



A protein database constructed from low-coverage genomic sequence of *Bacillus megaterium* and its use for accelerated proteomic analysis

Jibin Sun^{a,1}, Wei Wang^{b,1}, Claudia Hundertmark^c, An-Ping Zeng^{a,*},
Dieter Jahn^c, Wolf-Dieter Deckwer^b

^a Group Systems Biology, GBF-German Research Center for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany

^b Group of TU-BCE, GBF-German Research Center for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany

^c Institute of Microbiology, Technical University of Braunschweig, Germany

Received 26 October 2005; received in revised form 27 December 2005; accepted 13 January 2006

Abstract

Peptide mass fingerprint (PMF) matching is a high-throughput method used for protein spot identification in connection with two-dimensional gel electrophoresis (2DE). However, the success of PMF matching largely depends on whether the proteins to be identified exist in the database searched. Consequently, it is often necessary to apply other more sophisticated but also time-consuming technologies to generate sequence-tags for definitive protein identification. On the other hand, modern sequencing technologies are generating a large quantity of DNA sequences, first in unfinished form or with low genome coverage due to the time-consuming and thus limiting steps of finishing and annotation. We recently started to sequence the genome of *Bacillus megaterium* DSM 319, a bacterium of industrial interest. In this study, we demonstrate that a protein database generated from merely three-fold coverage, unfinished genomic sequences of this bacterium allows a fast and reliable protein spot identification solely based on PMF from high-throughput MALDI-TOF MS analysis. We further show that the strain-specific protein database from low coverage genomic sequence greatly outperforms the commonly used cross-species databases constructed from 13 completely sequenced *Bacillus* strains for protein spot identification via PMF.

© 2006 Elsevier B.V. All rights reserved.

Keywords: *Bacillus megaterium*; Protein database; Low-coverage; Genomic sequence; Peptide mass fingerprinting; Proteomics

1. Introduction

Proteomic study by combining two-dimensional gel electrophoresis (2-DE) and mass spectrometry has been well established during the last decade. Benefited from the rapid progress of mass spectrometric

* Corresponding author at: GBF, SBIO, Mascheroder Weg 1, D-38124 Braunschweig, Germany. Tel.: +49 531 6181 188; fax: +49 531 6181 751.

E-mail address: AZE@GBF.de (A.-P. Zeng).

¹ These authors contributed equally.

technology, protein identification via matrix-assisted laser desorption/ionisation-time of flight mass spectrometry (MALDI-TOF MS) data, also called peptide mass fingerprinting (PMF) matching, has become a real high-throughput method and is currently widely used. However, PMF matching largely depends on the presence of the proteins in the database searched. Since protein sequences of the specific organism under study are not always available, it is often necessary to resort to cross-species protein identification (Barrett et al., 2005; Mathesius et al., 2002; Cordwell et al., 1995; Kim et al., 2004). This means that homologous proteins from other organisms are used for spot identification. Depending on the organisms studied, the possibility and reliability of cross-species spot identification via PMF matching varies very much. In most cases, the results of cross-species protein identification are not satisfying. Consequently, it is necessary to apply other more sophisticated but also time-consuming technologies, such as sequence-tags using peptide fragment information obtained from nanoelectrospray ionisation quadrupole-time-of-flight tandem mass spectrometry (ESI-QqTOF MS/MS) or from MALDI two-stage time-of-flight tandem mass spectrometry (MALDI-TOF/TOF MS/MS), etc. to ultimately identify the spots on 2-D gels (Barrett et al., 2005; Marrero et al., 2004; Kim et al., 2004).

On the other hand, modern sequencing technologies and projects are generating a large quantity of DNA sequences in an ever astonishingly faster speed (<http://www3.ebi.ac.uk/Services/DBStats/>). Although the finishing step and annotation process are still time consuming, the generation of low coverage of genomic sequences has become very fast and relatively inexpensive. Many early-stage low-coverage genomic sequences are actually publicly accessible. As reported by NCBI (state of October 2005), 263 microbial genomes were completed and 520 are still in progress. The situation is even more drastic for eukaryotes: only 18 of the 266 eukaryotic genome sequencing projects were complete. How to make use of these low-coverage unfinished genomic sequences to accelerate proteomic study is still a challenge. Previously, we showed how high-quality and nearly finished raw genomic sequences of *Klebsiella pneumonia* can be used to help protein identification (Wang et al., 2003). To our knowledge, no work has been so far reported on the use of even lower-coverage unfinished genomic sequences in proteomic analysis.

The gram-positive bacterium *Bacillus megaterium* is a promising host for the production of diverse heterologous proteins and vitamins due to its intrinsic favourable properties such as low protease activity and high secretion capability (Vary, 1994; Malten et al., 2005a,b). Recently, we investigated *B. megaterium* as one of the preferred protein production organisms, for which the protein expression profiling in fermentation processes was studied by 2-DE analysis (Wang et al., 2005). One-hundred sixty-seven spots were identified via peptide sequencing using peptide fragment information obtained from ESI-QqTOF MS/MS. To better understand the metabolism and its genetic regulation at a genome level we recently started to sequence the genome of the *B. megaterium* strain DSM 319. In this study, we generate a protein database for *B. megaterium* based on genomic sequence of a 3.47-fold coverage. We further demonstrate that this protein database allows a fast and reliable protein spot identification solely based on peptide mass fingerprinting from high throughput MALDI-TOF MS experiments. A comparison with a reference protein database generated from other 13 sequenced bacilli strains shows that the use of this strain-specific database greatly outperforms the commonly used cross-species spot identification.

2. Material and methods

2.1. Genomic sequence

The genome of the bacterium strain *B. megaterium* DSM 319 was sequenced using whole genome shotgun approach with 3.47-fold genomic coverage in cooperation with GATC Biotech AG, Germany. The genomic sequence was assembled into 1025 contigs (>1 kb) (status: November 2004). The size of the largest contig is 47,077 bp and the average size is 4269 bp. The total size of these sequences is 4,375,263 bp in comparison to the estimated genome size of 4.8 Mbp.

2.2. Construction of protein database from unfinished genomic sequence

In our previous work (Sun and Zeng, 2004), a homolog-based method called “IdentiCS” was devel-

oped to identify protein coding sequences from unannotated low-coverage bacterial genome sequence. This method was demonstrated to be well tolerant to sequence errors that can cause major problems such as unusual stops (fragmented CDSs) and transcription frame shifts (wrong protein sequences) in gene prediction with many other gene-finding programs. Therefore, it is particularly well suited for the analysis of the low-coverage, unfinished genomic sequence of *B. megaterium*. All sequences from Swiss-Prot Release 45.0 of 25 October 2004 and TrEMBL Release 28.0 of 25 October 2004, obtained from the ftp server of the Swiss Institute of Bioinformatics (SIB) (<ftp://ftp.expasy.org/databases/uniprot>), were used as reference sequences in “IdentiCS” to annotate the unfinished genomic sequences. 4845 potential CDSs were found with a cut off of BLAST (program tblastn, Altschul et al., 1997) E-value at $1e^{-5}$ and annotated. These proteins form a database for our local Mascot server with the name “bmeg”.

2.3. Improved protein database “bmegMEC”

Like other homolog-based CDS prediction programs, “IdentiCS” cannot predict organism-specific CDSs that have no similarities to any proteins in the available protein databases. To complement this point, another computer program “CRITICA” (Badger and Olsen, 1999) was also used for CDS prediction from the raw genomic sequences. This program combines comparative analysis with more common non-comparative methods such as dicodon bias. It was considered to be a method well suited for analyzing novel genomes. 335 CDSs that were additionally found by CRITICA were transferred to the database “bmeg”, resulting in an extended database.

The protein database of *B. megaterium* was further improved by a treatment of partial proteins locating at the end of contigs. Because of the incompleteness of the genomic sequence, some proteins were separated as two fragments on the ends of two contigs (Fig. 1A). As a result, it is difficult to identify the corresponding protein spots from 2DE by comparing with either of the two fragments using MALDI-TOF MS data, i.e. none of these two fragments can give significant scores for an unambiguous identification. As described more in detail below, merging two fragments together can result in an increase of the Mascot score

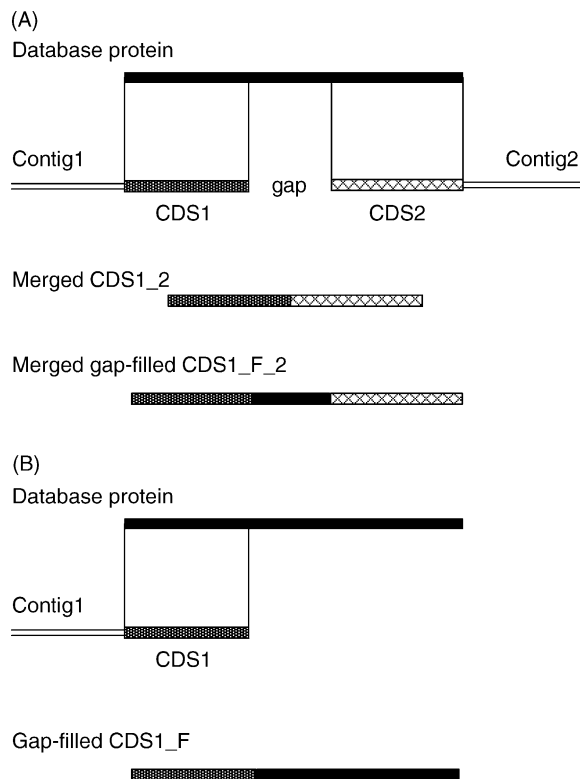


Fig. 1. Handling the incompleteness of genomic sequences. Peptide fragments locating at the end of different contigs were merged together to form a more complete protein by using its most identical homologous protein of known sequence as a key. By incorporation of the corresponding sequence of the key protein, the “missing” sequences can be filled within a merged protein (gap) (A) or at end of an incomplete protein (B) to form a hybrid protein.

and therefore facilitate the identification of the related proteins.

The scoring algorithm of the commercial software MASCOT is not published. However, it was mentioned that the MASCOT score is a probability-based MOWSE score (http://www.matrixscience.com/help/scoring_help.html). MOWSE score is calculated as:

$$\text{score} = \frac{50,000}{M_{\text{prot}} \cdot \prod_n m_{i,j}}$$

where M_{prot} is the molecular weight of the corresponding protein and the product term $\prod_n m_{i,j}$ is calculated from the MOWSE factor elements $m_{i,j}$ for each match between the experimental data and theoretical pep-

ptide masses calculated from the protein. The merge of two end fragments increases both the molecular weight M_{prot} and the total number n of peptide mass matches. Since the value of $m_{i,j}$ corresponding to each match is usually much smaller than one (Pappin et al., 1993), the overall effect of merging can lead to increases of MOWSE scores in most cases.

From the original prediction of IdentiCS, 304 CDSs are identified as partial fragments and therefore merged into 152 more complete proteins. Finally an improved database “bmegMEC” (improved bmeg database by Merging Ends and Critica) was created that includes 4541 proteins directly from the original prediction of IdentiCS, 152 merged proteins and 335 proteins from the prediction of CRITICA.

2.4. Reference protein databases used for comparison

Complete genome sequences of other 13 bacilli strains are currently available. It was desired to test whether these genome sequences can be useful for spot identification in our 2DE study with *B. megaterium*. The protein sequences of these organisms were downloaded from KEGG (<ftp://ftp.genome.jp/pub/kegg/genomes/>) and formatted as a reference database “13bacilli”.

The non-redundant database NCBI nr from NCBI is a frequently used protein database for protein identification. The taxonomic division “Bacteria” of NCBI nr was also used here as a reference database. The database “13bacilli” is a subset of this division. It should be mentioned that 572 proteins from the species *B. megaterium* (279 non-redundant ones with less than 95% identity to each other) are also included in this division.

2.5. Preparing protein spots from 2-DE experiments for MALDI-TOF MS analysis

In a previous works we demonstrated that *B. megaterium* is a promising host for the production of a heterologous dextranucrase (Malten et al., 2005a,b; Wang et al., 2005), in which *B. megaterium* strain MS941, developed from the *B. megaterium* wild-type strain DSM 319 by gene replacement of the major extracellular protease (Wittchen and Meinhardt, 1995), was transformed with the plasmid pMM1520*dsrS*,

which carries the gene of a dextranucrase from *Leuconostoc mesenteroides*. To further characterize the physiological responses of this strain to a high cell density cultivation (HCDC) process as well as to the induction of the heterologous protein production, 2-DE experiments were carried out with cell samples taken from a HCDC cultivation at different times. For MALDI-TOF MS analysis protein spots excised from 2-DE gels were subjected to tryptic digestion according to a method described previously (Wang et al., 2005, 2003) with slight modifications. Briefly, protein spots were washed twice with Milli-Q water, dehydrated with acetonitrile, digested overnight with trypsin (sequencing grade modified, Promega Corp.) at 37 °C, and desalted using the Montage ZipPlateC18 (Millipore Corp.) that can parallel desalt 96 digested protein samples. Subsequently, the tryptic peptides obtained were analyzed by MALDI-TOF MS with a Bruker Ultraflex time-of-flight mass spectrometer (Bruker Daltonics GmbH, Germany) as described before (Wang et al., 2003).

2.6. Protein spot identification

Peptide masses obtained from MALDI-TOF MS analysis were used for protein identification by PMF using the strain-specific protein database “bmeg” as well as “bmegMEC”. In addition, two reference protein databases described above, namely the database “13bacilli” and the non-redundant database NCBI nr with restriction to taxonomic division “Bacteria”, were also used for comparison. The reliability of protein spot identification was validated using the MASCOT score, pI value, molecular weight and the ESI-QqTOF results from Wang et al. (2005).

3. Results

3.1. Spot identification with the databases “bmeg” and “bmegMEC”

Three hundred forty-four protein spots were cut and analyzed by MALDI-TOF MS. 50.9% (175 spots) of them were identified by PMF using at least one of the databases mentioned in Section 2 (Fig. 2). With the exception of six spots, all the other spots were identified by applying the database “bmegMEC”, indi-

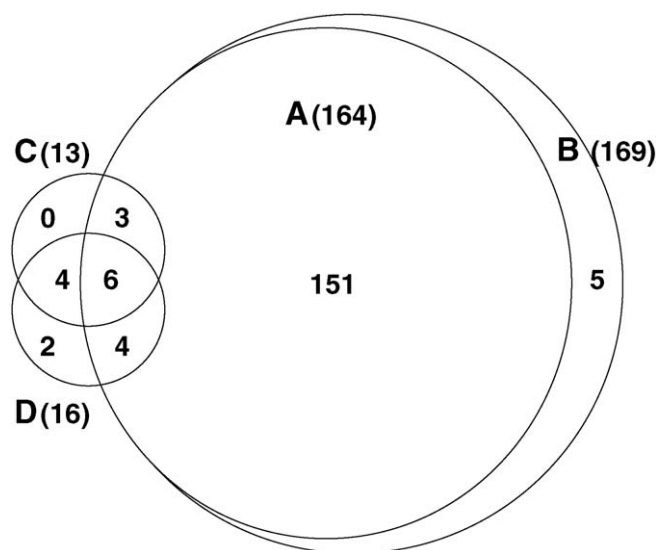


Fig. 2. Comparison of the number of spots identified by peptide mass fingerprint using different protein databases described in this work: A, bmeg; B, bmegMEC; C, 13bacilli; D, NCBIInr-Bacteria. All identified spots have a MASCOT score greater than the significance level corresponding to $p < 0.05$, being 49, 50, 61 and 73 for the four databases respectively. The numbers in parentheses show the total number of spots identified by the corresponding database. Other numbers show the number of spots identified by different combinations. Only true positive identifications are counted.

cating the usefulness of the improved strain-specific database. All protein spots identified with the database “bmeg” could also be identified with “bmegMEC”. The identification score of nine spots was, however, higher with “bmegMEC” than with “bmeg”. This is because all corresponding proteins in the “bmegMEC” but not in the “bmeg” database were merged from their partial fragments. As expected, the merge influenced the overall PMF matching scores positively. But there are also three spots, whose MASCOT score with “bmeg” was slightly higher than that with “bmegMEC”. We checked and confirmed that all the three proteins in “bmegMEC” were merged from a larger and a much smaller peptide fragments. By chance, none of the peptide masses determined by MALDI-TOF MS matched to the smaller peptide fragments (Fig. 3.). According to the score-calculating formula, the term M_{prot} became bigger while the term $\prod m_{i,j}$ remained unchanged, leading to the results of lower MASCOT score with “bmegMEC”. There are five spots that can only be identified with “bmegMEC” but not with “bmeg”. All of them were CDSs additionally predicted by the CDS prediction program CRITICA.

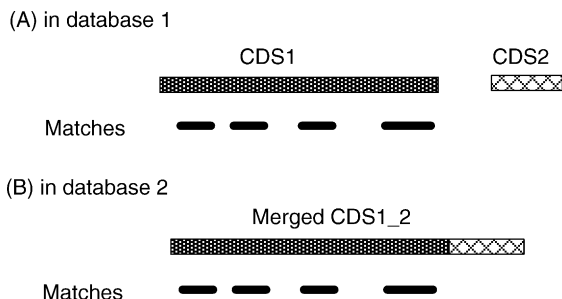


Fig. 3. Merging two protein fragments to form a protein of more complete sequence can in certain cases cause decreased MASCOT scores.

3.2. Cross-species protein spot identification

The presence of complete genome and proteome sequences for as many as 13 bacilli in the public database has been considered as a good starting point for the functional genomic analysis of *B. megaterium*. However, to our surprise, this did not help much for protein spot identification using peptide mass fingerprinting. Merely 19 protein spots, corresponding to

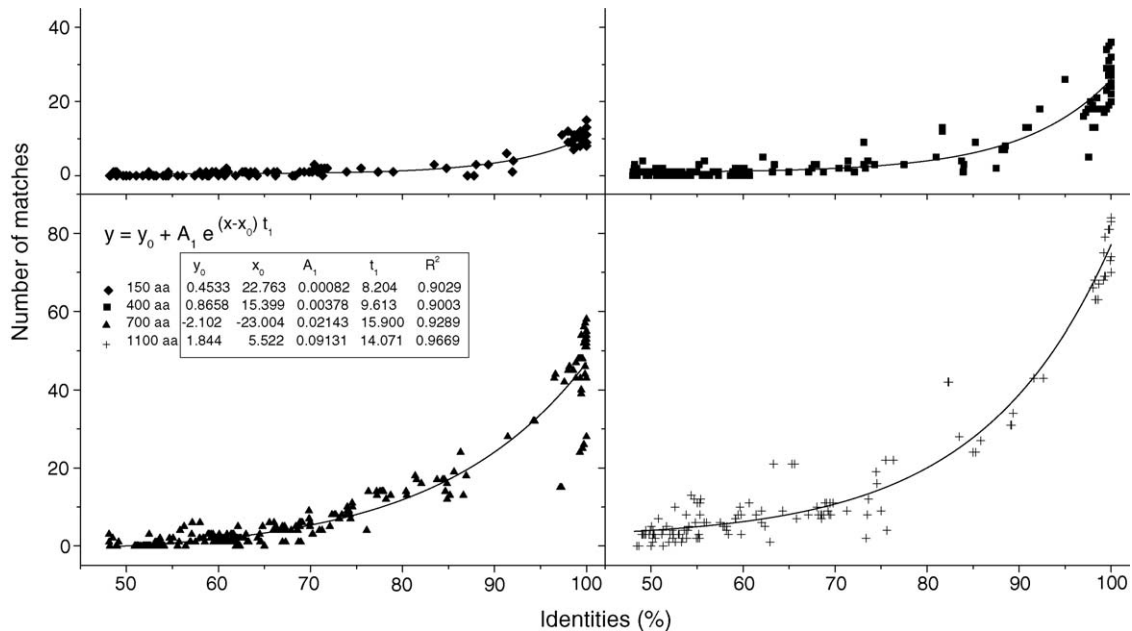


Fig. 4. Matches of peptide masses of a protein of interest to its homologous proteins rely on sequence identities and protein sizes. Results are from in silico experiments with four groups of proteins from *E. coli* (see text for detail). Variables in the regression equation: y , number of matches; x , identities; y_0 , x_0 , A_1 and t_1 , regression constants; R^2 , the square of the correlation coefficient.

10.8% of all the identified spots, were identified as true positive by using the protein database “13bacilli” or the NCBIInr-Bacteria database. Among them, six spots were only identified with NCBIInr-Bacteria but not with “13bacilli”, because all proteins corresponding to these six spots are exclusively from the species *B. megaterium* that were already included in the NCBIInr database. Three spots could be identified with “13bacilli” but not with NCBIInr-Bacteria, as the MASCOT scores for these three spots were not significant enough, when the NCBIInr-Bacteria database was applied.

Although nine spots were identified with both “13bacilli” and “bmegMEC”, the MASCOT scores using “13bacilli” were usually much lower than the scores using “bmegMEC”. Moreover, the cross-species spot identification generated a quite large number of false positives. With the databases “13bacilli” and NCBIInr-Bacteria, 13 and 12 spots, respectively, were identified as significant but the identification was proved false. Generally, the false positive spots can also be identified using the “bmegMEC” database, however, as different proteins with more significant scores. Some of the results were also verified by additional

peptide fragmentation through ESI-QqTOF MS/MS analysis. In several cases, the false “matched” homologous proteins are of much smaller molecular weight than calculated from the 2-DE gels. Therefore, this kind of cross-species matching was clearly a random false positive matching. Nevertheless, the relatively high number of false positive matches makes protein spot identification by cross-species database searching generally unreliable.

It should be mentioned that there were six spots, which could only be identified with “13bacilli” or NCBIInr-Bacteria but not with “bmegMEC”, as were verified by additional peptide fragmentation through ESI-QqTOF MS/MS analysis. The corresponding proteins were confirmed to be absent in the database “bmegMEC”, obviously because of the incompleteness of the genomic sequences of the *B. megaterium* strain studied.

3.3. Peptide mass matches strongly depend on sequence identity

The above results clearly showed that for protein spot identification through peptide mass fingerprint-

ing, using a strain-specific protein database generated from low-coverage genomic sequences is much more helpful than using protein sequences of even closely related species. In fact, for the first time the availability of the “bmegMEC” database has significantly facilitated our high-throughput protein identification via PMF.

To quantitatively understand the reason behind it, we did a further *in silico* experiment (Fig. 4.). Four groups of proteins were selected from the proteome of the well-sequenced *Escherichia coli* K12 (<http://www.ncbi.nih.gov/genomes/lproks.cgi>), representing proteins of different lengths, i.e. 1100 amino acids (aa), 700 aa, 400 aa and 150 aa, respectively. Each protein was queried against the NCBI nr database to find homologs with sequence identity greater than 50%. Each protein and its homologs were *in silico* digested with trypsin (the applied digestion rule was XXXX(K or R)(not P)XXXX). The generated peptide fragments with more than three amino acids were taken into account to calculate the number of matches between the protein and its homologs. Such matching resembles the real process of searching with experimentally measured peptide masses from MALDI-TOF MS analysis in a specific protein database through PMF searching programs such as MASCOT, ProFound, and PeptIdent. The number of matches was plotted against the identity level in Fig. 4. A statistical analysis showed that there is roughly an exponential relationship between the number of matches and the identity level. In other words, the number of matches dramatically decreases with the decrease of the identity. Each decrease of 10% in identities causes a decrease of about 50% in the number of matches. For instance, a protein of 1100 aa has 80 matches at identity 100%, 40 matches at identity 90%, 20 matches at identity 80%, and so on.

Fig. 4 also shows that the matching numbers decrease with the decline of the protein length as well. This is reasonable since smaller protein generates less number of peptide fragments during an enzymatic digestion. This fact makes the identification of small proteins rely much more on the presence of highly (if not completely) identical homologs in the protein database used. As an example, 80% of identity can sometimes produce above 20 matches for a protein with 1100 aa, that still makes the protein identification possible. In comparison, 90% of iden-

tity produces less than four matches for a protein with 150 aa, that makes an identification almost impossible. This is especially true when taking into consideration that not all peptide fragments that are obtainable from a theoretical digestion of a protein can be detected by mass spectrometric analysis, i.e. the sequence recovery of a protein from MS analysis can never reach 100%.

It is clear that match numbers, or the possibility to identify a protein using peptide masses, is highly dependent on the identity level between the protein studied and its homologous protein in the queried protein database. Consequently, the questions are how high the identities of homologous proteins among different species within the same genus or even among the subspecies of a specific species are and whether the identities are high enough for cross-species protein identification via PMF. Table 1 shows the number of proteins whose sequences are at least 90% identical between each two of the completely sequenced 14 bacilli strain. Between different subspecies (see the boxes in Table 1), such number corresponds to 58–96% of the proteins of the species in the first column, indicating a high possibility to identify these proteins by querying the protein database of another subspecies. Systematical comparison of proteins from variant *E. coli* subspecies gave the same results (data not shown). If comparing different species of the same genus (here *Bacillus*), some species are still highly related to each other (e.g. *B. anthracis*, *B. cereus* and *B. thuringiensis*), whereas the others demonstrate much remote relations (e.g. *B. licheniformis*, *B. clausii*, *B. halodurans*, *B. subtilis* and *B. megaterium*). This is consistent with the prevalent knowledge that all the aerobic spore-forming bacteria were lumped into the genus *Bacillus*, even though their GC content, physiology, sporulation, etc. are sometimes very different. In fact, only 4% of the 5028 CDSs of *B. megaterium* are at least 90% identical to CDSs of the other bacteria including all the 13 sequenced *Bacillus* species. Ninety percent of its CDSs have homologs from the public database with an identity level lower than 80%. It is thus understandable why the application of the cross-species protein database “13bacilli” only resulted in the identification of a few protein spots for *B. megaterium*.

A number of protein spots in our study can also be identified against partial proteins in the “bmeg-

Table 1
Percentage of proteins with at least 90% coding sequence identities between different species of the genus *Bacillus*

Strain of Bacilli	Abbr.	Number of CDSs	At least 90% identical to													
			ban	baa	bar	bat	bce	bca	bcz	btk	bli	bld	bcl	bha	bsu	bmg
<i>B. anthracis</i> Ames	ban	5502	88.1	96.4	89.9	61.5	70	78.6	77.2	0.47	0.49	0.10	0.10	0.10	0.36	0.34
<i>B. anthracis</i> A2012	baa	5874	87.3	92.5	91.2	61.2	69.5	80	78.6	0.47	0.47	0.15	0.13	0.13	0.45	0.37
<i>B. anthracis</i> Ames 0581	bar	5837	90.9	88.2	84.7	57.9	67	74.1	72.9	0.44	0.46	0.10	0.10	0.10	0.34	0.32
<i>B. anthracis</i> Sterne	bat	5287	93.7	95.7	93.6	66	73.9	86.4	85.1	0.49	0.51	0.11	0.11	0.11	0.37	0.35
<i>B. cereus</i> ATCC 14579	bce	5355	63.3	62.7	63.3	64.3	63.7	64.3	65.9	0.52	0.54	0.11	0.13	0.13	0.37	0.35
<i>B. cereus</i> ATCC 10987	bca	5880	65.9	65	66.8	65.6	58.1	66.7	65.4	0.45	0.47	0.1	0.11	0.11	0.4	0.32
<i>B. cereus</i> ZK	bcz	5134	83.7	84.3	83.7	86.7	66.8	75.9	85.1	0.5	0.52	0.11	0.13	0.13	0.38	0.37
<i>B. thuringiensis</i> ser.konkukian 97-27	btk	5197	81.5	82.3	81.6	84.7	67.8	73.7	84.2	0.5	0.51	0.11	0.13	0.13	0.38	0.34
<i>B. licheniformis</i> ATCC 14580	bli	4197	0.64	0.54	0.64	0.64	0.66	0.64	0.64	0.64	96.2	0.28	0.38	0.38	7.5	0.69
<i>B. licheniformis</i> DSM13	bld	4196	0.69	0.54	0.69	0.66	0.71	0.69	0.66	0.66	95.8	0.28	0.38	0.38	7.5	0.69
<i>B. clausii</i> KSM-K16	bcl	4108	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.29	0.29	0.29	1.5	0.17	0.09	
<i>B. halodurans</i> C-125	bha	4066	0.14	0.14	0.14	0.14	0.17	0.17	0.17	0.34	0.36	1.5	0.22	0.21	0.14	
<i>B. subtilis</i> 168	bsu	4106	0.51	0.48	0.51	0.51	0.51	0.58	0.51	7.7	7.7	0.17	0.17	0.21	0.75	
<i>B. megaterium</i> DSM 319	bmg	5028	0.49	0.57	0.49	0.49	0.49	0.51	0.49	0.71	0.71	0.13	0.13	0.17	0.77	

The boxes show comparisons between subspecies. Other highly related species are underlined. As an example, the number “88.1” (in the first line, fifth column) means 88.1% of the CDSs of *B. anthracis* Ames (ban) were found at least 90% identical to certain CDSs of the other *B. anthracis* subspecies A2012 (baa).

MEC” database. The existence of such partial proteins is because of the incompleteness of the raw genomic sequence. On the other side, these protein spots could not be identified by using cross-species protein database “13bacilli”. This means that more peptide mass matches can be achieved against partial but completely identical sequences than against homologous proteins, which are complete but not completely identical.

4. Discussion

A low-coverage genomic sequence has several negative impacts on spot identification via PMF. First, low-coverage means low quality. High quantity of sequencing errors, e.g. base exchange, deletion and insertion, often cause problems for coding sequence prediction, such as transcription frame shift or unusual stop codons. In our work, such problems were basically solved by applying the homology-based CDS prediction program “IdentICS”. In addition, the drawback of homology-based methods in finding strain-specific CDSs can be compensated by using ab initio gene-finding programs such as CRITICA and GeneMarks that apply non-homology information.

Second, low-coverage also means incomplete. If experimental gap closure such as by generating more coverage of genomic sequences or by variant finishing strategies (International Human Genome Sequencing Consortium, 2004) is not possible, the following concepts can be used for the improvement of the protein database quality:

- (1) Integration of known sequences of the organism of interest from public database. As mentioned in Section 2 of this paper, NCBI ENTREZ integrated database has included 572 proteins from *B. megaterium*, of which 279 are non-redundant. For 114 of them no corresponding coding sequences were found in the low-coverage genomic sequence of the *B. megaterium* strain DSM 319. They were therefore added directly to the “bmgMEC” database. Another 28 proteins partially overlapped with the genomic sequence. They were then used to substitute the incomplete partial proteins. The new protein database generated was named as “bmgMECI”. Several spots previously only identified

with NCBI-Bacteria are now able to be identified with bmegMECI.

- (2) Joining fragments of a protein that locate on the ends of two contigs by using homologous proteins from public database as a key (Fig. 1A). This was proved to generally increase the spot identification scores in this work. The gap inside a merged protein can be further artificially filled with the corresponding sequence of its most identical homologous protein (Fig. 1A). Similarly, a broken fragment, which cannot be joined to any other fragment can be extended by using the “missing” sequence from its most identical homologous protein (Fig. 1B).
- (3) Gaps in genomic sequence can be inferred by comparative genomics if the genomic sequences of its close relatives are available. Missing CDSs within such gaps can be taken from the related genomic sequences and integrated to the constructed protein database.

5. Conclusion

The results of this work demonstrated that the protein database generated from the low-coverage raw genomic sequence of *B. megaterium* is useful for a fast and reliable protein identification solely based on peptide mass fingerprint (PMF) from high-throughput MALDI-TOF MS analysis. Even partial proteins in the strain-specific database, which resulted from the incompleteness of the genomic sequence have been proven to be helpful for the identifications. The strain-specific database greatly outperformed protein databases of even closely related species. Due to high number of random matches (false positive), cross-species protein identification by PMF is not reliable. The protocols developed and used in this work to improve the protein database, such as end merge, are useful, especially when only genomic sequences of even lower coverage are available.

Acknowledgements

This work was financially supported by the Deutsche Forschungsgemeinschaft through the Sonderforschungsbereich SFB 578. The authors thank P.

Westphal for her excellent technical assistance in 2-DE experiments and protein sample preparation for MS analysis. The authors are grateful to the GBF MS analytic group, especially J. Majewski and A. Meier for their support in the MALDI-TOF MS analysis.

References

- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Badger, J.H., Olsen, G.J., 1999. CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* 16, 512–524.
- Barrett, J., Brophy, P.M., Hamilton, J.V., 2005. Analysing proteomic data. *Int. J. Parasitol.* 35, 543–553.
- Bird, C., Grafham, D., 2004. BAC finishing strategies. *Methods Mol. Biol.* 255, 255–277.
- Cordwell, S.J., Wilkins, M.R., Cerpa-Poljak, A., Gooley, A.A., Duncan, M., Williams, K.L., Humphrey-Smith, I., 1995. Cross-species identification of proteins separated by two-dimensional gel electrophoresis using matrix-assisted laser desorption ionisation/time-of-flight mass spectrometry and amino acid composition. *Electrophoresis* 16, 438–443.
- International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945.
- Kim, H.J., Lee, D.Y., Lee, D.H., Park, Y.C., Kweon, D.H., Ryu, Y.W., Seo, J.H., 2004. Strategic proteome analysis of *Candida magnoliae* with an unsequenced genome. *Proteomics* 4, 3588–3599.
- Malten, M., Hollmann, R., Deckwer, W.D., Jahn, D., 2005a. Production and secretion of recombinant *Leuconostoc mesenteroides* dextranucrase DsrS in *Bacillus megaterium*. *Biotechnol. Bioeng.* 89, 206–218.
- Malten, M., Nahrstedt, H., Meinhardt, F., Jahn, D., 2005b. Coexpression of the type I signal peptidase gene sipM increases recombinant protein production and export in *Bacillus megaterium* MS941. *Biotechnol. Bioeng.* 91, 616–621.
- Marrero, J., Gonzalez, L.J., Sanchez, A., Ayala, M., Paz-Lago, D., Gonzalez, W., Fallarero, A., Castellanos-Serra, L., Coto, O., 2004. Effect of high concentration of Co (II) on *Enterobacter liquefaciens* strain C-1: a bacterium highly resistant to heavy metals with an unknown genome. *Proteomics* 4, 1265–1279.
- Mathesius, U., Imin, N., Chen, H., Djordjevic, M.A., Weinman, J.J., Natera, S.H., Morris, A.C., Kerim, T., Paul, S., Menzel, C., Weiller, G.F., Rolfe, B.G., 2002. Evaluation of proteome reference maps for cross-species identification of proteins by peptide mass fingerprinting. *Proteomics* 2, 1288–1303.
- Pappin, D.J., Hojrup, P., Bleasby, A.J., 1993. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* 3, 327–332.
- Sun, J., Zeng, A.-P., 2004. IdentiCS—identification of coding sequence and in silico reconstruction of the metabolic net-

- work directly from unannotated low-coverage bacterial genome sequence. BMC Bioinformatics 5, 112.
- Vary, P.S., 1994. Prime time for *Bacillus megaterium*. Microbiology 140 (Pt 5), 1001–1013.
- Wang, W., Sun, J., Nimtz, M., Deckwer, W.-D., Zeng, A.-P., 2003. Protein identification from two-dimensional gel electrophoresis analysis of *Klebsiella pneumonia* by combined use of mass spectrometry data and raw genome sequences. Proteome. Sci. 1, 6.
- Wang, W., Hollmann, R., Furch, T., Nimtz, M., Malten, M., Jahn, D., Deckwer, W.-D., 2005. Proteome analysis of a recombinant *Bacillus megaterium* strain during heterologous production of a glucosyltransferase. Proteome. Sci. 3, 4.
- Wittchen, K.D., Meinhardt, F., 1995. Inactivation of the major extracellular protease from *Bacillus megaterium* DSM319 by gene replacement. Appl. Microbiol. Biotechnol. 42, 871–877.