when identified, proteins (SIBP) that are indicative for an individual analyte, e.g., a specific bacterial source of an environmental contamination (28), or certain taxon (14). For the discrimination of the three serotypes under investigation here, no such single ion or simple mass signatures could be pinpointed. Thus, prototype spectra representing the different serotype groups were constructed, and the similarity of sample spectra with the prototype spectra were used for the serotype assignment. For the required data reduction step, two simple and plausible statistical parameters were used to filter out less significant masses from the mass alignment: the frequency of occurrence in every serotype group and the *P* value of the Fisher test that was applied to test every mass of the alignment for differences in its frequency of occurrence in the serotype groups. As expected, removal of masses with low significance improved clustering (Fig. 2 and Table 3). However, similarity scores resulting from cluster analysis are difficult to interpret and require empirical cutoff values for the assignment of unknown samples to a certain taxon. For this reason, a bootstrap approach was implemented to evaluate the correctness of the assignment of a sample serotype to the serotype of the prototype spectrum with the smallest binary distance.

For the construction of informative prototype spectra removal of less significant masses was necessary and improved the false rate from over 30 to below 1% when parameters had been optimized. Surprisingly, application of highly stringent conditions were counterproductive in some combinations (Table 2). Recent publications on the discrimination of *Salmonella* subspecies (3) and *Pantoea* strains (23) using the SARAMIS software (Anagnostec, Germany) have reported minimum frequencies of 95 and 90%, respectively, as successful criterion for the selection of masses for the construction of taxon specific "superspectra" which are similar to the prototype spectra calculated here. For the discrimination of the three *E. coli* serotypes under investigation, a minimal frequency of only 30% was optimal when the *P* value filter was not applied. Most probably, the reduced discrimination power after application of highly stringent conditions was linked to the small number of remaining masses which varied in a wide range (ca. 270 with no filters applied, 110 under optimized conditions and 30 with *P* values of ≤ 0.001 and minimum frequency of >0.9), indicating that excessive data reduction may have unfavorable effects. Therefore, the program used here was designed to systemati-

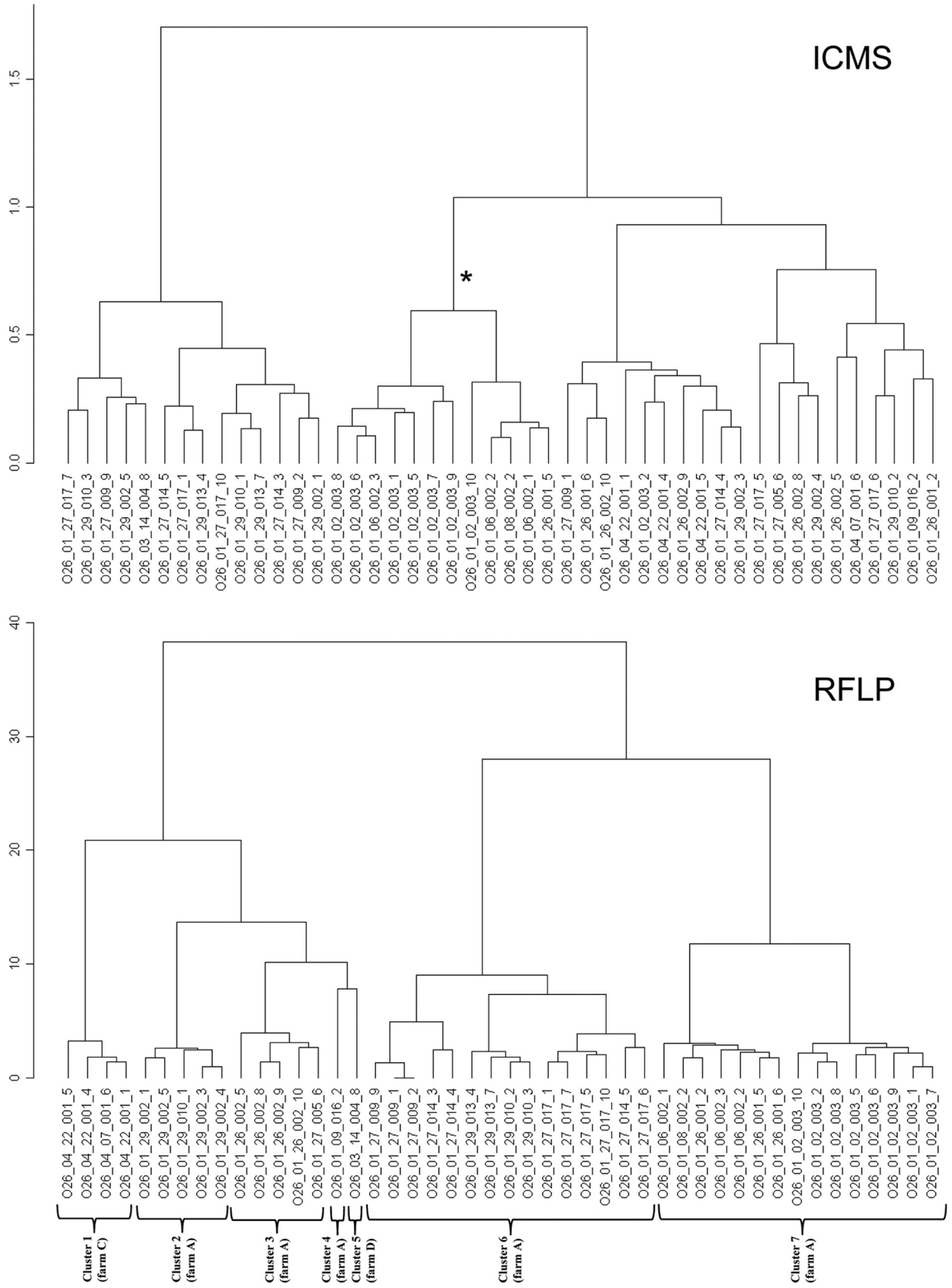

FIG. 4. Comparison of dendrograms calculated from RFLP and ICMS data using the masslists optimized for the discrimination of serotypes. Clusters obtained with RFLP data reflected the temporally and spatially related isolate groups (genetic clusters) indicated. As an example for rudimental cluster formation according to clusters defined by RFLP analysis, the highlighted (*) branch of the ICMS-based dendrogram contains 12 of the 15 representatives of the genetically defined cluster 7. The other clusters could not be reconstructed from the ICMS based dendrogram. Sample designation follows the pattern given in the footnote of Table 1, hyphens and slashes have been replaced by underscores.

cally vary minimum frequency and *P* values in order to optimize serotype assignment.

Optimal results were obtained only when both filters were set, although variation of the minimum frequency was more influential on the correctness of the serotype assignment than the variation of the *P* value. Although a single parameter set was found that allowed to reliably distinguish all three serotypes, the discrimination of O156 from the other two serotypes could be still enhanced by application of more stringent conditions, indicating that the parameters need to be customized to meet the requirements of a specific diagnostic demand in an optimal way. Under the most stringent conditions in Fig. 3, a binary distance between approximately 0.41 and 0.52 marks the range in which a complete and unequivocal discrimination of O156 from the other two serotypes was achieved, that is, any cutoff distance chosen in this range correctly assigned the O156 serotype to all samples with that serotype. Although wrong serotype assignments did occur with the parameters optimized for the discrimination of all three serotypes, ICMS with parameters tailored to the respective diagnostic challenge may be the method of choice for the discrimination of closely related bacteria isolates, e.g., when high throughput or low cost are important. The procedure for the calculation of prototype spectra representing subgroups of closely related microorganisms and the assignment of unknown samples to these groups is generally applicable under the condition that a sufficient number of reference samples is available.

With the exception of one case where rudimental clustering of mass spectra according to the given genetically defined clusters was observed, ICMS generally failed to reproduce genetic clusters within the serotypes (Fig. 4). The small genetic differences between these clusters did not seem to be sufficient to significantly affect MALDI mass spectra, indicating the limitations of ICMS.

REFERENCES

1. **Bielaszewska, M., et al.** 2007. Shiga toxin-mediated hemolytic-uremic syndrome: time to change the diagnostic paradigm? PLoS One **2:**e1024.
2. **Coimbra, R. S., et al.** 2000. Identification of *Escherichia coli* O-serogroups by restriction of the amplified O-antigen gene cluster (*rfb*-RFLP). Res. Microbiol. **151:**639–654.
3. **Dieckmann, R., R. Helmuth, M. Erhard, and B. Malorny.** 2008. Rapid classification and identification of salmonellae at the species and subspecies levels by whole-cell matrix-assisted laser desorption ionization-time of flight mass spectrometry. Appl. Environ. Microbiol. **74:**7767–7778.
4. **Everitt, B.** 1974. Cluster analysis. Heinemann Educational Books, London, United Kingdom.
5. **Geue, L., et al.** 2009. Analysis of the clonal relationship of serotype O26:H11 enterohemorrhagic *Escherichia coli* isolates from cattle. Appl. Environ. Microbiol. **75:**6947–6953.
6. **Geue, L., et al.** 2010. Rapid microarray-based genotyping of Enterohemorrhagic *Escherichia coli* (EHEC) serotypes O156:H25/H-/Hnt isolated from cattle and analysis of the clonal relationship. Appl. Environ. Microbiol. **76:**5510–5519.
7. **Geue, L., et al.** 2002. A long-term study on the prevalence of Shiga toxin-producing *Escherichia coli* (STEC) on four German cattle farms. Epidemiol. Infect. **129:**173–185.
8. **Geue, L., T. Selhorst, C. Schnick, B. Mintel, and F. J. Conraths.** 2006. Analysis of the clonal relationship of Shiga toxin-producing *Escherichia coli* serogroup O165:H25 isolated from cattle. Appl. Environ. Microbiol. **72:** 2254–2259.
9. **Giebel, R., et al.** 2010. Microbial fingerprinting using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) applications and challenges. Adv. Appl. Microbiol. **71:**149–184.
10. **Hartigan, J. A., and M. A. Wong.** 1979. A K-means clustering algorithm. Appl. Stat. **28:**100–108.
11. **Heller, D. N., R. J. Cotter, C. Fenselau, and O. M. Uy.** 1987. Profiling of bacteria by fast atom bombardment mass spectrometry. Anal. Chem. **59:** 2806–2809.
12. **Johnson, K. E., C. M. Thorpe, and C. L. Sears.** 2006. The emerging clinical importance of non-O157 Shiga toxin-producing *Escherichia coli.* Clin. Infect. Dis. **43:**1587–1595.
13. **Karger, A., B. Bettin, M. Lenk, and T. C. Mettenleiter.** 2010. Rapid characterisation of cell cultures by matrix-assisted laser desorption/ionisation mass spectrometric typing. J. Virol. Methods **164:**116–121.
14. **Krishnamurthy, T., P. L. Ross, and U. Rajamani.** 1996. Detection of pathogenic and non-pathogenic bacteria by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. **10:**883–888.
15. **Leopold, S. H., et al.** 2009. A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. Proc. Natl. Acad. Sci. U. S. A. **106:**8713–8718.
16. **Liebisch, B., and S. Schwarz.** 1996. Evaluation and comparison of molecular techniques for epidemiological typing of *Salmonella enterica subsp. enterica* serovar dublin. J. Clin. Microbiol. **34:**641–646.
17. **Marinach-Patrice, C., et al.** 2009. Use of mass spectrometry to identify clinical *Fusarium* isolates. Clin. Microbiol. Infect. **15:**634–642.
18. **Marklein, G., et al.** 2009. Matrix-assisted laser desorption ionization-time of flight mass spectrometry for fast and reliable identification of clinical yeast isolates. J. Clin. Microbiol. **47:**2912–2917.
19. **Mellmann, A., et al.** 2008. Analysis of collection of hemolytic-uremic syndrome-associated enterohemorrhagic *Escherichia coli.* Emerg. Infect. Dis. **14:**1287–1290.
20. **Nataro, J. P., and J. B. Kaper.** 1998. Diarrheagenic *Escherichia coli*. Clin. Microbiol. Rev. **11:**142–201.
21. **Platt, J. A., O. M. Uy, D. N. Heller, R. J. Cotter, and C. Fenselau.** 1988. Computer-based linear regression analysis of desorption mass spectra of microorganisms. Anal. Chem. **60:**1415–1419.
22. **R Development Core Team.** 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
23. **Rezzonico, F., G. Vogel, B. Duffy, and M. Tonolla.** 2010. Application of whole-cell matrix-assisted laser desorption ionization-time of flight mass spectrometry for rapid identification and clustering analysis of pantoea species. Appl. Environ. Microbiol. **76:**4497–4509.
24. **Sammon, J. W.** 1969. A non-linear mapping for data structure analysis. IEEE Trans. Comput. C **18:**401–409.
25. **Sauer, S., et al.** 2008. Classification and identification of bacteria by mass spectrometry and computational analysis. PLoS One **3:**e2843.
26. **Schmidt, F., T. Fiege, H. K. Hustoft, S. Kneist, and B. Thiede.** 2009. Shotgun mass mapping of *Lactobacillus* species and subspecies from caries related isolates by MALDI-MS. Proteomics **9:**1994–2003.
27. **Seng, P., et al.** 2009. Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin. Infect. Dis. **49:**543–551.
28. **Siegrist, T. J., et al.** 2007. Discrimination and characterization of environmental strains of *Escherichia coli* by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS). J. Microbiol. Methods **68:**554–562.
29. **Tuszynski, J.** 2009. caMassClass: processing & classification of protein mass spectra (SELDI) data. R package version 1.7.2009. R Foundation for Statistical Computing, Vienna, Austria.
30. **Villegas, E. N., S. T. Glassmeyer, M. W. Ware, S. L. Hayes, and F. W. Schaefer III.** 2006. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry-based analysis of *Giardia lamblia* and *Giardia muris*. J. Eukaryot. Microbiol. **53:**179–181.