Available online at www.sciencedirect.com
ScienceDirect
International Journal of Mass Spectrometry 261 (2007) 218–226
International Journal of Mass Spectrometry
www.elsevier.com/locate/ijms

Universal bacterial identification by mass spectrometry of 16S ribosomal RNA cleavage products

George W. Jackson [a,b,*,1], Roger J. McNichols [b,1], George E. Fox [a,c], Richard C. Willson [a,c]

[a] *Department of Chemical Engineering, University of Houston, 4800 Calhoun Avenue, Houston, TX 77204-4004, United States*
[b] *BioTex, Inc., 8058 El Rio St., Houston, TX 77054, United States*
[c] *Department of Biology and Biochemistry, University of Houston, 4800 Calhoun Avenue, Houston, TX 77204-5001, United States*

Received 28 June 2006; received in revised form 13 September 2006; accepted 13 September 2006
Available online 12 October 2006

Abstract

The public availability of over 180,000 bacterial 16S ribosomal RNA (rRNA) sequences has facilitated microbial identification and classification using nucleic acid hybridization and other molecular approaches. Species-specific PCR, microarrays, and *in situ* hybridization are based on the presence of unique subsequences in the target sequence and therefore require prior knowledge of what organisms are likely to be present in a sample. Mass spectrometry is not limited by a pre-synthesized inventory of probe/primer sequences. It has already been demonstrated that organism identification can be recovered from mass spectra using various methods including base-specific cleavage of nucleic acids. The feasibility of broad bacterial identification by comparing such mass spectral patterns to predictive databases derived from virtually all previously sequenced strains has yet to be demonstrated, however. Herein, we present universal bacterial identification by base-specific cleavage, mass spectrometry, and an efficient coincidence function for rapid spectral scoring against a large database of predicted "mass catalogs". Using this approach in conjunction with universal PCR of the 16S rDNA gene, four bacterial isolates and an uncultured clone were successfully identified against a database of predicted cleavage products derived from over 47,000 16S rRNA sequences representing all major bacterial taxa. At present, the conventional DNA isolation and PCR steps require approximately 2 h, while subsequent transcription, enzymatic cleavage, mass spectrometric analysis, and database comparison require less than 45 min. All steps are amenable to high-throughput implementation.

Keywords: MALDI-TOF; Ribosomal RNA; Spectral coincidence; Rapid microbial identification

1. Introduction

To this day, determinative bacteriology often relies on culture-based methods involving time-consuming isolation, cultivation, and characterization of phenotypic traits. While there are a few cases in which a rapid identification can be made using phenotypic methods, the taxonomic resolution of such methods is usually quite low. Characterization of cells based on morphology, staining, and metabolic traits is often not discriminatory and can take days to weeks for unambiguous identification. Perhaps most importantly, many pathogens are fastidious or even uncultivable under laboratory conditions, so that culture-based methods are not applicable. Finally, such methods are labor-intensive, not amenable to automation, and require extensive "hands-on" time and interpretation by the trained microbiologist. In the "post-genome" era, molecular methods are rapidly supplanting phenotypic characterization.

Although a variety of nucleic acid-based approaches are in use, most current bacterial diagnostic research is focused on comparative sequencing of PCR-amplified genes, *in situ* hybridization with labeled probes or molecular beacons, and phylogenetic microarrays [1–7]. Methods that rely on hybridization effectively leverage genomic information, but they typically face the significant drawback of requiring advance synthesis of one or more probes based on *a priori* anticipation of the genus or species that needs to be detected. Complete or partial

* Corresponding author at: BioTex, Inc., 8058 El Rio St., Houston, TX 77054, United States. Tel.: +1 713 741 0111; fax: +1 713 741 0122.
E-mail addresses: bill@biotexmedical.com (G.W. Jackson), willson@uh.edu (R.C. Willson).
[1] These authors contributed equally to this work.

1387-3806/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijms.2006.09.021

genomic sequencing requires no such preliminary knowledge, but even the fastest sequencing separations are time-consuming compared to mass spectrometry. Finally, both sequencing and hybridization probing require a means for radioisotope- or fluorescence-labeling. Although the *in vitro* transcription and cleavage reactions proposed here are similar in time and effort to that required by conventional sequencing, mass spectrometric acquisition is on the order of seconds (*versus* minutes or hours for conventional sequencing) such that the greatest gains in overall efficiency are had when processing multiple samples.

Microbial identification based on mass spectrometry of characteristic proteins or peptides has been demonstrated [8,9], however these approaches do not take advantage of the well-established phylogenetic relationships determined by 16S rRNA alignments (and the large number of sequences therefore freely available), nor do they take advantage of the amplification options afforded to nucleic acid-based approaches. Furthermore, they typically involve training of expert systems on relatively small sets of organisms, making it difficult to predict the generality of the identified biomarkers [10]. In addition to proteomics, genomics applications are now also adopting "soft ionization" mass spectrometric methods such as matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF) and electrospray ionization (ESI) [11–16]. While MALDI-TOF mass spectrometry has been used for chain-termination sequencing of nucleic acids [17–21], the maximum read length using such an approach is ~60 nucleotides [17]. Very high resolution measurement of PCR product composition using ESI-Fourier transform ion-cyclotron resonance (ESI-FTICR) has been demonstrated [22–26]. Unfortunately, the resolution required for unambiguous compositional assignment (±1 ppm) of such large molecules requires instrumentation out of reach for many laboratories. Thus, it is advantageous to introduce a cleavage step, which reduces the resolution requirements while retaining valuable information. Additionally, analysis of single-stranded nucleic acids is generally preferred as the same information content is available at roughly half the mass. Endoribonucleases can be used to selectively cleave a single-stranded RNA after a particular base (for example, guanosine residues in the case of RNase T_1). Despite the information loss associated with compositional rather than sequential analysis, microbial identification based upon the composition of base-specific cleavage products appears extremely promising [27–30]. von Wintzingerode et al. described comparison of base-specific cleavage patterns derived from *Bordetella* species against the patterns predicted by virtual cleavage of 50 published 16S rDNA sequences, including 13 sequences which were known to be closely related [27]. Discriminating masses (non-degenerate between the strains under consideration) were compared and strains were typed by inspection. Lefmann et al. used similar methods to rank the identification of mycobacteria [30].

In contrast to previous work, here we describe successful organism identification by comparison of observed masses to those predicted for all previously sequenced taxa. Experimental protocols for reliable generation of cleavage products containing mass-modified uridine residues from universally amplifiable bacterial sequence regions facilitate the approach, and an automated, quantitative method for rapidly scoring entire spectra of RNA cleavage products (including those masses which are shared degenerately by more than one organism or strain) against large databases of predicted masses results in accurate organism identification. We have recently shown that a great many pathogens of interest cluster (and important species and strains are resolvable) based on these observable mass spectral patterns [31].

In order to quantitatively inter-compare mass spectral "fingerprints" produced by base-specific cleavage, we formulated the scalar- or inner-product defined by Eq. (1). We define a scalar product (often referred to as a 'dot-product') of two mass spectra as

$$\langle M, M' \rangle = M \cdot M' \equiv \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \delta(m_i - m'_j) \tag{1}$$

where m_i are the masses of each of the N_1 individual cleavage products in the spectrum for species 1 and m'_j are the masses of each of the N_2 cleavage products for species 2, and δ is the discrete (Kronecker) delta function defined as

$$\delta(k) = \begin{cases} 1, & k = 0 \\ 0, & \text{otherwise} \end{cases} \tag{2a}$$

It can be easily verified that the following commutative, distributive, and positive-definiteness conditions for an inner-product are satisfied:

$$M_1 \cdot M_2 = M_2 \cdot M_1 \tag{3a}$$

$$(\alpha_1 M_1 + \alpha_2 M_2) \cdot M_3 = \alpha_1 M_1 \cdot M_3 + \alpha_2 M_2 \cdot M_3 \tag{3b}$$

$$M_1 \cdot M_1 > 0 \,\forall\, M_1 \neq [0] \tag{3c}$$

Using this inner-product, we then define the following metric or "coincidence function":

$$c_{ij} = c(M_i, M_j) = \frac{2 \times M_i \cdot M_j}{(M_i \cdot M_i) + (M_j \cdot M_j)} \tag{4}$$

that is a normalized (*i.e.*, between 0 and 1) representation of the degree to which two spectra are similar. While other similarity metrics have certainly been investigated [32–34], the various methods vary widely in computational intensity. For our feasibility study, our similarity metric was easily implemented, few computations per identification are required, and it is compatible with using (or not using) relative peak height information. Using this metric then, a coincidence (or similarity) matrix, $\mathbf{C}$ with elements $\boldsymbol{c}_{ij}$ can be generated to tabulate the degree of similarity between the mass catalogs of every pair of organisms. Likewise, a matrix of distances, $\mathbf{D}$ with elements $\boldsymbol{d}_{ij} = (1 - \boldsymbol{c}_{ij})$ can be created, and used as input to conventional cluster analysis.

Conveniently, Eqs. (3a)–(3c) also can be used for rapid comparison of observed spectra to predicted masses. As an example of the method, we demonstrate the successful (concordant with sequencing) identification of four organisms from pure cultures and an uncultured clone by quantitative comparison using Eqs. (3a)–(3c) of acquired spectra to a database of cleavage product

masses derived from over 47,000 sequences from the publicly available Ribosomal Database Project [2].

2. Experimental

2.1. Universal PCR, mass-modified transcription, cleavage, and MALDI preparation

Genomic DNA was released by boiling lysis of glycerol cell stocks of four model organisms archived at −80 °C. For each of the four stocks, 10 μl of glycerol cell stock was lysed in 1 ml deionized water for 2 min to release genomic DNA. Two microlitres (∼25 ng total genomic DNA) of such lysate was used directly for subsequent PCR. Using primers (Sigma–Genosys, The Woodlands, TX) described by Lane as "A", "B", and "C" [35,36], we generated PCR products corresponding to two contiguous sequence regions of the 16S gene obtainable from ∼80% of all previously sequenced bacteria [31]. To limit the complexity of the ultimate spectra acquired, we used only the adjacent Lane "AB" and "BC" primer pairs for amplification of ∼400 and ∼500 bp of 16S rDNA from most organisms, respectively. All forward primers employed also contained a 5′-extension for incorporation of a T7 RNA polymerase promoter sequence, and all reverse primers were "5′-tailed" with the reverse compliment of a sequence used for single point internal mass calibration. Twenty-five microlitres PCRs for all primer pairs were optimized using a FailSafe™ optimization kit (Epicentre, Madison, WI). Conventional PCR thermal cycling conditions in a GeneAmp 2400 thermal cycler (Perkin-Elmer) were: 5 min denaturation at 95 °C, 30 cycles of 95, 55, and 72 °C for 30, 30, and 45 s, respectively, followed by a 7 min extension at 72 °C totaling approximately 105 min. Following PCR, reaction mixtures were treated directly with 1 μl (20 units) of DNA exonuclease I (Epicentre) at 37 °C for 5 min to digest any unincorporated single-stranded primers. (Without this step, we have occasionally observed low yields of RNA transcript, presumably due to interaction of the RNA polymerase with primers or primer-dimers instead of the double-stranded PCR product.) Exonuclease I was then deactivated at 80 °C for 5 min. Approximately 2 μl (typically ∼1 μg DNA) of the resulting mixture was then used directly without purification as template for *in vitro* transcription for 30 min using a T7-flash™ kit (Epicentre) containing the manufacturer's suggested T7 RNA polymerase and nucleotide concentrations, except that amino-allyl UTP (Fermentas, Hanover, MD) was used as a 100% substitute for the natural UTP substrate, thereby improving the cleavage product mass resolution of the experiment (see Section 3). Following transcription, 1 μl of ribonuclease T_1 (1000 units) was added to each 20 μl reaction, and transcripts were cleaved (after G residues) for 5 min at 37 °C. Finally, RNA cleavage product mixtures were desalted by reverse phase purification using ZipTips™ (Millipore, Billerica, MA) per manufacturer's instructions for nucleic acid treatment [37]. In the final step, RNA cleavage products were eluted from the ZipTip columns with 2 μl MALDI matrix. The MALDI matrix comprised an 8:1 mixture of 3-hydroxypicolinic acid (3-HPA) and diammonium citrate (DAC). 3-HPA was prepared as a saturated solution in 50:50 water:acetonitrile, and DAC as a 50 mg/ml solution in RNase-free water. Samples were spotted on the MALDI plate in duplicate samples of 1 μl.

Sequencing of PCR amplicons was performed by Lone Star Laboratories (Houston, TX) on an ABI Model 3130 sequencer using Big Dye terminator chemistry v. 3.1 (Applied Biosystems, Foster City, CA) per manufacturer's instructions.

To demonstrate flexibility of the method, an uncultured clone, identified as *Stenotrophomonas maltophilia* by sequencing, was also successfully identified using identical methods as above following plasmid isolation using a standard alkaline lysis miniprep. A number of such clones have been graciously provided by Drs. Kasthuri Venkateswaran and David Newcombe of NASA's Jet Propulsion Laboratory, Pasadena, CA.

2.2. MALDI-TOF acquisition and spectral processing

All spectra were acquired using a Voyager DE-STR MALDI spectrometer (Applied Biosystems) in linear, negative ion mode. Typically, four 50-shot acquisitions were summed (200 shots total) for each 1 μl MALDI spot corresponding to the digests of either the "AB" or "BC" sequence regions from each organism. All mass-modified 6-mers are at least 1900 Da, and the vast majority of the other expected masses will be 5000 Da or less. The range of acquisition was therefore 1900–5000 Da. A 400 ns delayed extraction was also employed. All spectra were processed in an identical fashion using the Data Explorer™ software packaged with the instrument. These steps, in order, were: baseline correction (optional, never more than once), noise-filtering, mass calibration, centroiding, and thresholding the peak detection at the 7% of the maximum level. For mass calibration, an internal product mass common to all reactions is generated from the reverse compliment of the antisense primer. (This mass also serves as a confirmation that the RNA transcription was full-length.) The centroiding operation reduces spectra to a relatively short list of peaks with zero-width, and the resulting spectrum is then sent in text format as a list of masses for comparison to a database currently containing over 1,300,000 masses.

2.3. Comparison of acquired spectra to predictive database of masses

The creation of our large databases of masses corresponding to base-specific cleavage products from various sequence regions of 47,257 16S rRNAs has been previously described in detail [31]. Briefly, the databases are implemented in Linux Version 2.2.13 and maintained as a large number of multiline text files of cataloged lists of isotope-average masses corresponding to a particular organism and inter-primer region (in this case a subregion of 16S rDNA). Using either a shell or CGI web-interface, the list of observed masses is entered along with the corresponding primer pair, ribonuclease enzyme used (*i.e.*, RNase T_1, A, etc.), nucleotide mass-spreading substitutions (*e.g.*, aaU), and target database (*i.e.*, "Lane-AB" or "BC" sequence region). The list of coincidence scores between the entered spectrum and predicted spectra for all organisms in the

specified database is generated and sorted in ca. 90 s. When two spectra are available for a single sample (for example, Lane "AB" and "BC" regions), the pair-wise scores are multiplied to obtain a single overall score for each organism. In order to facilitate comparison of "real" spectra, a selectable mass-accuracy tolerance parameter (tol) (typically 1.0 Da, see Section 3) was added to the inner product by redefining Eq. (2a) so that

$$\delta(k) = \begin{cases} 1, & |k| \leq \text{tol} \\ 0, & \text{otherwise} \end{cases} \quad (2b)$$

Because shorter nucleotides are highly degenerate, only 6-mers and above are used for coincidence analysis. Comparison of both an "AB" and "BC" amplicon to a database of 47,257 organisms currently takes less than 3.5 min on a moderately powered PC. Further increases in speed should be easily realizable because the coincidence analysis is trivially parallelizable.

3. Results and discussion

Determination of RNA cleavage product compositions is challenging because of the small mass difference between U and C. (The repeating G, A, U, and C monomer masses for RNA are, respectively, 345.2, 329.2, 306.2, and 305.2 Da with pair-wise differences ca. 16, 39, 40, 23, 24, and 1 Da.) A number of methods have been described for generating RNA transcripts which are mass-modified and certain 2′ modifications allow RNA to be cleaved mono-specifically after bases other than G [13,15,29,38,39]. Here, we have incorporated amino-allyl uridine (aaU) residues as a 100% substitute for natural U in RNA transcripts, increasing the 1 Da difference between U and C to 55 Da and thereby increasing the resolving power of the experiment. Because aaU incorporation is widely used for labeling of nucleic acids, that is readily accessible to most laboratories without the need for mutant enzymes or more expensive modified nucleotides.

Fig. 1 shows comparison of spectra acquired from digests of aaU-modified RNA from the "AB" sequence regions (those lying between positions ∼520–925 defined by the Lane A and Lane B primers [35]) of two different organisms, *Pseudomonas aeruginosa* and *Vibrio proteolyticus*. The spectra are shown as raw data with only the single point mass calibration performed (in each case producing calibration offsets of ∼3 Da). The resolution at full-width half-maximum (FWHM) of the major peaks ranged from 450 to 800 (typical for operation in linear mode). At this resolution, without amino-allyl U modification, many products having only a U/C difference in composition would be superimposed. Instead, U/C "compomers" differ by ∼55 Da due to the amino-allyl group located on the 5-position of the uridine base. Table 1 shows expected masses and measured masses following noise filtering, single point calibration, centroiding, and a 7% of maximum intensity threshold applied to the calibrated raw spectrum from *P. aeruginosa* in Fig. 1, Panel A. Cleavage products having the closest masses under 100% amino-allyl U

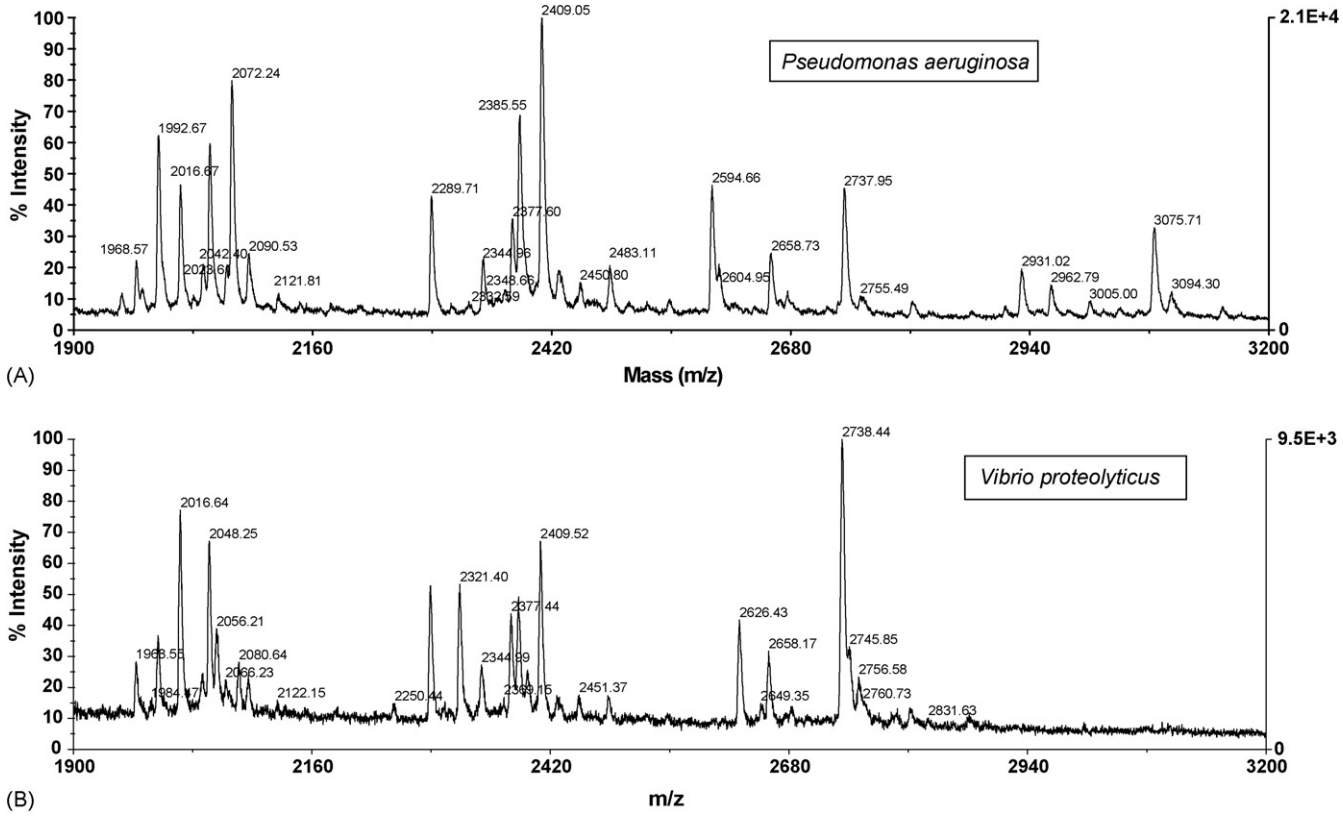

Fig. 1. (A and B) Comparison of typical spectra of base-specific cleavage patterns acquired from two different organisms. Only single point calibration has been performed on the raw spectra presented. The masses 2377 and 2385 Da corresponding to the mass-modified RNA oligonucleotides, UAAUACG and AUCCUUG, respectively, are illustrative of nearest mass-neighbors under conditions of 100% amino-allyl U substitution.

Table 1
Predicted masses for cleavage products length 6 and longer *vs.* all peaks observed for a typical digest of the "Lane AB" interprimer region in *P. aeruginosa*

Oligonucleotide sequence	Expected mass (Da)	Measured mass (Da)	Diff. (Da)
*U*CCACG	1968.23	1968.60	−0.36
ACAC*U*G	1992.26	1992.75	−0.49
C*U*AACG	1992.26	1992.75	−0.49
*U*AAACG	2016.28	2016.71	−0.43
	–	2041.10	–
A*U*AC*U*G	2048.32	2048.63	−0.31
A*U*A*U*AG	2072.35	2072.66	−0.32
	–	2091.01	–
CAAACAG	2289.43	2289.70	−0.28
	–	2345.69	–
*U*AA*U*ACG[a]	2377.53	2377.53	0.01
A*UCCUU*G[a,b]	2385.55	2385.55	0.00
*UU*AA*U*CG	2409.57	2409.74	−0.16
A*UCUU*AG	2409.57	2409.74	−0.16
	–	2428.65	–
	–	2483.91	–
AACACCAG[a]	2594.61	2594.63	−0.01
ACCACC*U*G[a]	2602.63	2601.82	0.81
A*U*ACCC*U*G	2658.69	2658.73	−0.03
AA*UU*AC*U*G	2738.78	2738.91	−0.12
AAA*U*CCCCG	2931.84	2931.76	0.08
C*U*CAACC*U*G	2963.88	2963.74	0.14
AA*UUU*CC*U*G	3076.01	3075.81	0.20
	–	3094.49	–
*UU*AAAAC*U*CAAA*U*G	4722.06	n.o.	–
CA*U*CCAAAAC*U*AC*U*G	4947.16	n.o.	–

Peaks are those detected above a threshold of 7%. U's are italicized to emphasize the amino-allyl U modification; n.o., expected but not observed.

[a] Has a nearest isotopic-average neighboring mass 8 Da away as predicted under amino-allyl U substitution (see text).

[b] Internal calibration standard mass (see text).

(aaU) modification are separated by ∼8 Da. This 8 Da difference is not attributable to a single monomer difference in composition but rather when a product of otherwise equivalent composition has one C and one aaU residue *versus* two A monomers. For instance cleavage product 1, AA*UU*CG = 2048.4 Da, while product 2, C*UUU*CG = 2056.4 Da, where the difference between the products is underlined and the U's are italicized to emphasize the amino-allyl modification. Two such situations are highlighted in Table 1 with the footnote 'a'. Most importantly, this minimum 8 Da difference for any two nearest G-specific cleavage products allows the experiment to be performed at this modest resolution and to set a reasonable tolerance (see Section 2) on the coincidence function without concern that organism identification will be hampered by U/C ambiguity. For organism identification, spectra are processed beyond the raw data depicted in Fig. 1. Note that in all cases, mass accuracy of the peaks picked by centroiding is within ±0.5 Da of the expected values. In the example of Table 1, four unexpected masses remain after spectral processing. Such mass signals are not merely noise and could be attributed to a number of phenomena including incomplete digestion by RNase T_1, products having a 2′–3′ cyclic monophosphate (a known intermediate product of endoribonucleases), cleavage products with remaining cation adducts, fragmentation during the MALDI process itself (*i.e.*, gas-phase dissociation), or products from incomplete transcriptions or infidelity of T7 RNA polymerase. Additionally, in contrast to most other prokaryotic genes, many bacteria have multiple copies of the ribosomal RNA operon that exhibit microheterogeneity in their sequence. Many of the ribosomal RNA sequences found in the RDP are therefore actually "composite" sequences [40]. Toward developing a rapid, universal assay for identifying bacteria, and based on our promising predictive results, the concern was not to determine the cause of each and every unexpected mass, but rather to determine if their occasional occurrence might prevent a correct identification. We therefore took the two relatively short lists of masses (similar to the list shown in Table 1) for each organism corresponding to digestion of the Lane "AB" and "BC" sequence regions of 16S and scored them against a database of mass catalogs corresponding to digestion of over 47,000 sequences. A final score for observed spectra was obtained by multiplying the "AB" and "BC" coincidence scores. For example, a database organism which exhibited a 70% coincidence score for both the "AB" and "BC" spectra from a given sample would receive an overall score of 0.49. We chose a ±1.0 Da tolerance (Eq. (2b)) on the coincidence function based on typical results for mass accuracy as shown in Table 1.

Table 2 gives the resulting organism identifications compared to those obtained by BLAST (against the entire NCBI nucleotide-nucleotide database) of sequences determined by conventional capillary electrophoresis sequencing of the same amplicons. All BLAST scores (bits) were taken as reported by the website without further modification. For all BLAST "hits" the expected value, E was zero indicating significant sequence matches. (A more detailed discussion of BLAST scoring can be found at http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html and elsewhere [41,42]. As can be seen, mass spectrometry was quite successful in identifying these bacteria when compared to full sequence analysis. For brevity, only the top eight scores for each method are included, however, many of the ambiguities resulting from "tie" scores were the same for both the mass- and sequence-based method. Furthermore, we should stress that only a single cleavage reaction and only the sense-strand of each "AB" or "BC" amplicon was used to obtain this result. In the case of *Escherichia coli*, correct identification to the strain-level was achieved (we knew *a priori* to be using a K-12 strain).

For both sequencing/BLAST and mass spectrometry, as yet uncultured *Acinetobacter* accessions received top scores. While our *Acinetobacter* strain, ATCC 33604 has no record in the RDP and is therefore not found in our database of masses, ATCC reports this organism as "*Acinetobacter* sp.; deposited as *Acinetobacter anitratus*". Table 2 shows that the organism with the second highest score for this strain by mass spectrometry is *Acinetobacter calcoaceticus* subspecies *anitratus*. The top scoring organism by mass spectrometry, "uncultured bacterium AY700608", is however classified in the RDP under the genus *Acinetobacter*.

P. aeruginosa was correctly identified to the species level with a relative separation of (0.471 − 0.455)/0.471 = 3.4% in "signal"

Table 2
Comparison of bacterial identification by conventional sequencing *vs.* base-specific mass spectrometric coincidence analysis

Sequencing/BLAST	Score, "bits" (rank)	Mass spectrometric coincidence analysis	Combined AB/BC score (rank)
Sample: *Escherichia coli* (K-12, MG1655)			
***E. coli* K-12 MG1655**	**930 (1)**	***E. coli*; K-12; M87049**	**0.476 (1)**
***E. coli* strain RW-29 16S...**	**930 (1)**	***E. coli*; K-12; U18997**	**0.476 (1)**
***E. coli* O157:H7 EDL933**	**930 (1)**	***E. coli*; O157:H7; BA000007**	**0.476 (1)**
***E. coli* 16S rRNA gene**	**930 (1)**	***E. coli*; AU1713; AY043392**	**0.476 (1)**
***E. coli* C2 16S rRNA**	**930 (1)**	***E. coli*; L10328**	**0.476 (1)**
***E. coli* 16S rRNA gene**	**930 (1)**	***E. coli*; CCCO4; AF511430**	**0.476 (1)**
Escherichia albertii strain 10457...	930 (1)	*Shigella flexneri* 2a str. 301; AE005674	0.476 (1)
E. albertii strain 12502...	930 (1)	*S. flexneri* 2a str. 2457T; AE016989	0.476 (1)
Sample: *Acinetobacter* sp. (ATCC 33604)			
Uncultured *Acinetobacter* sp. 16S rRNA	**759 (1)**	Uncultured bacterium; AY700608 (genus *Acinetobacter*)	0.303 (1)
***Acinetobacter* sp. H1 16S rRNA**	**759 (1)**	*Acinetobacter calcoaceticus* subsp. *anitratus*	0.298 (2)
***Acinetobacter* sp. phenon 10**	**759 (1)**	***Acinetobacter* sp. ATCC 31012; AF542963**	**0.288 (3)**
Uncultured clone ELB19-080 16S rRNA	759 (1)	*Acinetobacter grimontii* (T); AF509828	0.283 (4)
Acinetobacter junii 16S rRNA	759 (1)	*Acinetobacter johnsonii*; 5B02; AF509831	0.267 (5)
***Acinetobacter* sp. PAMU-1.11**	**759 (1)**	***Acinetobacter* sp. PAMU-1.11; AB118222**	**0.254 (6)**
***Acinetobacter* sp. phenon 3**	**759 (1)**	***Acinetobacter* sp. ADP1; CR543861**	**0.248 (7)**
Acinetobacter junii 16S rRNA	759 (1)	***Acinetobacter* sp. ADP1; 93A2; AJ812656**	**0.248 (7)**
Sample: *Pseudomonas aeruginosa* (ATCC 25102)			
***P. aeruginosa* gene for 16S rRNA**	**946 (1)**	***P. aeruginosa*; AT10; AJ549293**	**0.471 (1)**
Pseudomonas sp. pDL01 16S rRNA	938 (2)	*Pseudomonas alcaligenes* (T); LMG 1224T	0.455 (2)
Pseudomonas sp. Bxl-1	938 (2)	*P. alcaligenes*; M4-7; AY835998	0.446 (3)
***P. aeruginosa* ATCC BAA-1006**	**938 (2)**	***P. aeruginosa*; ATCC BAA-1006;**	**0.439 (4)**
Pseudomonas sp. BWDY-42 16S rRNA	938 (2)	***P. aeruginosa*; PAO1; AE004949**	**0.439 (4)**
***P. aeruginosa* partial 16S rRNA**	**938 (2)**	***P. aeruginosa*; ATCC 27853;**	**0.439 (4)**
Pseudomonas sp. HY-7 16S rRNA	938 (2)	***P. aeruginosa*; SCD-13;**	**0.439 (4)**
Pseudomonas sp. LQG-3 16S	938 (2)	*Pseudomonas* sp. pDL01; AF125317	0.439 (4)
Sample (clone): *Stenotrophomonas maltophilia*			
***S. maltophilia* AY748889.1**	**811 (1)**	***S. maltophilia*; ATCC 19861T**	**0.251 (1)**
***S. maltophilia* AY748888.1**	**811 (1)**	Uncultured beta proteobacterium; AF529323	0.246 (2)
***S. maltophilia* strain TKW2**	**811 (1)**	*P. aeruginosa*; SCD-1; AF448038	0.245 (3)
***S. maltophilia* strain B25R**	**811 (1)**	*P. aeruginosa*; AF225956	0.245 (3)
***S. maltophilia* strain B8R**	**811 (1)**	*Pseudomonas geniculata* (*T*); ATCC 19374T	0.245 (3)
***S. maltophilia* DQ141193.1**	**811 (1)**	MTBE-degrading bacterium PM1; AF176594	0.244 (6)
***S. maltophilia* AY360340.1**	**811 (1)**	Uncultured beta proteobacterium; AJ422152	0.244 (6)
Uncultured bacterium clone PDB-OTU11	811 (1)	Uncultured bacterium; W33; AY770973	0.243 (8)
Sample: *Vibrio proteolyticus* (ATCC 15338T)			
***V. proteolyticus* (ATCC 15338T)**	**944 (1)**	*Vibrio* sp.; VI1067/44; X97989	0.493 (1)
Uncultured bacterium clone PDC-OTU7	918 (2)	Uncultured bacterium; PDC-OTU7	0.489 (2)
Vibrio alginolyticus 16S rRNA	918 (2)	***V. proteolyticus* (T); ATCC 15338T**	**0.481 (3)**
Vibrio alginolyticus 16S rRNA	918 (2)	*Vibrio alginolyticus*; LA6; AF513447	0.481 (3)
Vibrio parahaemolyticus RIMD 2210633	918 (2)	*Vibrio parahaemolyticus* RIMD 2210633; O3:K6	0.481 (3)
Vibrio parahaemolyticus 16S rRNA	918 (2)	*Vibrio parahaemolyticus*; ATCC 17802	0.481 (3)
Vibrio sp. NLEP97-1598 16S	918 (2)	*Vibrio* sp. NLEP97-1598; AF410778	0.481 (3)
Vibrio sp. AB 16S rRNA gene	918 (2)	*Vibrio* sp. NAP-4; AF064637	0.481 (3)

The top eight scores for each method are presented. Species-level "hits" of the sample organism for each method are shown in bold, regardless of rank.

over the next ranked organism. The 16S sequence of our particular *Pseudomonas* strain, ATCC 25102 is also not found in the RDP (nor is it available in GenBank), however, our sequencing results indicate only 11 nt differences over the combined 918 nt "Lane AB+BC" sequence region between ATCC 25102 and the top strain identified by mass spectrometry (GenBank accession AJ549293).

As mentioned previously, a sequence cloned into *E. coli* from an uncultured organism was also identified by our methods. For this sample, the plasmid containing the cloned sequence was isolated prior to PCR using the Lane "AB" and "BC" primer pairs. Both conventional sequencing/BLAST and mass spectrometry returned the identification of the cloned bacterial sequence as *S. maltophilia*.

3.1. Effect of duplicate expected compositions and current form of the coincidence function

For the results presented in Table 2, duplicate cleavage product masses were maintained in the predictive catalogs for every organism. For example, referring to Table 1, if an organism in the database is expected to have both the oligonucleotides,

Table 3
Identification ranking of *V. proteolyticus* following removal of duplicate masses from database

Mass spectral coincidence	ID rank	Bacteria name	Notes
0.6417	**1**	***V. proteolyticus* (T); ATCC 15338T; X7472**	
0.6417	**1**	***V. proteolyticus*; PH8; AF513463**	
0.6417	1	*Vibrio parahaemolyticus*; MP-2; AY911391	
0.6417	1	*Vibrio natriegens*; 01/097; AJ874352	
⋮	⋮	⋮	39 other *Vibrio* species tied for 1st
0.6417	1	Uncultured bacterium; PDC-OTU1; AY700616	Was ranked 2nd with duplicate masses included in database (Table 2)
0.6412	45	*Vibrio alginolyticus*; AY373027	Was ranked 2nd with duplicate masses included in database (Table 2)
0.6050	107	*Vibrio* sp.; VI1067/44; X97989	Was ranked 1st with duplicate masses included in database (Table 2)
0.5862	126	*Listonella anguillarum* serovar O2a; AY069971	1st out-of-genus species, (in family *Vibrionaceae*)
0.5556	201	*Pseudoalteromonas* sp. RE1-12a; AF539781	1st out-of-family species
0.5498	252	*Vibrio cholerae*; CECT 514 T; X76337	1st *V. cholerae*, an important *Vibrio* pathogen
0.0170	18,062	uncultured bacterium; oc30; AY491573	lowest non-zero score
0.0000	18,063	*Bordetella pertussis* (T); ATCC 9797	Zero mass spectral coincidence (3809 other strains also had zero)

V. proteolyticus was also ranked 1st by conventional sequencing/BLAST (see Table 2).

ACACUG and CUAACG (1992.26 Da each), then two entries for that mass were maintained. Unfortunately, for observed spectra, it is difficult to quantitatively correlate relative MALDI peak height with the observance of such situations with any confidence (especially if the organism is unknown). We therefore count each *observed* mass only once regardless of peak height. Under our current formulation of the coincidence function, organisms with the largest number of predicted cleavage products are therefore penalized. This is precisely the situation for *V. proteolyticus* ATCC 15338T when compared to the two other *Vibrio* species ranked higher by our methods. That is, although the same masses are expected and measured for all three organisms, several duplicate masses are expected for *V. proteolyticus* (the actual organism) resulting in its lower ranking. Table 3 shows a re-ranking of organisms when duplicate expected compositions are ignored in the coincidence calculation. Not surprisingly less specificity results (44 organisms receive the top score), however the strain ranked first by sequencing, *V. proteolyticus* ATCC 15338T is among them. Such situations argue for separate cleavage of the antisense strand which is easily accomplished by choice of a different promoter on the reverse PCR primer. For example, the duplicate sense-strand RNase T_1 compositions 5′-ACACUG-3′ and 5′-CUAACG-3′ become 5′-. . .CAG/UG/U. . .-3′ and 5′-. . .CG/UUAG/. . .-3′, respectively, under RNase T_1 cleavage of the antisense strand. We are currently generating the necessary databases derived from the antisense sequences to take advantage of this complimentary information, and we fully expect that increased specificity will result. Due to the rapidity of mass spectral acquisition, very little time penalty would be associated with obtaining this complimentary information.

These brief examples may be indicative of the limits of organism-resolution of the method; while 16S sequence (and more recently, compositional) analysis is the prevailing molecular standard for determining phylogenetic relatedness, it may not be sufficient in all cases for strain-level identification [43], nor will it predict the presence of organisms expressing virulence factors. On the other hand, the methods described here, as well our coincidence function are certainly compatible with rapid analysis of other conserved genomic regions and should be extendable to both eukaryotes and viruses [44–47]. Finally, 16S rRNA sequences remain the largest dataset of gene-specific sequences, and, thus far, have proven as good as or better than other conserved genomic regions in determining relatedness [48].

The methods presented here are suitable for high-throughput identification of any organism-pure source, *i.e.*, cloned sequences, single colonies, cell stocks, or materials enriched for a single dominant organism such as a weaponized biomaterial. For the organisms under consideration, the patterns represented by Fig. 1 and Table 1 were information-rich enough to correctly identify organisms by quantitative comparison to a large database of masses, even with small numbers of unexpected masses included in the spectral scoring routine. Currently, we are working to further improve the fidelity of the spectra by repeated acquisition from model organisms, and we are assessing in greater detail the effects of observing spurious masses arising from minority chemical events, spectral processing (especially threshold levels in peak detection), and automated algorithms for “desalting” spectra. Most importantly, we have shown that there is no fundamental limitation of the database searching technique or misidentification due to cleavage product mass degeneracies, even when very large databases of masses are used. Furthermore, due to the rapidity of mass spectrometric acquisition and the ever-increasing amount of publicly available genomic information, other conserved sequence regions could be analyzed with little time penalty, thereby resolving any ambiguities among a particular group of organisms. We are currently investigating performance of the method on organism mixtures such that environmental and clinical samples can be fully characterized without cloning (currently the preferred method for “molecular sorting”). Simple binary mixtures of varying relative content should demonstrate the dynamic range of mass spectrometry in identifying a minority population amongst a majority, and previously described analytical and experimental

techniques [8,25,26,49–51] should facilitate extensive, rapid molecular characterization of microbial communities. While both preparative and data processing techniques will certainly aid the analysis of complex mixtures, we currently feel that the mass basis space and typical microbial community are likely too complex to be addressed only by computational methods (such as principal components analysis, etc.).

In general, these results indicate that complete cleavage after just one base should provide at least genus-level resolution of most bacteria, and that species- or strain-level identification may be achieved for some organisms using only the presented sequence regions of 16S rRNA. This can be improved by transcription and cleavage of the antisense strand, and/or cleavage after an alternative or additional base. For the purposes of developing a broad-based "sentinel" bacterial assay, this level of resolution may be acceptable. In situations, for example, in which an enriched unknown substance is presented, whether the rapid analysis indicates *Bacillus anthracis*, *Bacillus cereus*, or several other near phylogenetic neighbors may be irrelevant to the near-term prophylactic steps to be taken if an assay is sufficiently rapid. In the case of clinical diagnostics and response, many antibiotics have broad organism activity, so a rapid, genus-level diagnostic may be of higher resolution than the drugs actually available for certain infections.

4. Conclusions

We were able to generate highly characteristic mass spectra of base-specific cleavage products from bacterial 16S rRNA subsequences. While this has been demonstrated previously using various experimental methods, exhaustive scoring of such spectra against large predicted databases of masses has not. Using our coincidence analysis resulted in organism identification in ∼3 h with all methods amenable to high-throughput implementation, and requiring no pre-synthesis of probes or primers for organisms predicted to become of interest. While we have recently reported on the feasibility of organism identification by observing small signatures of just one, two, or three unique masses derived from a large sequence dataset [52], here we have taken entire patterns of RNA oligoribonucleotides length 6 and longer under consideration.

Segmentation of the analysis of the 16S gene into universally amplifiable subregions ultimately yields spectra of manageable complexity for accurate identification. With U and C residues better differentiated by 100% amino-allyl U substitution, acquired mass spectra can be "centroided" with increased confidence that strain-distinguishing masses differing by only a U or C residue are not convolved. The resulting spectra approximate a high resolution "bacterial barcode" of minimal data and maximum information content.

Acknowledgements

The work reported here was supported in part by grants from NASA (cooperative agreement NCC 9-58 and successor grant NNJ04HF43G) and the Welch Foundation to GEF and RCW and the Institute of Space Systems Operations to GEF, as well as Small Business Innovation and Research (SBIR) grants, from NASA (NNM06AA44C) and the USDA (#2006-33610-16775) to GWJ.

References

[1] J.J. Cannone, S. Subramanian, M.N. Schnare, J.R. Collett, L.M. D'Souza, Y. Du, B. Feng, N. Lin, L.V. Madabusi, K.M. Muller, N. Pande, Z. Shang, N. Yu, R.R. Gutell, BioMed Cent. Bioinform. 3 (2002) 2 (Correction: BioMed Cent. Bioinform. 3, 15).
[2] J.R. Cole, B. Chai, R.J. Farris, Q. Wang, S.A. Kulam, D.M. McGarrell, G.M. Garrity, J.M. Tiedje, Nucleic Acids Res. 33 (Database issue) (2005) D294.
[3] E.F. DeLong, G.S. Wickham, N.R. Pace, Science 243 (1360–1363) (1989).
[4] R. Amann, B.M. Fuchs, S. Behrens, Curr. Opin. Biotechnol. 12 (2001) 231.
[5] M. Wagner, M. Horn, H. Daims, Curr. Opin. Microbiol. 6 (2003) 302.
[6] E. Busti, R. Bordoni, B. Castiglioni, P. Monciardini, M. Sosio, S. Donadio, C. Consolandi, L. Rossi Bernardi, C. Battaglia, G. De Bellis, BMC Microbiol. 2 (2002) 27.
[7] D.P. Chandler, G.J. Newton, J.A. Small, D.S. Daly, Appl. Environ. Microbiol. 69 (2950–2958) (2003).
[8] A. Hu, P.-J. Tsai, Y.-P. Ho, Anal. Chem. 77 (2005) 1488.
[9] C. Fenselau, P.A. Demirev, Mass. Spectrom. Rev. 20 (2001) 157.
[10] O. Schmid, G. Ball, L. Lancashire, R. Culak, H. Shah, J. Med. Microbiol. 54 (2005) 1205.
[11] P.F. Crain, J.A. McCloskey, Curr. Opin. Biotechnol. 9 (1) (1998) 25.
[12] A. Null, L. Benson, D. Muddiman, Rapid Commun. Mass Spectrom. 17 (24) (2003) 2699.
[13] S. Krebs, I. Medugorac, D. Seichter, M. Forster, Nucleic Acids Res. 31 (7) (2003) e37.
[14] M. Ehrich, S. Böcker, D. van den Boom, Nucleic Acids Res. 33 (4) (2005) e38.
[15] P. Stanssens, M. Zabeau, G. Meersseman, G. Remes, Y. Gansemans, N. Storm, R. Hartmer, C. Honisch, C.P. Rodi, S. Bocker, D. van den Boom, Genome Res. 14 (1) (2004) 126.
[16] T. Sasayama, M. Kato, H. Aburatani, A. Kuzuya, M. Komiyama, J. Am. Soc. Mass Spectrom. 17 (2005) 3.
[17] Y. Kwon, K. Tang, C. Cantor, H. Koster, C. Kang, Nucleic Acids Res. 29 (3) (2001) E11.
[18] M.T. Roskey, P. Juhasz, I.P. Smirnov, E.J. Takach, S.A. Martin, L.A. Haff, Proc. Natl. Acad. Sci. U.S.A. 93 (10) (1996) 4724.
[19] B. Spottke, J. Gross, H.J. Galla, F. Hillenkamp, Nucleic Acids Res. 32 (12) (2004) e97.
[20] H. Koster, et al. DNA diagnostic (sic) based on mass spectrometry. United States patent 5,605,798 and continuations (1997).
[21] H. Koster, et al. DNA diagnostics based on mass spectrometry. United States patent 6,043,031 and continuations (2000).
[22] D. Muddiman, D. Wunschel, C. Liu, L. Pasa-Tolic, K. Fox, A. Fox, G. Anderson, R. Smith, Anal. Chem. 68 (1996) 3705.
[23] D. Wunschel, D. Muddiman, K. Fox, A. Fox, R. Smith, Anal. Chem. 70 (1998) 1203.
[24] D. Muddiman, A.P. Null, J.C. Hannis, Rapid Commun. Mass Spectrom. 13 (1999) 1201.
[25] S.A. Hofstadler, R. Sampath, L.B. Blyn, M.W. Eshoo, T.A. Hall, Y. Jiang, J.J. Drader, J.C. Hannis, K.A. Sannes-Lowery, L.L. Cummins, Int. J. Mass Spectrom. 242 (1) (2005) 23.
[26] D.J. Ecker, R. Sampath, L.B. Blyn, M.W. Eshoo, C. Ivy, J.A. Ecker, B. Libby, V. Samant, K.A. Sannes-Lowery, R.E. Melton, K. Russell, N. Freed, C. Barrozo, J. Wu, K. Rudnick, A. Desai, E. Moradi, D.J. Knize, D.W. Robbins, J.C. Hannis, P.M. Harrell, C. Massire, T.A. Hall, Y. Jiang, R. Ranken, J.J. Drader, N. White, J.A. McNeil, S.T. Crooke, S.A. Hofstadler, Proc. Natl. Acad. Sci. U.S.A. 102 (22) (2005) 8012.
[27] F. von Wintzingerode, S. Bocker, C. Schlotelburg, N.H. Chiu, N. Storm, C. Jurinke, C.R. Cantor, U.B. Gobel, D. van den Boom, Proc. Natl. Acad. Sci. U.S.A. 99 (10) (2002) 7039.
[28] R. Hartmer, N. Storm, S. Boecker, C.P. Rodi, F. Hillenkamp, C. Jurinke, D. van den Boom, Nucleic Acids Res. 31 (9) (2003) e47.

[29] S. Hahner, H.C. Ludemann, F. Kirpekar, E. Nordhoff, P. Roepstorff, H.J. Galla, F. Hillenkamp, Nucleic Acids Res. 25 (10) (1997) 1957.
[30] M. Lefmann, C. Honisch, S. Bocker, N. Storm, F. von Wintzingerode, C. Schlotelburg, A. Moter, D. van den Boom, U.B. Gobel, J. Clin. Microbiol. 42 (1) (2004) 339.
[31] G.W. Jackson, R.J. McNichols, G.E. Fox, R.C. Willson, BioMed Cent. Bioinform. 7 (321) (2006).
[32] S. Stein, D. Scott, J. Am. Soc. Mass Spectrom. 5 (1994) 859.
[33] K.X. Wana, I. Vidavskya, M.L. Grossa, J. Am. Soc. Mass Spectrom. 13 (1) (2002) 85.
[34] M.J. MacCoss, C.C. Wu, J.R. Yates III, Anal. Chem. 74 (2002) 5593.
[35] D.J. Lane, B. Pace, G.J. Olsen, D.A. Stahl, M.L. Sogin, N.R. Pace, Proc. Natl. Acad. Sci. U.S.A. 82 (20) (1985) 6955.
[36] W.G. Weisburg, S.M. Barns, D.A. Pelletier, D.J. Lane, J. Bacteriol. 173 (2) (1991) 697.
[37] Technical Note, TN225, Sample Preparation of Oligonucleotides Prior to MALDI-TOF MS Using ZipTipC18 and ZipTipμ-C18 Pipette Tips, 2000. http://www.millipore.com/ziptip.
[38] R. Sousa, R. Padilla, EMBO J. 14 (18) (1995) 4609.
[39] R. Padilla, R. Sousa, Nucleic Acids Res. 27 (6) (1999) 1561.
[40] J. Klappenbach, P. Saxman, J.R. Cole, T. Schmidt, Nucleic Acids Res. 29 (1) (2001) 181.
[41] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucleic Acids Res. 25 (17) (1997) 3389.
[42] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, J. Mol. Biol. 215 (1990) 403.
[43] G.E. Fox, J.D. Wisotzkey, P. Jurtshuk Jr., Int. J. Syst. Bacteriol. 42 (1) (1992) 166.
[44] L. Fulop, A.D.T. Barrett, R. Phillpotts, K. Martin, D. Leslie, R.W. Titball, J. Virol. Methods (179–188) (1993) 44.
[45] S.R. Santos, H. Ochman, Environ. Microbiol. 6 (7) (2004) 754.
[46] N. Scaramozzino, J.-M. Crance, A. Jouan, D.A. DeBriel, F. Stoll, D. Garin, J. Clin. Microbiol. 39 (5) (2001) 1922.
[47] R. Kappe, C. Fauser, C.N. Okeke, M. Maiwald, Mycoses 39 (1–2) (1996) 25.
[48] H. Teeling, F.O. Gloeckner, BMC Bioinform. 7 (66) (2006).
[49] S. Gygi, B. Rist, S. Gerber, F. Turecek, M. Gelb, R. Aebersold, Nat. Biotechnol. 17 (1999) 994.
[50] J. Peng, S. Gygi, J. Mass Spectrom. 36 (2001) 1083.
[51] R. Tsuchihashi, F. Loge, J. Darby, Water Environ. Res. 75 (4) (2003) 292.
[52] Z. Zhang, G.W. Jackson, G.E. Fox, R.C. Willson, BioMed Cent. Bioinform. 7 (117) (2006).