Mass Spectrometry-Based Proteomics Combined with Bioinformatic Tools for Bacterial Classification

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · FEBRUARY 2006

Impact Factor: 4.25 · DOI: 10.1021/pr050294t · Source: PubMed

Jacek Dworzanski

Leidos, Inc.

41 PUBLICATIONS **805** CITATIONS

SEE PROFILE

Rabih E Jabbour

United States Army

24 PUBLICATIONS **239** CITATIONS

SEE PROFILE

A. Peter Snyder

112 PUBLICATIONS **1,582** CITATIONS

SEE PROFILE

research articles Journal of proteome research

Mass Spectrometry-Based Proteomics Combined with Bioinformatic Tools for Bacterial Classification

Jacek P. Dworzanski,*,† Samir V. Deshpande,‡ Rui Chen,§ Rabih E. Jabbour,† A. Peter Snyder,‖ Charles H. Wick,‖ and Liang Li§

Geo-Centers, Inc., Aberdeen Proving Ground, Maryland 21010-0068, Science and Technology Corporation, Edgewood, Maryland 21040, Department of Chemistry, University of Alberta, Edmonton, Alberta T6G 2G2, and U. S. Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, Maryland 21010-5424

Received September 1, 2005

Timely classification and identification of bacteria is of vital importance in many areas of public health. We present a mass spectrometry (MS)-based proteomics approach for bacterial classification. In this method, a bacterial proteome database is derived from all potential protein coding open reading frames (ORFs) found in 170 fully sequenced bacterial genomes. Amino acid sequences of tryptic peptides obtained by LC−ESI MS/MS analysis of the digest of bacterial cell extracts are assigned to individual bacterial proteomes in the database. Phylogenetic profiles of these peptides are used to create a matrix of sequence-to-bacterium assignments. These matrixes, viewed as specific assignment bitmaps, are analyzed using statistical tools to reveal the relatedness between a test bacterial sample and the microorganism database. It is shown that, if a sufficient amount of sequence information is obtained from the MS/MS experiments, a bacterial sample can be classified to a strain level by using this proteomics method, leading to its positive identification.

Keywords: classification of bacteria • proteomics • tandem mass spectrometry • LC−MS/MS • bioinformatics

Introduction

Rapid and accurate classification of microorganisms including pathogenic organisms plays a critical role in areas of public health and is important for bacterial taxonomy. Moreover, detection and identification of pathogenic agents is of paramount importance in response to unintentional or terrorist outbreaks of infectious diseases. For example, many of the pathogenic organisms are considered as potential biowarfare agents (USA CDC, Categories A−C).[1]

There are several different ways of classifying microorganisms based on the analysis of chemical signatures, or by using molecular microbiology methods.[2] With the availability of genome sequences of an increasing number of microorganisms, classification of microorganisms, such as bacteria, can be done based on the relatedness of their genome sequences.[2,3] Sequence similarities are determined based on DNA−DNA hybridization results, mainly to determine groupings on the species/genus level (e.g., DNA−DNA reassociation values higher than 70% between strains are considered as the same species), and by sequencing of 16S rDNA and selected protein coding genes, like *gyrA*.[2−4] In addition, a multilocus sequence typing (MLST) approach, based on sequence alignments of several genes coding housekeeping enzymes, is also being implemented in microbiology reference laboratories to provide information on genetic relationships between bacterial strains.[5]

In contrast to genome sequence comparison, the proteome of a microorganism serves as a unique and informative readout of both its phenotypic state, which results from cell responses to physiological and environmental perturbations, and genomic information reflected in the amino acid sequences of expressed proteins. For example, multilocus enzyme electrophoresis (MLEE) is a popular method used to characterize differences in electrophoretic mobilities of cellular enzymes to detect amino acid substitutions, and mobility variants of such enzymes are equated with alleles at the corresponding structural gene locus. MLEE is considered to be a classical method for bacterial population genetics and systematics, and has been used for decades to estimate the levels of genetic relatedness among isolates, populations and species.[6] However, a more reliable proteomics method to determine the overall genomic similarities among bacteria would be to infer them from a set of confidently identified peptide sequences mapped to diverse chromosomal locations.

With recent advances in mass spectrometry (MS) and proteome database searching methods, peptide sequencing can be done with high speed and accuracy. Tandem mass spectrometry (MS/MS) combined with electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI) for the analysis of peptides generated from cellular proteins has been developed as a potential tool for identification of targeted

* To whom correspondence should be addressed. Tel: (410) 436-6681. Fax: (410) 436-3764. E-mail: jacek.dworzanski@us.army.mil.

† Geo-Centers, Inc.

‡ Science and Technology Corporation.

§ University of Alberta.

‖ U. S. Army Edgewood Chemical Biological Center.

10.1021/pr050294t CCC: $33.50

PAGE EST: 11.5

Published on Web 00/00/0000

microorganisms.[9−13] For example, Warscheid et al.[14,15] demonstrated that acid extraction of bacterial spore proteins followed by peptide sequencing using MALDI MS/MS could be used to discriminate species from the genus *Bacillus* in spore mixtures. This approach was recently improved by using an ion trap mass spectrometer with an atmospheric pressure MALDI source to increase the speed of analysis and accuracy of ion selection for unique peptides.[16] Hu et al.[17] used selective MS/MS analysis of a few species-unique peptide ions from three bacteria for capillary electrophoresis (CE)−MS/MS-based detection of targeted species. VerBerkmoes et al.[18] evaluated the MS/MS method for identification of targeted bacteria present in a simulated environmental matrix by detecting the species-unique peptides from these bacteria. However, they also noted that, with an increase in the number of species in the database, there will be a smaller number of 'unique' peptides with their sequences matching to only one database strain for positive identification of the targeted species.

Recently, we reported a method of bacterial identification based on the number of peptide-to-bacterium assignments. The bacterium in the database with the highest number of matched peptides is deemed to be the correct one present in the sample.[19] This simplistic approach worked well for identifying a test bacterium whose proteome information was included in a small database. As the number of database bacteria increases, this approach, as in the case of detecting species-unique peptides for bacterial identification, will become difficult for unequivocal identification of an unknown bacterium.

In this work, we describe a proteomic approach to classify bacteria based on a set of peptide sequences determined by MS/MS. The peptide sequence information is used to classify an unknown bacterium by determining similarities between the investigated strain and database bacteria already grouped in accordance with their established taxonomic position, such as phylum, class, order, family, genus, and species. This general classification approach can be applied to any size of database and any number of experimentally determined peptide sequences. Depending on the sequence information obtained from a given MS/MS experiment, a bacterial sample can be classified to a particular taxonomic position, and may ultimately be identified at the species level.

Experimental Section

Strains and Media. *Escherichi coli* K-12 and *Bacillus cereus* were obtained from the American Type Culture Collection (ATCC) and were stored frozen before culturing. A starter culture of *E. coli* MG1655 (ATCC 47076) was first grown in 1065 LB medium at 37 °C without shaking for ∼20 h. After centrifugation (3500 × *g* for 5 min), the bacterial pellet of the starter culture was washed and suspended in 1550 mL of LB broth solution. After 2 h, 200 mL was transferred to a centrifuge tube and centrifuged for 20 min. The supernatant was then removed and the cells were rinsed twice with phosphate buffered saline and water (Milli-Q). The same procedure was then repeated after 6 and 53 h of incubation. *B. cereus* (ATCC 14579*)* cells were grown in Nutrient Broth (Difco BD 234000) according to the manufacturer's instructions. At the beginning of the stationary phase cells were harvested, washed with water (Milli-Q), lyophilized, and stored at −25 °C before analysis.

Preparation of Tryptic Peptides. Proteins were extracted from bacterial cells after lysis by sonication and were processed prior to LC−MS/MS analysis, as described previously.[19]

Liquid Chromatography−Mass Spectrometry of Peptides. LC−MS/MS analyses were carried out on an LCQ DECA Surveyor LC−MS system (ThermoFinnigan, San Jose, CA). Chromatographic separation was performed on a Vydac C18 column (300 Å, 5 *μ*m, 150 *μ*m i.d. × 150 mm) with a flow rate of 1 *μ*L/min.[19] The mobile phase consisted of water and acetonitrile, both containing 0.5% (v/v) acetic acid. Each MS data acquisition cycle consisted of a full-scan MS over the mass range 400−1400 *m*/*z*, followed by three data-dependent MS/MS scans over *m*/*z* 200−2000 on the three most intense precursor ions from the survey scan. The normalized collision energy was set at 35% with a 30 ms activation time. Dynamic exclusion was enabled with a repeat duration of 0.5 min, repeat count of 2, and a 3 min exclusion duration.

Proteome Database. Genomes and annotated theoretical proteomes were downloaded from the National Institutes of Health National Center for Biotechnology Information (NCBI) ftp server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria) on October 5, 2004. They comprised the fully assembled genomes of 170 bacteria, including their sequenced plasmids, representing 85 genera, 61 families, 41 orders, 20 classes and 12 phyla, and are detailed in Table S1 (see Supporting Information). A proteome database was assembled in a FASTA format and each protein header line was modified by adding genomic positions of respective ORFs. The database was comprised of sequences translated from 461 881 ORFs recognized by gene finding algorithms. Assignments of peptide sequences shared by multiple bacterial proteomes were further validated by using BLAST-P searches[20] against the NCBI protein database.

Data Processing and Analysis Algorithms. We have developed a suite of algorithms for the analysis of bacteria similarities and their classification that were implemented in Microsoft Visual Basic.NET and PERL. Figure 1 shows a schematic representation of the data processing procedures performed by the bioinformatics tools. The SEQUEST output files[21] were read and processed to retrieve peptide sequences that match MS/MS spectra, and assigned scores as a measure of the quality of fit between such matches. Identified peptides were matched against the proteome database and mapped onto respective chromosomes. This proteogenomic mapping was used for a reconstructed display of expressed proteins in a 2D-plot. Such 2D-plots show the chromosomal location of predicted genes (ORFs), indicate which DNA strand encodes a given ORF, and the potential length of the encoded protein expressed as the number of codons.

The SEQUEST generated matching scores were analyzed using a PeptideProphet algorithm, developed at the Institute of Systems Biology[22], to provide probabilities that sequence assignments to MS/MS spectra are correct. Sequence unique peptides with the probability of correct identification higher than 98% were retained in the dataset, and were used to generate a binary matrix of sequence-to-bacterium assignments (SBAs). The resulting matrixes were further processed to classify and potentially identify an analyzed bacterium using multivariate statistical techniques.

The logic of the data processing workflow shown in Figure 1 can be described as follows. During the analysis of an unknown bacterium *u*, database searches with uninterpreted MS/MS spectra of peptide ions give peptide sequences, which can then be validated using probability criteria. A set of *m* accepted peptide sequences s_i where $i = 1, 2, 3,..., m$, can be considered as elements of a column vector $\mathbf{b}_u$ that represents the peptide profile of *u* composed of *m* assignments a_{iu} ($\boldsymbol{b}_u =$

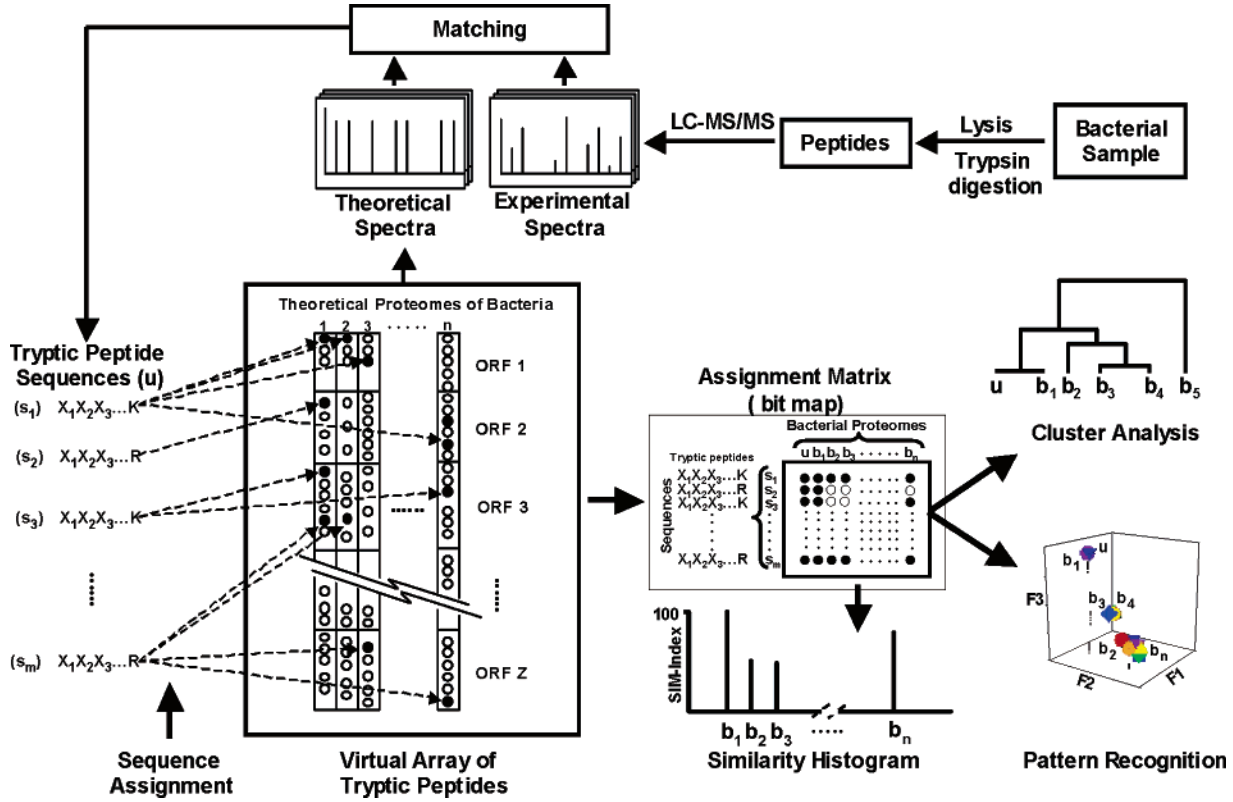

Figure 1. Schematic representation of sample analysis and data processing workflow for proteomics-based bacterial classification.

a_{1u}, a_{2u}, a_{3u}, ..., a_{mu}). Accordingly, sequence-to-bacterium assignments a_{i1} are elements of a column vector $\mathbf{b}_1$ that represents a peptide profile of a database bacterium assigned as number 1, and in general, assignments a_{ij} are elements of column vectors $\mathbf{b}_j$, where j represents the theoretical proteome of a j^{th} bacterium in the database ($j = 1, 2, 3,..., n$). All these column vectors form a binary matrix of assignments $\boldsymbol{A}_{m \times (n+1)}$ that can be represented as a virtual array of m peptide sequences assigned to $n = 170$ theoretical proteomes of database bacteria and an unknown microorganism. Each column vector represents a peptide profile of a bacterium, while each row vector represents a phylogenetic profile $\boldsymbol{s}_\mathbf{i}$ ($\boldsymbol{s}_\mathrm{i} = a_{iu,}\ a_{i1}, a_{i2}, a_{i3}, ..., a_{in}$) of a peptide sequence. Thus, for each LC−MS/MS analysis, a matrix of sequence-to-bacterium assignments is created with entries representing the presence or absence of a given sequence in each bacterial theoretical proteome. The absence of a given peptide sequence s_i in a proteome of a microorganism b_j (coded by the $j^{-\text{th}}$ genome) is represented by zero ($a_{ij} = 0$), while its presence is reflected by the entry of unity ($a_{ij} = 1$).

For comparisons of an unknown strain u to r similar strains (e.g., family or genus), a new matrix of differences $\boldsymbol{D}_{\text{diff}}$ ($\boldsymbol{D}_{\text{diff}} = \mathbf{b}_\mathrm{u} - \mathbf{b}_\mathrm{j}$, where $j = 1, 2, 3,..., r$) was generated and used to visualize differences in sequence assignments. Functional annotation of proteins was obtained from the NCBI web site.

Similarities between an unknown sample and database bacteria for construction of similarity histograms were calculated by using a Jaccard's coefficient. In the first test, the difference between similarity coefficient values for the two highest scoring taxa is compared to the known error rate for the dataset ('noise level'). When it exceeds a preset threshold value (e.g., $S/N = 3$), such difference is deemed to be significant and a taxon with the highest similarity is considered as a potential candidate for the correct sample classification.

To verify the correctness of the above classification test, a second test was devised[19] which is based on the following premises. Namely, if a result of the first test is correct, then peptides assigned to other taxa at the same classification level (i.e., phylum, class, order, etc.) should comprise only subsets of those matching the highest scoring taxon. Conversely, ifa result of the first test is incorrect, it means that some peptides have unique sequences for other taxa. However, the presence of such unique peptides is considered as statistically significant only if the number of sequences unique for a given taxon exceeds the noise level ($S/N = 3$).

Hierarchical clustering analysis (HCA) was performed using diverse linkage methods (single, complete, Ward's) with squared Euclidean distances as the similarity measure. Principal component analysis (PCA) of an $m \times m$ covariance matrix obtained from an SBA matrix and projections of microorganisms into the dataspace of the three principal components with the highest eigenvalues was used to evaluate the observed groupings. Principal component and cluster analyses were performed using STACluster and STAFactor libraries from Statistica (release 6, StatSoft, Inc., Tulsa, OK).

Results and Discussion

Figure 1 illustrates the method that combines LC−ESI MS/MS analyses of bacterial protein digests and several bioinformatics tools to process the peptide sequence information for bacterial classification. The key components of the method are described below using illustrative examples of model gram-negative and gram-positive microorganisms.

Processing of Product Ion Mass Spectra. Hundreds of product ion mass spectra of peptide ions are generated during one-dimensional (1D) LC−MS/MS analysis of the bacterial cell extract digest. To identify matching sequences, product ion mass spectra were searched against the protein database composed of proteomes of 170 bacteria using SEQUEST. The validity of each match has to be evaluated to distinguish between real matches (i.e., correct identities) and random matches (i.e., incorrect identities). Moreover, for high throughput analysis of peptide sequence assignments to bacterial proteomes, the discrimination between correctly and incorrectly identified peptides has to be performed automatically. To this end, SEQUEST computed matching scores were interpreted using discriminant function (DF) analysis.[19,22] Multivariate DF analysis transformed the SEQUEST generated scores into DF scores, and distributions of correctly and incorrectly identified peptides were modeled and analyzed to determine the probabilities of correct peptide identification (P-values). The peptides with low *P*-values were filtered out by setting a user-selected probability threshold. The peptides with high *P*-values were considered to be the correctly identified peptides and they were retained to generate SBA matrixes that could be viewed as assignment bitmaps. These bitmaps were analyzed using multivariate projection and clustering techniques to determine similarities between an unknown bacterial sample and database microorganisms as described below.

Concept of the Classification Method. In contrast to many genomic techniques utilizing a limited number of genetic loci to determine similarities between an unknown and reference strains, we attempt to classify an unknown bacterium by using confidently identified peptide sequences derived from hundreds of loci. In this case, a peptide sequence can be assigned to at least one bacterial proteome in the database. Furthermore, the protein sequences can be traced to segments of protein encoding ORFs located on chromosomes or plasmids of database bacteria. Although some sequences are unique, i.e., they are mapped to the proteome of only one database bacterium, many sequences are mapped to more than one bacterial proteome. Because such shared peptides are not rare, the sequence-to-bacterium assignments have to be sorted and processed to extract useful taxonomic information. In our approach, yes or no is assigned to individual proteomes in the database for a detected peptide based on its sequence match to the proteomes. Hence, SBAs can be interpreted as binary characters (1's and 0's) and a pattern indicating the status of SBA for a given peptide sequence across the entire proteomes of database bacteria forms a phylogenetic profile of the peptide. Consequently, the profiles of all identified peptides form a binary matrix of SBAs. Such a data matrix can be further analyzed using a set of algorithmic approaches to carry out bacterial classification.

In our method, all database bacterial strains are classified in accordance with the established taxonomy of prokaryotic microorganisms. Similar bacterial strains are grouped into species while groupings of very similar species form genera. This species/genus level in the taxonomic position of each organism within the classification scheme is reflected in the binomial name of bacteria. However, groupings do not stop at this level, but also include broader taxonomic arrangements of organisms into hierarchical classifications based on similarities. Namely, similar genera are placed in the same family of microorganisms; similar families in the same order; similar orders in the same class; similar classes in the same phylum; and finally all bacterial phyla form the kingdom of bacteria (also known as eubacteria).

Classification of an unknown bacterium involves the analyses of sequence similarities between the investigated strain and database bacteria grouped in accordance to their established taxonomic position in a classification hierarchy.[23] Every bacterium may be assigned to one and only one taxon at each classification level. Therefore, the assignment of an unknown bacterium to a correct taxon should be associated with the highest similarity between a test sample and database bacteria placed in that taxon.

In reality, there are many sources of error that may cause deviation from this ideal pattern of assignments. These may include proteome sequence errors in a DNA-translated proteome database and experimental errors introduced at the sample preparation or spectra processing stages. Therefore, in our approach, an experimentally determined error rate associated with a set of accepted peptide sequences is used in testing the significance of any observed matching discrepancies. During the classification process, the algorithm applies decision criteria based on the significance of computed similarities between a test sample and taxons at each classification level.

A taxon with peptide assignments fulfilling these requirements (passing both tests) is considered as a valid identity of an unknown bacterium at a given classification level. The classification process is then repeated at the lower taxonomic level. Otherwise, this classification process is terminated and the data subset retained at the failed classification level is further analyzed using unsupervised statistical techniques, like PCA and HCA, to arrive at the final grouping of the unknown bacterium to one or a set of database bacteria.

Classification of a Test Sample. To illustrate the classification process, *E. coli* K-12 cells harvested during the stationary growth phase were analyzed and, from the digest of the cell extract, a set of 129 unique peptide sequences having a probability of correct identification greater than 99.5% were accepted for classification. The first task was to classify the test sample on the highest classification level of bacteria, i.e., to determine which phylum of database bacteria the test sample belongs to. Currently, all bacteria represented in the in-house database belong to 12 phyla (see the legend in Figure 2A and Supporting Table S1). They are demarcated mainly on the basis of sequence similarities of genes encoding their small subunit ribosomal RNA (SS rRNA).[4] The similarity results obtained for the test sample are displayed in Figure 2A. They were measured using similarity coefficient values computed for each phylum and the test sample (the coefficient values represent a fraction of identified peptides that match bacteria classified as members of a given phylum).

The histogram presented in Figure 2A indicates that 100% of identified sequences were assigned to the phylum *Proteobacteria,* while only a very small fraction was assigned to *Firmicutes* (1.6%) and *Bactoreidetes/Chlorobi* (0.8%). The error rate for these dataset was determined to be 0.5%; therefore the difference between similarity coefficients for *Proteobacteria* and *Firmicutes* (98.4%) is statistically significant. On the other hand, the assignment of 100% of peptides to *Proteobacteria* means that there is no unique peptide sequence matched to other phyla. Therefore these results allowed for an unambiguous identification of the test sample as a *Proteobacterium*.

Next, the identification process was repeated at the class level of the phylum *Proteobacteria* that is comprised of four groups, named α-, β-, γ-, and δ/ϵ-*Proteobacteria.* As in the phylum level classification, the SBA results obtained from this sample identified it as a member of γ-*Proteobacteria*. Continuation of

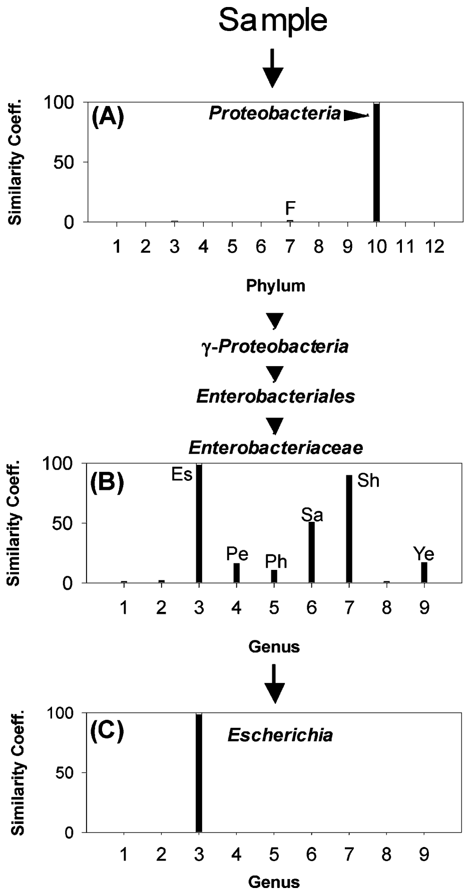

Figure 2. Hierarchical classification scheme of the stationary phase *E. coli* K-12 cells based on assignments to database proteomes of 129 tryptic peptide sequences identified during LC–MS/MS analysis. (A) Histogram of similarity coefficients (number of positive matches × 100/ total number of sequences identified for the test organism) calculated for 12 phyla obtained by merging database bacteria according to their taxonomic position[23] [Phyla: 1- *Actinobacteria*, 2- *Aquificae*, 3-*Bacteroidetes/ Chlorobi* group, 4- *Chlamydiae*, 5- *Cyanobacteria*, 6- *Deinococcus-Thermus*, 7- *Firmicutes* (F), 8- *Fusobacteria*, 9- *Planctomycetes*, 10- *Proteobacteria*, 11-*Spirochaetes*, 12- *Thermatogae*]. Further analysis of sequence assignments to *Proteobacteria* revealed a 100% match to *γ-Proteobacteria*, *Enterobacteriales* and *Enterobacteriaceae*. The assignments at the genus level are shown before (B) and after (C) the removal of shared (degenerate) peptides matching *Escherichia*; (for detailed explanation see text) [genera: 1- *Shewanella*, 2- *Buchnera*, 3- *Escherichia* (Es), 4- *Pectobacterium* (Pe), 5- *Photorhabdus* (Ph), 6- *Salmonella* (Sa), 7- *Shigella* (Sh), 8- *Wiagglesworthia*, 9- *Yersinia* (Ye)].

this process at the order and family levels indicated that the test sample belongs to *Enterobacteriales* and *Enterobacteriaceae*, respectively, and results of the analyses, represented as a hierarchy of similarity histograms, are summarized in Figure 2. However, at the genus level, the histogram (Figure 2B) revealed that as well as a very high level of similarity of the test sample to *Escherichia*, there was substantial similarity to *Shigella* and *Salmonella*. Nevertheless, the differences between the similarity coefficients for the genera *Escherichia* and those for *Shigella* and *Salmonella* were still significant and, in addition, there was a lack of any assignments uniquely specific for other genera (Figure 2C). Thus the analyzed sample could be classified as a highly probable member of *Escherichia*. The genus *Escherichia* is actually represented in the database by four strains of *E. coli* (K-12, CF073, and two O157:H7). By applying the same classification process at the strain level, we were able to identify the test sample as the K-12 strain of *E. coli.*

Analysis of Classification Results. The above example illustrates that the classification method may lead to the correct identification of the bacterium up to the strain level. However, due to the limited number of fully sequenced strains and the unsettled taxonomy of many species, the results obtained by using the above method should be viewed as a screening procedure aimed to find the most similar strains in the database. Nevertheless, this screening procedure is still very useful because it focuses the final classification process on a group of most similar database bacteria. Relationships between a test sample and such bacteria may be further analyzed using both multivariate statistical analyses and comparative studies of sequences associated with the annotated proteins.

An example of initial classification, followed by statistical analysis of the classification results, is a test sample containing *B. cereus* ATCC 14579 where digested peptides were sequenced by LC–MS/MS. The application of the above algorithmic procedure for the classification analysis of this sample indicates that it can be classified as *Firmicutes* → *Bacilli* → *Bacillales* → *Bacillaceae* → *Bacillus* → *B. cereus* group → *B. cereus* ATCC 14579 strain. The classification result was obtained in a high throughput fashion by analysis of an SBA matrix constructed using a set of 125 accepted peptides. The peptide sequences were characterized by a probability of correct identification greater than 99.7% (see Supporting Information, Table S2). Further analyses of these peptides indicated that they were derived from 73 predicted proteins of the *B. cereus* strain ATCC 14579. DNA sequences encoding these proteins map to 2101 codons from 124 loci. The mapping of identified ORFs on the chromosome (see gene positions in Table 1) indicates that, although some of them form clusters (mainly genes coding ribosomal proteins), the majority of coding sequences are uniformly distributed along the genome of this bacterium, and products of these genes are associated with a wide variety of cellular processes (Table 1). Only one identified sequence was associated exclusively with hypothetical proteins of *B. anthracis* Ames strains; however, subsequent manual inspection of the MS/MS spectrum indicated that this match was not correct due to mass spectral irregularities with the expected fragmentation pattern of the peptide ion.

Functions of identified proteins (Table 1) comprise a broad range of cellular processes that include metabolism, transcription, translation, membrane structure and transport, cellular signaling and so on. Because it would be beyond the scope of this report to provide a comprehensive analysis of every protein contributing to differences between an analyzed strain and reference microorganisms in the database, we have chosen to highlight only various proteins whose identity logically suggests specific testable predictions regarding phenotypic differences between the test sample and its closest relatives.

The similarity histograms (data not shown) indicated substantial taxonomic similarities of the test sample to database bacteria classified as *Bacillaceae* family. Therefore to validate the results of the initial classification, all database *Bacillaceae* bacteria, 11 strains from the genus *Bacillus* and one from *Oceanobacillus* (*O. iheyensis*), were selected for further analysis. The SBA matrix, limited to the test sample and *Bacillaceae* strains, was investigated using principal component (PCA) and hierarchical cluster analysis (HCA). These unsupervised methods were applied to discover groupings of the test strain with

Table 1. Open Reading Frames (ORFs)/Proteins Identified from the Proteome of *Bacillus cereus* ATCC 14579

	position	length			
ORF	(bp)	(codons)	strand	function	accession
BC0015	18388	298	+	Pyridoxine biosynthesis protein	30018288
BC0042	40611	95	−	Transcription state regulatory protein abrB	30018308
BC0120	112582	120	+	LSU ribosomal protein L12P (L7/L12)	30018370
BC0129	124894	396	+	Protein Translation Elongation Factor Tu (EF-TU)	30018378
BC0150	135218	147	+	LSU ribosomal protein L15P	30018399
BC0294	257497	97	+	10 kDa chaperonin GROES	30018514
BC0295	257826	545	+	60 kDa chaperonin GROEL	30018515
BC0377	359038	188	−	Alkyl hydroperoxide reductase C22	30018585
BC0479	469178	53	+	Hypothetical protein	30018685
BC0501	490217	53	−	IG hypothetical 18102	30018707
BC0548	526137	632	+	Serine protein kinase	30018736
BC0875	853236	68	−	Small acid-soluble spore protein	30019030
BC0880	858102	119	+	Putative transcriptional regulator	30019035
BC1155	1134174	489	+	Catalase	30019310
BC1216	1195214	257	+	Enoyl-[acyl-carrier-protein] reductase (fabL) (NADPH)	30019369
BC1225	1201911	173	−	2′-5′ RNA ligase	30019378
BC1279	1256509	198	+	Spore coat-associated protein N	30019430
BC1483	1440583	83	−	Ferredoxin	30019631
BC1506	1460632	87	+	Hypothetical protein	30019654
BC1509	1462383	493	+	Stage IV sporulation protein A	30019657
BC1510	1464089	115	+	DNA-binding protein HU	30019658
BC1515	1467912	149	+	Nucleoside diphosphate kinase	30019663
BC1603	1561849	66	−	Cold shock protein	30019750
BC1657	1611140	273	−	Flagellin	30019803
BC1810	1766446	403	+	Nonhemolytic enterotoxin lytic component L1	30019952
BC1984	1933068	71	+	Small acid-soluble spore protein	30020123
BC2235	2182024	87	+	Hypothetical protein	30020369
BC2287	2229799	303	+	Methylisocitrate lyase	30020419
BC2313	2261556	91	−	DNA-binding protein HU	30020445
BC2484	2453061	446	+	Propionyl-CoA carboxylase biotin-containing subunit	30020612
BC2848	2810502	541	−	Oligopeptide-binding protein oppA	30020966
BC2849	2812145	334	−	Cell wall-associated hydrolase	30020967
BC3010	2971955	443	−	Microbial collagenase	30021125
BC3539	3506145	67	−	Cold shock protein	30021641
BC3555	3525565	495	+	Aldehyde dehydrogenase	30021657
BC3584	3557857	557	−	Oligopeptide-binding protein oppA	30021685
BC3585	3559993	567	−	Oligopeptide-binding protein oppA	30021686
BC3616	3589995	908	−	Aconitate hydratase	30021715
BC3682	3648476	681	−	Transketolase	30021779
BC3728	3693603	91	+	DNA-binding protein HU	30021822
BC3770	3741218	181	−	Spore coat protein E	30021862
BC3784	3756449	82	−	IG hypothetical 16623	30021876
BC3800	3775853	200	−	Dipicolinate synthase, B chain	30021891
BC3806	3782585	90	−	SSU ribosomal protein S15P	30021897
BC3824	3801917	241	−	Protein Translation Elongation Factor Ts (EF-Ts)	30021914
BC3842	3819088	91	−	SSU ribosomal protein S16P	30021932
BC3848	3826807	81	−	Acyl carrier protein	30021938
BC3922	3906120	169	−	Prespore specific transcriptional activator rsfA	30022011
BC4049	4021492	88	−	Phosphocarrier protein HPr	30022136
BC4061	4033184	146	−	Peptidyl-prolyl cis−trans isomerase	30022148
BC4162	4125770	367	−	Leucine dehydrogenase	30022246
BC4281	4226271	85	+	IG hypothetical 17696	30022363
BC4312	4254033	612	−	Chaperone protein dnaK	30022393
BC4313	4255895	189	−	GrpE protein	30022394
BC4419	4359525	215	−	Hypothetical protein	30022500
BC4467	4413537	327	−	Stage VI sporulation protein D	30022548
BC4480	4428541	426	−	Trigger factor, ppiase	30022561
BC4521	4466210	105	−	Thioredoxin	30022599
BC4524	4468603	259	−	3-hydroxybutyryl-CoA dehydratase	30022601
BC4639	4580400	167	−	Thiol peroxidase	30022714
BC4646	4587202	66	−	Small acid-soluble spore protein	30022721
BC4801	4727816	222	−	Phage shock protein A	30022872
BC4952	4867135	79	+	NifU protein	30022992
BC4955	4868966	164	+	Low temperature requirement C protein	30022995
BC5101	5002781	513	−	Perfringolysin O precursor	30023138
BC5135	5037710	432	−	Enolase	30023172
BC5191	5090740	68	−	Cold shock protein	30023224
BC5196	5095050	436	−	*N*-acetylmuramoyl-L-alanine amidase	30023229
BC5320	5228047	166	−	PTS system, glucose-specific IIA component	30023351
BC5333	5239820	322	−	Fructose-1,6-bisphosphatase	30023364
BC5335	5242524	286	−	Fructose-bisphosphate aldolase	30023366
BC5445	5369862	209	+	Superoxide dismutase [Mn]	30023475
BC5475	5398259	171	−	Single-strand DNA binding protein	30023497

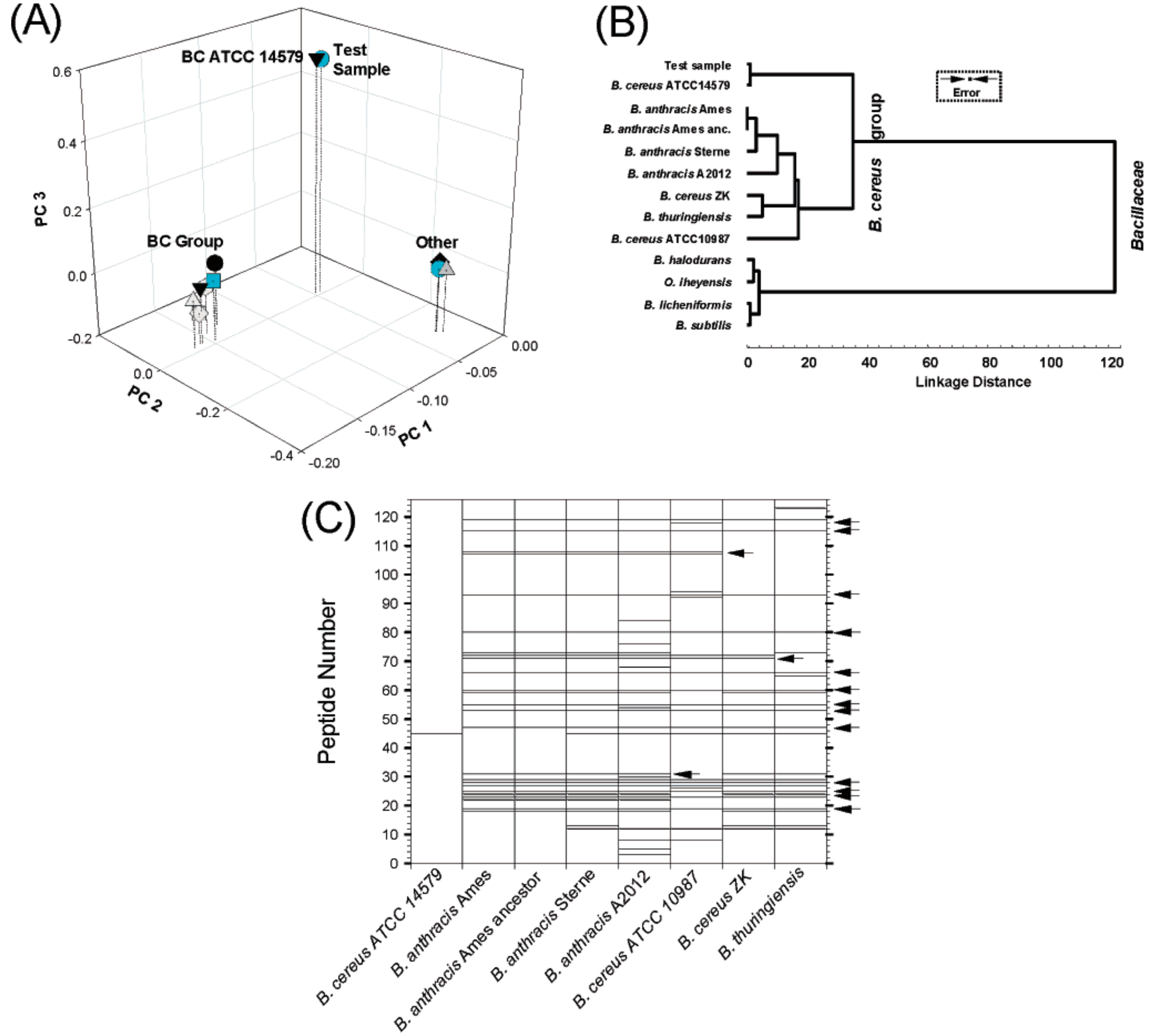

Figure 3. Analyses of a sequence-to-bacterium assignment (SBA) matrix of 125 peptide sequences assigned to the test sample (*B. cereus* ATCC 14579) and database strains from the *Bacillaceae* family. (A) Principal component analysis (PCA). Projection of PC scores in the dataspace reflecting 80.7% of the total variance included in the SBA matrix. PC1 contributes 45.6%; PC 2, 22.7%; and PC 3, 12.4% of the total variance. The 'BC group' cluster includes: *Bacillus thuringiensis* serovar konkukian, strain 97−27; *Bacillus anthracis* strains: Ames, Ames 0581, A2012, and Sterne; *Bacillus cereus* strains: ATCC 10987 and ZK. The cluster marked as 'Other' includes: *Bacillus licheniformis* ATCC 14580, *Bacillus halodurans* C-125, *Bacillus subtilis* subsp. *subtilis* strain 168, and *Oceanobacillus iheyensis* HTE831. 'BC ATCC 14579' is the database strain of *B. cereus*. (B) Dendrogram obtained after applying a hierarchical cluster analysis (HCA) to the binary SBA matrix. The dendrogram was constructed using squared Euclidean distances and the complete linkage joining method. Linkage distances are equivalent to the number of peptide sequences that differentiate organisms or their clusters. Error rate of sequence assignments: 0.003. (C) Sequence differences between the test sample and database strains from the *B. cereus* group. Each numbered row corresponds to a specific sequence of a tryptic peptide (see Supporting Information, Table S2), while the columns represent the indicated reference strains. Horizontal bars denote the absence of a sequence in a theoretical proteome of the specified organisms and these are listed in Table 2. Arrows positioned outside of the graph point to sequences that differentiate the test organism from the remaining *B. cereus* group strains, while arrows inside of the graph identify sequences with a significant discriminatory power.

reference database strains and to measure consistencies between various approaches. The obtained results were interpreted for classification and identification purposes and used to represent graphically inter-strain similarities or distances.

The rationale of the PCA method is the linear transformation of the original variables into a new set of variables called principal components (PCs). They are uncorrelated with each other and may be represented as an orthogonal system of axes, denoted PC1, PC2...PCn, that respectively correspond to a decreasing order of the amount of variance (information) in the data set. Figure 3A shows a spatial representation of inter-strain similarities or distances of analyzed bacteria in the data space of principal components PC1−PC3. This 3D plot reflects 80% of the variance in the data and provides the evidence of

Table 2. Selected Peptide Sequences Discriminating between the Test (*B. cereus* ATCC 14579) and Database *B. cereus* Group Strains

peptide no.	sequence	protein	accession
8	IEDALNSTR	60 kDa chaperonin GROEL	30018515
12	DVNEHTLEEEELPVNIEAYK	Hypothetical protein BC0479	30018685
13	RPNGTINTHPQER	IG hypothetical 18102	30018707
18	MMQGQEITEEDNQQAQEVVAR	Putative transcriptional regulator	30019035
19	ESYAAVQADTASK	Putative transcriptional regulator	30019035
22	IIMKPLEELYSAQQQA	Putative transcriptional regulator	30019035
23	AIAENEDFK	Putative transcriptional regulator	30019035
24	SFSKEEQDNLIANLTNDLK	Catalase	30019310
25	SFSKEEQDNLIANLTNDLKDVNER	Catalase	30019310
27	TPFEAQDEQLESIVNELHTIASK	2′-5′ RNA ligase	30019378
28	GDTLTAVDNDLSAWFWDEK	Spore coat-associated protein N	30019430
29	LKGDTLTAVDNDLSAWFWDEK	Spore coat-associated protein N	30019430
31	AIINNNMPTDFGSLSK	Hypothetical protein BC1506	30019654
45	KQPNFDDSSNFAK	Hypothetical protein BA 3347 [*B. anthracis* Ames]	30263256
47	TGDAALGSISNILLR	Flagellin	30019803
53	GVIENLASSVENLAELQISKDENAEDR	Hypothetical protein BC2235	30020369
55	DASAAVQSVFDTIANALQSGDK	DNA-binding protein HU	30020445
57	TGRNPQTGEEIQIAAGK	DNA-binding protein HU	30020445
59	TNTPMLLQVLEDEVFK	Propionyl-CoA carboxylase biotin-containing subunit	30020612
60	GLADSFLNDGSVAANYYVPK	Oligopeptide-binding protein oppA	30020966
65	LPLLSEDTISYR	Cell wall-associated hydrolase	30020967
66	WYQIPELFQFNSDSLK	Microbial collagenase	30021125
71	ATDQVSFLALNNVMEGLYR	Oligopeptide-binding protein oppA	30021685
72	GLTNVILNDGSTPADYLVPK	Oligopeptide-binding protein oppA	30021685
73	STDTLGAQILGNTMEGLYR	Oligopeptide-binding protein oppA	30021686
80	VGDYLANEVEGR	IG hypothetical 16623	30021876
93	EYEVPITAAQADQIVLLMK	IG hypothetical 17696	30022363
107	LVSLAEQQLGGYQK	Small acid-soluble spore protein	30022721
108	RLVSLAEQQLGGYQK	Small acid-soluble spore protein	30022721
115	SLTTSPVDISIIDSVVNR	Perfringolysin O precursor	30023138
119	VGSPQPGDLVFFQGTYK	*N*-acetylmuramoyl-L-alanine amidase	30023229
123	VIDYYFNTFDNLKDQLSK	Superoxide dismutase [Mn]	30023475

distinct clusters of points representing bacteria. It is clear from Figure 3A that all strains accepted for analysis were segregated into three groups. One contains seven database bacteria classified as *B. cereus sensu lato*, the second contains the *B. cereus* test microorganism and the *B. cereus* ATCC 14579 database strain, while the third group included the other bacilli of the *Bacillaceae* family represented by genera *Bacillus* (*B. subtilis*, *B. licheniformis*, *B. halodurans*) and *Oceanobacillus* (*O. iheyensis*).

The binary SBA matrix can be represented as a bitmap, i.e., with every sequence treated as a character that can assume only two states (1-present or 0-absent); hence every bacterium is represented as a point in the 125-dimensional space of peptides. Under these circumstances squared Euclidean (or city block) distances between bacteria are equivalent to the number of sequences that differentiate them. Such distances were used as to calculate a similarity matrix that was analyzed by HCA to reveal groupings among bacterial strains.

A few agglomerative HCA methods were applied to a distance matrix and in all cases the obtained hierarchical classifications (dendrograms) consisted of two large clusters. However, the comparison of the sub-clusters indicated that the nearest neighbor algorithm gave a much less structured classification than other methods and resulted in a phenomenon that is referred to as chaining.[24] The dendrograms obtained from the Ward's method and the furthest neighbor strategy were very similar and both consisted of identical sub-structures. The clustering tree from the application of the furthest neighbor method is shown in Figure 3B. The first cluster includes bacteria known as *B. cereus* (BC) group, while the second comprises the remaining bacteria from the *Bacillaceae* family. This classification confirms the results of the PCA (Figure 3A) and the identification of the test sample as a member of the BC group. Furthermore, it can be seen that the test sample is included in a BC group subcluster that contains only a *B. cereus* ATCC 14579 strain while the second BC group sub-cluster agglomerates other BC strains, that is, ZK and ATCC 10897.

Hence, we can conclude that the test sample is most likely identical with the *B. cereus* ATCC 14579 strain and substantially differs from *B. cereus* ATCC 10987 and *B. cereus* ZK. Although this classification tree is the outcome of phenotypic, hierarchical classification, it carries a strong phylogenetic signature that can be used to understand relationships among the genomes of the test and database bacteria. Moreover, it is noteworthy that the tree and its overall clustering topology are remarkably similar to phylogenies revealed by multilocus enzyme electrophoresis and nucleotide sequence analysis of many chromosomal genes.[8]

To further clarify these results at the sequence level it is useful to perform direct comparison between the test sample and other related species. Although 124 of the 125 accepted sequences were matched to the theoretical proteome of *B. cereus* strain ATCC 14579, 75−80% were shared with the remaining members of the *B. cereus* group, while only 4−5% of peptides matched to other *Bacillaceae* strains. To present these sequence differences (SDs) in a compact format, the algorithm calculates a matrix of differences between the test sample and database bacteria and generates a series of horizontal bar graphs as shown in Figure 3C. Each bar in this plot indicates the lack of a peptide in a bacterial proteome in comparison to the test strain. Note that, for our discussion, the phrase 'sequence differences' has been used in the generic

Table 3. Proteins Identified from the Proteome of *E. coli* K-12 at Different Growth Phases

		growth phase	
functional assignment	early exponential	late exponential	stationary
Information storage and processing			
Translation, ribosomal structure and biogenesis	RpsT, RpsB, Tsf, RpsA, AspS, RpsP, RpsU, InfB, RplQ, RpsK, RplR, RplF, RplE, RplX, RpmC, RplP, RplD, TufA, FusA, RpsG, RpmG, RpmB, RplA, RplJ, RplL, RpsF, RplI, PrfB, RplK	Tsf, Frr, RpsA, RpsV, b1809, YfiA, RplE, RpmC, RplP, RplD, TufA, RplL, RpsF, YjgF	Tsf, Frr, RpsA, YfiA, RpmC, TufA, FusA, RplL
Transcription	MarA, NusA, RpoC, RpoZ	CspC, RpoZ	CspC, RpoZ
DNA replication, recombination and repair	HupA; HupB	XseB, Dps, HupA, HupB	XseB, Dps, HupA, HupB
Cellular processes			
Post-translational modification, protein turnover, chaperones	DnaK, Tig, AhpC, GrpE, KpA, TrxA, MopA	DnaK, Tig, AhpC, Tpx, OsmC, YeaA, FkpA, TrxA, MopB	DnaK, Tpx, OsmC, FkpA, GrxC, TrxA, MopA, MopB
Cell envelope biogenesis, outer membrane	HlpA	HlpA	HlpA, OmpA, b1377, OmpC
Inorganic ion transport and metabolism		SodB, Ftn, yffB, b2495	ChaB, ChaC, SodB, KatE, Bfr, PhnA
Signal transduction mechanisms		YbdQ uspA	YbdQ
Metabolism			
Energy production and conversion	AceE, AceF, PflB, AtpA		LpdA, AceA
Carbohydrate transport and metabolism	GapA, PtsH, Crr, Eno, Fba, Pgk, RbsB, DeoB	GapA, GatA, MglB, Crr, Eno, RbsB, MalE	GapA, MglB, PtsI, Crr, Eno, Fba, Pgk, RbsB
Amino acid transport and metabolism	DapD	GlnH, ArtI, PotD, OppA, TnaA	GlnH, ArtI, OppA, GadB, TnaA
Nucleotide transport and metabolism	Adk	YcfF	
Lipid metabolism	AcpP	AcpP, Psd	AcpP, Psd
Poorly characterized			
General function prediction only	b0753, Hns, YfiD, StpA	Hns, YgaD, OsmY	YahK, YbeL, b0753, WrbA, MsyB, Hns, YedU, OsmY
Function unknown	YccJ, YqjD, HdeB	YkfE, YahO, YccJ, YebF, YggE, YgiW, YqjD, HdeB, YihI, YiiU, YjbJ	YkfE, yccJ, b1586, b1836, YebF, b1953, YgaU, YgiW, YqjD, HdeB, HdeA, YiiU, YjbJ

sense to reflect the status of various genome traceable amino acid sequences, and not in the specific evolutionary sense to imply that the differential presence of a particular sequence among unknown and reference species arose through gene loss or acquisition.

The sequences marked with arrows in Figure 3C (positioned outside of the graph) originate from 15 different proteins listed in Table 2, and indicate peptides that differentiate the *B. cereus* test organism from the remaining *B. cereus* group strains. Some of these proteins are associated with cell wall and information processing while the other diagnostic sequences are derived from proteins conferring pathogenicity for *B. cereus*, e.g., those annotated as microbial collagenase and perfringolysin O precursor.[25]

However, it should be pointed out that, although the presence of a given sequence in other proteomes strongly suggests the potential to express a similar protein, the absence of a sequence in any proteome may reflect a broad range of genomic modifications. They may include a single nucleotide polymorphism caused by nonsynonymous point mutations in the investigated gene segment, indels or the lack of a particular gene due to loss through genome rearrangements. For instance, TGDAALGSISNILLR (#47) mapped to ORF BC1656 is exclusively associated with the proteome of *B. cereus* ATCC 14579 and its presence strongly suggests that the motility of the *B. cereus* test bacterium is due to its origin from a flagellin (Table 2). However, the absence of this sequence in the remaining proteomes may reflect: (a) the lack of functional flagella, and hence the motility of cells (e.g., *B. anthracis*), or (b) the polymorphism of genes encoding this protein (*B. cereus* and *B. thuringiensis*) caused by nonsynonymous mutations in this particular segment of their genomes.

Effect of Growth Conditions on Bacterial Classification. The above examples indicate that a relatively small number of peptide sequences can provide meaningful taxonomic information about the analyzed bacteria. However, relationships among bacteria revealed by a hierarchical clustering of strains may be potentially altered by peptides from proteins associated with different phylogenetic profiles. To test this possibility, the *E. coli* MG1665 strain was grown to three different phases, and the tryptic peptides from the extracts of the individual samples were analyzed by LC−ESI MS/MS. The results obtained allowed the identification of 117 to 129 peptides with a probability of correct sequence assignments higher than 98%, and error rates in each dataset below 0.5% (see Supporting Information, Tables S3−S5). These peptides were mapped to sequences of diverse proteins expressed under different growth stages of the analyzed bacterium (Table 3). The chromosomal locations of ORFs encoding these proteins are presented in Figure 4A-C. In this figure the transcription origin on the circular chromosome is placed in the zero position, while numbers of codons assigned to a given ORF are shown as negative or positive to indicate

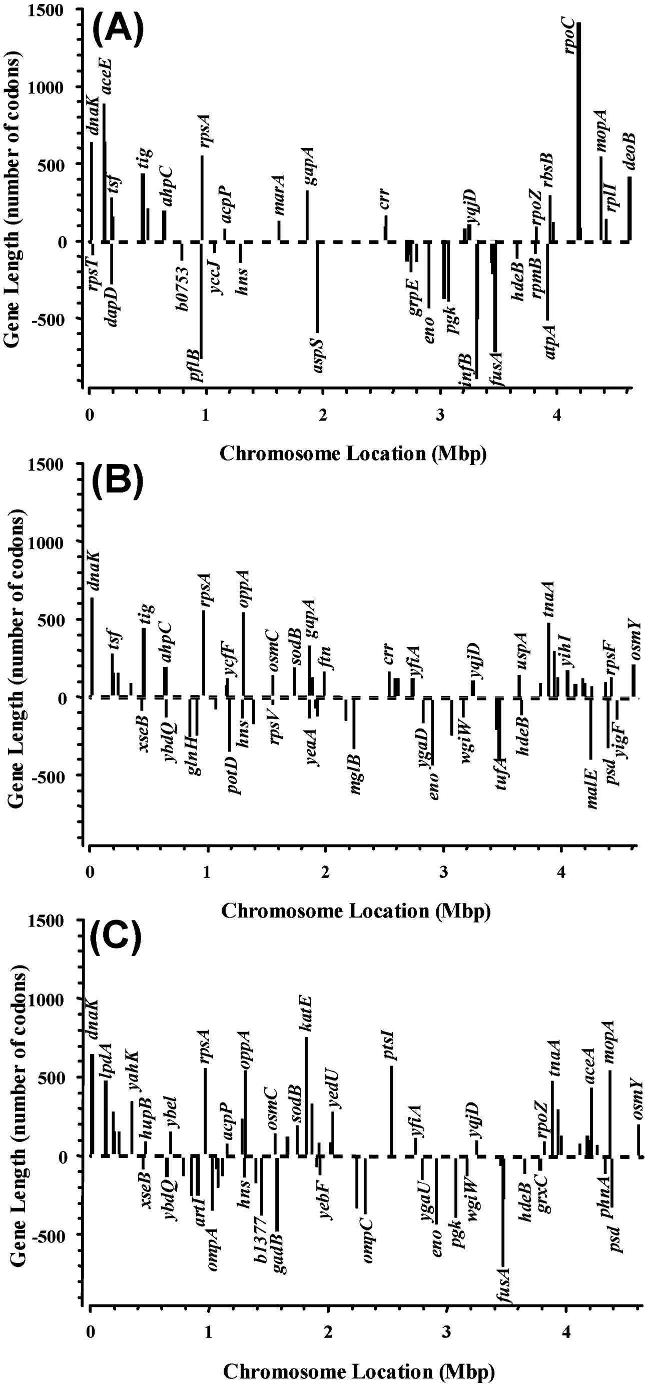

Figure 4. Locations of genes encoding proteins that contributed peptides identified during LC–MS/MS analysis of *E. coli* K-12 samples harvested during: (A) early logarithmic, (B) late logarithmic, and (C) stationary growth phases. The horizontal axis represents chromosomal locations,[26] and the vertical bars indicate the lengths of ORFs shown as the number of codons; where positive or negative values refer to the positive or negative DNA coding strand, respectively.

that they are coded by a negative or positive DNA strand, respectively.[26] The analysis of Figure 4A-C indicates that ORFs encoding identified proteins (see Supporting Information, Tables S6–S8) are scattered along the whole circular chromosome. However, 55% of the peptides detected from the beginning of the exponential growth and only 25% of the peptides from the stationary phase are from the proteins involved in information storage and processing (Figure 5). As a consequence, the contribution of other proteins increases with the cultivation time.

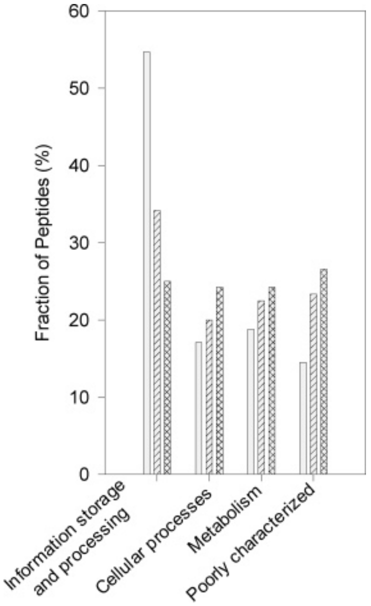

Figure 5. Functional distribution of *E. coli* proteins obtained from different growth phases for the *E. coli* strain K-12, and identified by sequencing peptides accepted for the classification analyses: open bars – denote the early logarithmic phase; hatched bars – late logarithmic, and crosshatched bars – the stationary growth phase.

The results from taxonomy-based classification (e.g., Figure 2) and cluster analysis (Figure 6) indicate that in all cases the analyzed samples are most similar to the database *Escherichia/Shigella/Salmonella* (ESS) strains. Furthermore, the dendrograms displayed in Figure 6 indicate that the test samples are identical to a database strain *E. coli* K-12 (Figure 6A, B) or differ by only one peptide sequence (Figure 6C). Taking into account the error rate (0.4%) and the number of accepted peptides (129), it is obvious that the analyzed *E. coli* K-12 sample is indistinguishable from the database strain K-12. Moreover, in each case, the cluster of the *E. coli* K-12 sample with the database K-12 strain can be linked with other *E. coli* and *S. flexneri* strains into a group that is substantially different in comparison to the next closest cluster of *Salmonella* strains. These results are consistent with reports based on whole genome comparisons of both species[27,28] suggesting that *Shigella* is phylogenetically indistinguishable from *E. coli* strains.

In all three cases shown in Figure 6, the relative distances between clusters remain stable. Nevertheless, among ESS strains, sequences of proteins involved in informational processes, e.g., ribosomal proteins and factors involved in protein biosythesis, as well as co- and post-translational chaperones, are more conserved in comparison to other proteins. Therefore, the substantially higher contribution of such proteins during the beginning of the exponential growth phase (Figure 5) translates into shorter absolute linkage distances, as documented in Figure 6A, in comparison to those shown in Figure 6B and C.

During the exponential and stationary growth phases, the contribution of proteins associated with metabolite transport and stress regulation increases (Table 3), and a substantial fraction of SD's originate from a set of gene products that are

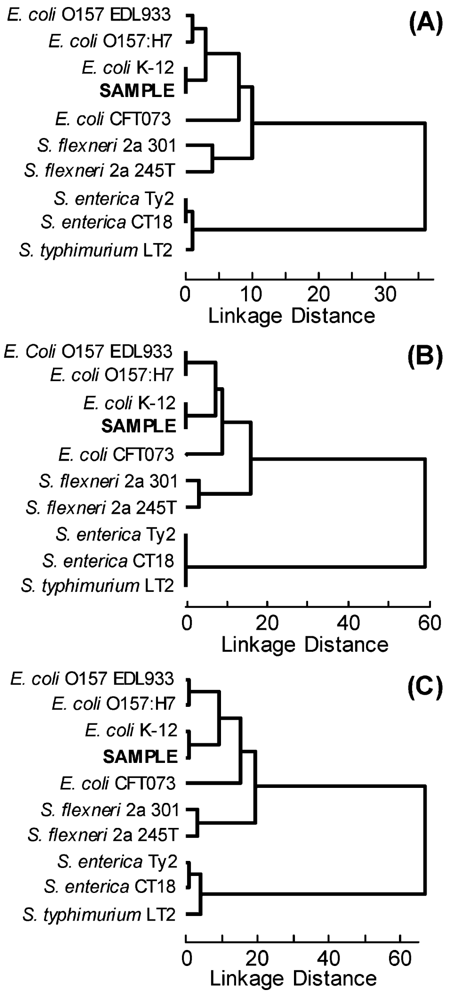

Figure 6. Cluster analysis of sequence-to-bacterium assignment matrixes for *E. coli* K-12 samples taken from: (A) early logarithmic, (B) late logarithmic, and (C) stationary growth phases (only fragments of dendrograms, limited to the *Escherichia, Salmonella* and *Shigella* database strains, are presented). The dendrograms were constructed using the complete linkage (furthest neighbor) joining method and squared Euclidean distances in the multidimensional space of: (A) 117, (B) 120, and (C) 129 peptide sequences. In these tree topologies, the linkage distances are equivalent to the number of peptide sequences that differentiate organisms or their clusters. Error rates for sequence assignments: (A) 0.005; (B) 0.005, (C) 0.004.

under the transcriptional control of the alternative sigma factor σ^S that accumulates in response to stress conditions.[29] However, the lower discriminatory power of peptides identified during the onset of the exponential growth, in comparison to those derived from later growth phases, is also due to the use of a mass spectrometer in the so-called 'data dependent' mode of operation. Under such circumstances, the mass spectrometer preferentially detects peptides derived from proteins of highest abundance, and therefore the proteomic analyses of whole-cell lysates are biased toward highly abundant proteins. Nevertheless, a comparison of ORF sequences from CFT073, O157:H7 EDL933, and K-12 strains[30] indicate that they share 2,996 genes, representing the backbone of the *E. coli* chromosome, and most of the remaining sequences specific for pathogenic organisms are horizontally transferred foreign DNAs. Hence, peptide sequences derived from the K-12 cells can be used to probe only a chromosomal backbone of the pathogenic strains, and as a consequence, differences between *E. coli* K-12 and pathogenic strains revealed by the MS-based proteomic approach are relatively small.

The above results demonstrate that the analysis of bacteria harvested under different growth stages produces no major changes in the overall classification, as documented in Figure 6. This confirms the robustness of phylogenetic peptide profiles associated with identical sequences found across diverse bacteria for classification of strains, because different growth conditions only affect the internal branching of the furthest relatives of the analyzed species. This method of bacterial classification can be used to explore subtle proteome differences between the *E. coli* K-12 sample and the related *Enterobacteriaceae* database strains, and to infer the functionality of some of poorly characterized gene products.

Conclusions

We have described a bacterial classification method based on the peptide sequence information generated from LC–ESI MS/MS analysis of a bacterial protein digest. This method may function as a strong complement to the alternative approaches of comparing microbial genomes based on DNA sequencing or microarray hybridization techniques.[31] The results shown in this work indicate that the proteomics method can provide a relatively rapid classification of the analyzed bacteria, even though data analysis was based on a small number of peptide sequences that reflect 21–99 base long segments of ORFs scattered throughout the genome (chromosome) and cumulatively represent about 6000 bases.

A major advantage of the proteomics method is that no prior knowledge is required of the sample; although, it is obvious that taxa under-represented in the database will not provide a sufficiently high resolution to accurately classify the unknown sample to the strain level. However, with more than seven hundred sequencing projects in progress and the increasing rate of sequencing of bacterial genomes, this method should be applicable to a wide range of microorganisms in the future. In addition, advances in separation and mass spectrometric technologies will allow rapid sequencing of a great number of peptides and proteins soon. Thus we envisage that the proteomics method will become useful in many bacterial classification and identification applications.

Supporting Information Available: Taxonomy of database bacteria with fully sequenced genomes used in this study [Table S1]. Peptide sequences accepted for the classification analysis of a test sample (*B. cereus* ATCC 14579) [Table S2]. Peptides identified from the analysis of *E. coli* K-12 proteome expressed at the early logarithmic [Table 3], the late logarithmic [Table 4], and at the stationary growth phase [Table 5]. Open reading frames/proteins identified from the analysis of *E. coli* K-12 proteome at the onset of the exponential growth phase [Table 6], at the late exponential growth phase [Table 7], and at the stationary growth phase [Table 8]. This material is available free at http://pubs.acs.org.

References

(1) Rotz, L. D.; Khan, A. S.; Lillibridge, S. R.; Ostroff, S. M.; Hughes, J. M. *Emerg. Infect. Dis.* **2002**, *8*, 225–230.

(2) Rossello-Mora, R.; Amann, R. *FEMS Microbiol. Rev.* **2001**, *25*, 39–67.

(3) Pershing, D. H.; Tenover, F. C.; Versalovic, J.; Tang, Y.-W.; Unger, E. R.; Relman, D. A.; White, T. J. *Molecular Microbiology: Diagnostic Principles and Practice*, American Society for Microbiology, Washington, D. C., 2003.

(4) Woese, C. R.; Stackebrandt, E.; Macke, T. J.; Fox, G. E. *Syst. Appl. Microbiol.* **1985**, *6*, 143−151.

(5) Maiden, M. C. J.; Bygraves, J. A.; Feil, E.; Morelli, G.; Russell, J. E.; Urwin, R.; Zhang, Q.; Zhou, J.; Zurth, K.; Caugant, D. A.; Feavers, I. M.; Achtman, M.; Spratt, B. G. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 3140−3145.

(6) Selander, R. K.; Caugant, D. A.; Ochman, H.; Musser, J. M.; Gilmour, M. N.; Whittam, T. S. *Appl. Environ. Microbiol.* **1986**, *51*, 873−84.

(7) Scortichini, M.; Natalini, E.; Angelucci, L. *Microbiology* **2003**, *149*, 2891−2900.

(8) Helgason, E.; Økstad, O. A.; Caugant, D. A.; Johansen, H. A.; Fouet, A.; Mock, M.; Hegna, I.; Kolstø, A−B. *Appl. Environ. Microbiol.* **2000**, *66*, 2627−2630.

(9) Chen, W.; Laidig, K. E.; Park, Y.; Park, K.; Yates, J. R., III.; Lamont, R. J.; Hackett, M. *Analyst* **2001**, *126*, 52−57.

(10) Harris, W. A.; Reilly, J. P. *Anal. Chem.* **2002**, *74*, 4410−4416.

(11) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242−247.

(12) Wolters, D. A.; Washburn, M. P.; Yates, J. R., 3rd. *Anal. Chem.* **2001**, *73*, 5683−5690.

(13) Yao, Z.-P.; Alfonso, C.; Fenselau, C. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 1953−1956.

(14) Warscheid, B.; Fenselau, C. *Anal. Chem.* **2003**, *75,* 5618−5627.

(15) Warscheid, B.; Fenselau, C. *Proteomics* **2004**, *4*, 2877−2892.

(16) Pribil, P. A.; Patton, E.; Black, G.; Doroshenko, V.; Fenselau, C. *J. Mass Spectrom.* **2005**, *40*, 464−474.

(17) Hu, A.; Tsai, P.-J.; Ho, Y.-P. *Anal. Chem.* **2005**, *77,* 1488−1495.

(18) VerBerkmoes, N. C.; Hervey, W. J.; Shah, M.; Land, M.; Hauser, L.; Larimer, F. W.; Van Berkel, G. J.; Goeringer, D. E. *Anal. Chem.* **2005**, *77*, 923−932.

(19) Dworzanski, J. P.; Snyder, A. P.; Chen, R.; Zhang, H.; Wishart, D.; Li, L. *Anal. Chem.* **2004**, *76*, 2355−2366.

(20) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(21) Eng, J. K.; McCormack, A. L.; Yates, J. R. 3rd. *J. Am. Soc. Mass Spectrom.* **1994**, *5,* 976−989.

(22) Keller, A.; Nesvizhskii, A. I.; Kolker, I.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.

(23) Wheeler D. L.; Chappey, C.; Lash, A. E.; Leipe, D. D.; Madden, T. L.; Schuler, G. D.; Tatusova, T. A.; Rapp, B. A. *Nucleic Acids Res.* **2000**, *28*, 10−14.

(24) Everitt, B. S. *Cluster Analysis*, 3rd ed.; Edward Arnold: London, 1993.

(25) Ivanova, N.; Sorokin, A.; Anderson, I.; Galleron, N.; Candelon, B.; Kapatral, V.; Bhattacharyya, A.; Reznik, G.; Mikhailova, N.; Lapidus, A.; Chu, L.; Mazur, M.; Goltsman, E.; Larsen, N.; D'Souza, M.; Walunas, T.; Grechkin, Y.; Pusch, G.; Haselkorn, R.; Fonstein, M.; Ehrlich, D. S. D.; Overbeek, R.; Kyrpides, N. *Nature* **2003**, *423*, 87−91.

(26) Blattner, F. R.; Plunkett, G., 3rd.; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y. *Science* **1997**, *277*, 1453−1462.

(27) Jin, Q.; Yuan, Z.; Xu, J.; Wang, Y.; Shen, Y.; Lu, W.; Wang, J.; Liu, H.; Yang, J.; Yang, F.; Zhang, X.; Zhang, J.; Yang, G.; Wu, H.; Qu, D.; Dong, J.; Sun, L.; Xue, Y.; Zhao, A.; Gao, Y.; Zhu, J.; Kan, B.; Ding, K.; Chen, S.; Cheng, H.; Yao, Z.; He, B.; Chen, R.; Ma, D.; Qiang, B.; Wen, Y.; Hou, Y.; Yu, J. *Nucleic Acids Res.* **2002**, *30*, 4432−4441.

(28) Wei, J.; Goldberg, M. B.; Burland, V.; Venkatesan, M. M.; Deng, W.; Fournier, G.; Mayhew, G. F.; Plunkett, G., III.; Rose, D. J.; Darling, A.; Mau, B.; Perna, N. T.; Payne, S. M.; Runyen-Janecky, L. J.; Zhou, S.; Schwartz, D. C.; Blattner, F. R. *Infect. Immun.* **2003**, *71*, 2775−86.

(29) Hengge-Aronis, R. *Microbiol. Mol. Biol. Rev.* **2002**, *66*, 373−395.

(30) Welch, R. A.; Burland, V.; Plunkett, G., 3rd; Redford, P.; Roesch, P.; Rasko, D.; Buckles, E. L.; Liou, S. R.; Boutin, A.; Hackett, J.; Stroud, D.; Mayhew, G. F.; Rose, D. J.; Zhou, S.; Schwartz, D. C.; Perna, N. T.; Mobley, H. L.; Donnenberg, M. S.; Blattner, F. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 17020−17024.

(31) Fukiya, S.; Mizoguchi, H.; Tobe, T.; Mori, H. *J. Bacteriol.* **2004**, *186*, 3911−3921.

PR050294T