Proteome Science

BioMed Central

Research

Open Access

Protein identification from two-dimensional gel electrophoresis analysis of *Klebsiella pneumoniae* by combined use of mass spectrometry data and raw genome sequences

Wei Wang†[1], Jibin Sun†[2], Manfred Nimtz[3], Wolf-Dieter Deckwer[1] and An-Ping Zeng*[2]

Address: [1]TU-BCE, GBF – German Research Centre for Biotechnology Mascheroder Weg 1, 38124 Braunschweig, Germany, [2]Department of Genome Analysis, GBF – German Research Centre for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany and [3]Department of Structure Study, GBF – German Research Centre for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany

Email: Wei Wang - wwa@GBF.de; Jibin Sun - JSU@GBF.de; Manfred Nimtz - MNI@GBF.de; Wolf-Dieter Deckwer - WDD@GBF.de; An-Ping Zeng* - aze@GBF.de

* Corresponding author †Equal contributors

Published: 03 December 2003

Proteome Science 2003, **1**:6

Received: 23 September 2003
Accepted: 03 December 2003

This article is available from: http://www.proteomesci.com/content/1/1/6

Abstract

Separation of proteins by two-dimensional gel electrophoresis (2-DE) coupled with identification of proteins through peptide mass fingerprinting (PMF) by matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) is the widely used technique for proteomic analysis. This approach relies, however, on the presence of the proteins studied in public-accessible protein databases or the availability of annotated genome sequences of an organism. In this work, we investigated the reliability of using raw genome sequences for identifying proteins by PMF without the need of additional information such as amino acid sequences. The method is demonstrated for proteomic analysis of *Klebsiella pneumoniae* grown anaerobically on glycerol. For 197 spots excised from 2-DE gels and submitted for mass spectrometric analysis 164 spots were clearly identified as 122 individual proteins. 95% of the 164 spots can be successfully identified merely by using peptide mass fingerprints and a strain-specific protein database (ProtKpn) constructed from the raw genome sequences of *K. pneumoniae*. Cross-species protein searching in the public databases mainly resulted in the identification of 57% of the 66 high expressed protein spots in comparison to 97% by using the ProtKpn database. 10 *dha* regulon related proteins that are essential for the initial enzymatic steps of anaerobic glycerol metabolism were successfully identified using the ProtKpn database, whereas none of them could be identified by cross-species searching. In conclusion, the use of strain-specific protein database constructed from raw genome sequences makes it possible to reliably identify most of the proteins from 2-DE analysis simply through peptide mass fingerprinting.

Background

The identification of proteins and protein expression patterns under given physiological conditions by proteomic analysis has gained fundamental importance for functional study of cellular processes in recent years. Mass spectrometry (MS) has become a central element for proteomic analysis [1-6]. It is used in combination with various protein separation methods and bioinformatic tools for large-scale protein identification and characterization of various organisms [reviewed in [7-16]]. Among

different MS techniques peptide mass fingerprinting (PMF) remains the most simple and powerful technique for high-throughput protein identification, in which peptide mass fingerprints acquired by matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) are compared with theoretical peptide mass fingerprints calculated for all the proteins in a given protein sequence database. For bacterial strains PMF is found even more reliable for species-specific protein identification than N-terminal sequencing [17]. This approach can be successfully applied for organisms the genomes of which are fully sequenced and annotated, or for proteins having sequences well-conserved for cross-species identification. Otherwise, additional information such as amino acid sequences is often needed for an unambiguous identification [1,3,10,17-19]. Therefore, systematic identification of protein expressions of an organism can be greatly enhanced with genome sequence data.

Presently, the genomes of more than 100 organisms have been completely sequenced http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html, NCBI 06.Sept.2003). More than 100 genome sequencing projects are in progress. However, the annotation (function assignment) of these genome sequences is a time and resource consuming process. So far, the genome sequences of about 90 organisms have been extensively annotated. In fact, there is a relatively long time-delay between the completion of genome sequencing and the full annotation of the sequences. This means a time-delay in the full use of genome sequences for protein identification.

In this work, we propose to directly use raw genome sequences for identifying proteins separated by 2-DE and analyzed by mass spectrometry. The method is demonstrated for proteomic analysis of *Klebsiella pneumoniae*, an organism with biomedical importance (e.g. respiratory tract infection, urea tract infection and biofilm formation) and many potential biotechnological applications (e.g nitrogen fixation, production of enzymes and biochemicals). To our knowledge a large-scale proteomic analysis of *K. pneumoniae* has not been reported in literature so far. One of our objectives is to identify as many as possible proteins involved in the anaerobic bioconversion of glycerol to 1,3-propanediol by *K. pneumoniae*. The microbial production of 1,3-propanediol has recently attracted a great deal of industrial attention as an emerging new biochemical and *K. pneumoniae* is a major model organism for understanding the metabolic and regulatory pathways of this bioconversion process [20,21]. The large-scale profiling of expressed proteins under given physiological conditions, especially those directly involved in the enzymatic conversion of glycerol into intermediates and the final product is desirable [22]. The proteomic method should be also useful for studying the medical aspects of this organism.

Results and Discussion

Protein database preparation and genome sequence annotation

Because the web version of GeneMarkS does not accept multiple-sequence FASTA format file as input and does not consider partial ORF lying on the ends of a contig, the unfinished genome sequences must be modified before submission for the prediction of ORFs. A short artificial linker DNA sequence (CATTCATTCA*TAAATAAATAA*ATGAATGAATG*TTATTTATTTA*) that includes the start codons (underlined) and stop codons (italic) with all six possible transcription frame shifts in two directions was used to link all the 341 contigs of *K. pneumoniae* into one. All possible partial ORFs were then predicted by GeneMarkS. A total of 5616 ORFs were found and translated into proteins numbered as Kpn1, Kpn2...Kpn5616, respectively. These proteins were further compared to the genomic sequences to remove additional amino acid groups caused by the linker sequence. The functions of these proteins were assigned based on similarity comparison to proteins in SWISS-PROT and TrEMBL (see the protein sequence database at http://genome.gbf.de/bioinformatics/index.html).

The number of predicted ORFs for *K. pneumoniae* is much higher than the number of ORFs in *Salmonella typhimurium* LT2 (4451 ORFs) and *E. coli* K12 MG1655 (4289 ORFs), two close relatives of *K. pneumoniae*. This indicated the existence of gaps between contigs and/or sequencing errors inside contigs of the genome sequences of *K. pneumoniae* even with the 8 times coverage of the genome data. As a consequence, a real and biologically functioning ORF could be *in silico* predicted as several partial ORFs. In some cases, the sequencing errors (gaps or extensions) can lead to translation shifts and thus wrong predictions of protein sequences. In fact, near 20% of these proteins have a length shorter than half length of their most similar proteins in SWISS-PROT and TrEMBL. 759 proteins can be clearly classified as partial proteins belonging to 343 intact proteins. If an ORF is predicted as several partial ORFs or its translation is wrong, the chance to identify the corresponding protein merely by searching with PMFs in this strain-specific database will be reduced. In such a situation, PMF can be performed by searching in public databases for cross-species identification or additional information such as partial amino acid sequences from ESI-QqTOF-MS analysis should be considered. An alternative solution is to identify the partial ORFs and sequencing errors (mainly gaps or extensions) and to correct them *in silico*. A program is now under development to identify partial proteins through integrity comparison between the predicted protein and its most similar proteins found in

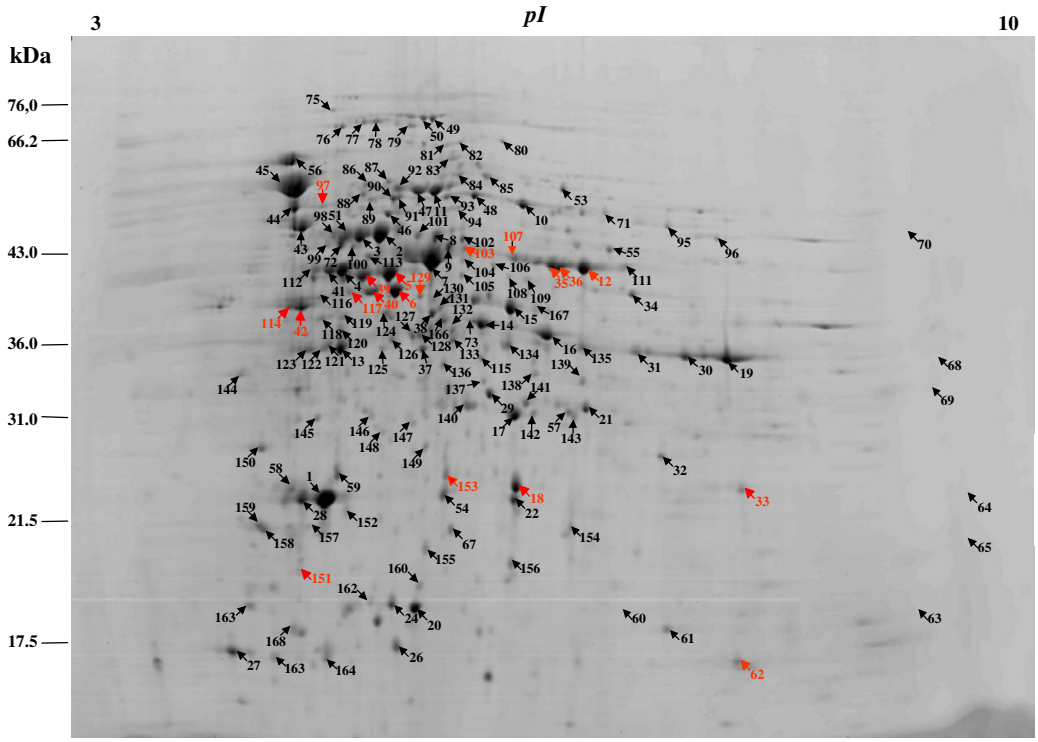

Figure 1
Typical image of 2-DE gel electrophoresis analysis of intracellular proteins of *K. pneumoniae* grown on glycerol. Spots numbers refer to those given in Tables 1 and 1. Red marked are proteins related to the *dha* regulon that is essential for the anaerobic glycerol metabolism in this organism.

the public protein databases (Swiss-Prot and TrEMBL). This program will also further identify sequencing errors such as base pair gaps or extensions causing transcription frame shifting and important base pair substitutions causing abnormally terminating or reading through, and then correct them. The curated protein database should enhance the identifiability especially when a lot of sequence errors exist, e.g., in the case of the low coverage of genome sequences.

Identification of proteins separated by 2-DE based on PMF and a strain-specific protein database

The proteomic analysis of *K. pneumoniae* was intended to elucidate some unusual dynamic behaviour of this organism anaerobically grown on glycerol [23]. To this end, samples were taken for 2-DE analysis to identify and quantify protein expressions in different stages of the fermentation process. Fig. 1 shows a typical protein expression pattern of *K. pneumoniae* obtained from the 2-DE analysis. 197 different protein spots were excised from 2-DE gels and analyzed by peptide mass fingerprinting using MALDI-TOF MS. In some cases ESI-QqTOF MS/MS analysis was also applied to verify the PMF results. As shown in Fig. 1, after searching in the specific ProtKpn database for *K. pneumoniae* and, in some case in the two public protein databases NCBInr and SWISS-PROT/TrEMBL, 163 spots that correspond to 83% of the proteins submitted for MS analysis, were identified as 122 individual proteins. That means some proteins appeared as several spots on the 2-DE gels because of co- and posttranslational modifications [24]. Table 1 [see Additional file 1] summarizes information on the identified proteins using **Mascot** as search program [25]. All the proteins listed in Table 1 [see Additional file 1] are the first protein candidate in the search result lists and their scores are generally much higher than the significant level (significant level is 50 when using the ProtKpn database) defined by **Mascot** and are normally significantly higher

than the scores for the second candidates. In most cases, these proteins are also the only candidate having significant score, leading to their unambiguous identifications.

Using the specific ProtKpn database 156 protein spots from 2-DE gels, corresponding to 95.7% of the 163 protein spots identified, can be successfully identified simply with their peptide mass fingerprinting data obtained from the MALDI-TOF MS analysis. These results were also confirmed by sequence similarity searching with partial amino acid sequences obtained from ESI-QqTOF MS/MS analysis as shown in Table 1 [see Additional file 1] for some of the proteins. This clearly demonstrates that protein identification of *K. pneumoniae* by simply using the peptide mass fingerprints and the ProtKpn database is reliable and sufficient enough for an identification when the score is significant.

Only 7 protein spots (spot 73, 75, 82, 84, 145, 151 and 167 in Table 1 [see Additional file 1]) were not identified by searching with PMFs in the ProtKpn database. 4 of them, spots 73, 75, 82 and 84, were submitted to ESI-QqTOF MS/MS analysis and identified by amino acid sequence similarity searching in the ProtKpn database and confirmed by comparison of their apparent *pI* and *Mr* values with those of the predicted. Another 2 of them, spot 145 and spot 167 can be identified by cross-species searching in the NCBInr database. The reason for the failure to identify these two proteins with the ProtKpn database is the appearance of these proteins as partial proteins in the ProtKpn database as mentioned above by genome sequence annotation. As a consequence of this, the scores were not significant enough for an unambiguous identification.

However, some proteins predicted as partial proteins in the ProtKpn database can still be identified. 15 Spots listed in Table 1 [see Additional file 1] belong to this category. For example, cysteine synthase is predicted as four partial proteins, namely Kpn 3706, Kpn 3707, Kpn 3708 and Kpn 3709 in the ProtKpn database. Three spots (spots 14, 16 and 166) have PMFs all matched separately to Kpn 3706, Kpn 3707, Kpn 3708 and/or Kpn 3709 (Table 1 [see Additional file 1]). So it is to expect that scores would be much higher when these partial proteins were curated to give one intact protein, as evidenced by the high score values of 147, 173 and 132 for the three spots by cross-species searching in the NCBInr database, respectively.

Most spots on the 2-DE gels were identified as a single protein with high scoring. However, we found that it is possible to identify proteins that were not separated by 2-DE. For example, the spots 121 and 122 were found to contain two proteins, transaldolase B (*pI* 4.99, *Mr* 35.4) and elongation factor Ts (*pI* 5.15, *Mr* 30.4), respectively. These two protein showed very similar apparent *Mr* values on the 2-DE gels (36,0 and 36.2, respectively) and both existed as different protein isoforms on the 2-DE gels, namely spots 121, 122, 123 for transaldolase B and spots 13, 120, 121 and 122 for elongation factor Ts (Figure 1). As a consequence, one isoform of each of the two proteins appeared coincidently at the same positions on the 2-DE gels.

34 proteins spots were not identified probably because of their low concentrations or of low MS spectra qualities. Most of them were obvious proteins of low expression levels on the 2-DE gels. In addition, they might belong to partial proteins or wrong predicted proteins in the ProtKpn database. All of them were just measured once by MALDI-TOF MS. With additional measurements they could be identified as well.

Comparison of protein identifications with public database and with strain-specific database

Peptide mass fingerprinting is the method of choice for straightforward high-throughput protein identification, when the amino acid sequence of a protein exists in any kind of annotated protein databases. For organisms whose genomes are not yet fully sequenced or annotated, it is desirable for cross-species protein identification of protein spots from 2-DE gels [26]. However, a theoretical study of Wilkins and Williams showed that peptide masses were not well conserved across species boundaries, with few or no peptides being conserved when sequence identity between two proteins was below 75% [27]. This means that cross-species protein identification by PMF is not reliable.

In order to make a judgment of how advantageous the specific ProtKpn database is in comparison with public databases, cross-species protein identification by PMF was carried out for 66 spots that belong to the high expressed proteins. Only 57% could be identified by searching in the public databases, whereas 97% were identified using the ProtKpn database (Data not shown). Using the three database search programs **Mascot**, **PeptIdent** and **Knexus** to interrogate the public databases NCBInr and SWISS-PROT/TrEMBL resulted in similar identifications.

Many house-keeping proteins of *K. pneumoniae* were identified with high scoring simply by peptide mass fingerprinting through both cross-species homologue protein searching and searching in the ProtKpn database. These are proteins mainly involved in the glycolysis and gluconeogenesis pathways, energy metabolism, amino acid metabolism, protein biosynthesis and anti-stress processes. Nearly all these identified proteins of *K. pneumoniae* show high sequence similarities to the sequences of homologue proteins from *S. typhimurium*, *S. typhi* and *E.*

coli. The genomes of these closely related microorganisms were sequenced and to a large extent annotated and therefore presented in the public databases. As a result, a cross-species protein identification of some of these housekeeping proteins is possible due to the high sequence conservation of these proteins between *K. pneumoniae* and these microorganisms.

Using the smaller strain-specific database significantly decreased the noises and uncertainties caused by the large number of sequences in the public databases. In contrast, searching with PMFs in public databases often provide many probabilistic protein candidates. A clear identification of the target protein is not always possible. In such a case, further fragmentation of selected peptides of a protein was inevitable to gain partial amino acid sequences for an unambiguous protein identity. There are different techniques for peptide fragmenting such as electrospray tandem quadrupole time-of-flight mass spectrometer (ESI-QqTOF MS/MS) [28,29] or MALDI-quadrupole time-of-flight mass spectrometer (MALDI-QqTOF MS) [30] as well as tandem time-of-flight mass spectrometer (MALDI-TOF/TOF MS) [31,32]. Except the requirement of additional resources and works, these techniques are more complex and less scalable than MALDI/MS peptide mass fingerprinting. ESI-QqTOF MS/MS which was used in this study is much more susceptible to contaminants of small molecules in digested peptide mixtures than MALDI-TOF MS and desalting with ZipTip is needed. For our experience it is often difficult to obtain MS/MS-spectra of high quality for performing successful sequencing.

As expected, when the proteins or their homologues are not present in the public databases, the search in public databases did not lead to their identification. This is especially the case for the identification of proteins of the *dha* regulon which encodes enzymes for the initial assimilation of glycerol and for the formation of 1,3-propanediol that were specially interesting for us [33]. Most proteins of the *dha* regulon in *K. pneumoniae* were identified with significant scores by searching the ProtKpn database but could not be identified using the public databases (Table 1). Except *K. pneumoniae*, the *dha* regulon is known to exist only in a few organisms like *Citrobacter freundii*, *Clostridium perfringens*, *Clostridium pasteurianum* and *Clostridium butyricum* [35]. Except for *C. perfringens,* the genomes of these organisms have not yet been sequenced. Using PMF only 1,3-propanediol oxidoreductase (PDOR) was identified both in the SwissProt/TrEMBL database and the NCBInr database and two subunits of glycerol dehydratase (GDHt) (beta and small subunits) were found in the NCBInr database. However, the identifications were not resulted from a cross-species identification but due to the existence of these two proteins of *K. pneumoniae* in the public databases. For the identification of dihydroxyacetone kinase (DHAK I), DHAK I from *C. freundii* and the hypothetical oxidoreductase yqhD (HOR) from *E. coli* were found as possible candidates by cross-species searching with **PeptIdent** in the SwissProt/TrEMBL database. But DHAK I of *C. freundii* was the third score in the protein candidates list and HOR of *E. coli* the sixteenth so that a definite identification of this enzyme was not possible with this approach.

Function assignment of identified proteins and their biological interpretations

The functions of identified proteins were assigned by comparing their sequences to public protein database SWISS-PROT/TrEMBL through a NCBI-BLAST local server. Homologue proteins with the highest sequence similarities are listed under the corresponding annotation for each identified proteins in Table 1 [see Additional file 1]), and wherever possible, well studied homologue proteins are preferred to be included.

The identified 122 proteins can be classified into 9 categories and 38 subcategories as shown in Table 1 [see Additional file 1] based on KEGG http://www.genome.ad.jp/kegg/. The categories cover from energy metabolism, catabolism of small molecules and anabolism of building blocks to genetic and environmental information processing such as transcription, translation, transportation and stress response. The first category, carbohydrate metabolism, is the biggest category and contains about 25% of all identified proteins or peptides. It includes all the enzymes of the *dha* regulon as well as enzymes of near-complete glycoslysis and gluconeogenesis pathways, enzymes of the pentose phosphate pathway and partial of the TCA cycle. The carbohydrate metabolism plays an essential role not only by delivering metabolic precursors but also by supplying energy. In the anaerobic glycerol bioconversion by *K. pneumoniae* substrate-level phosphorylation is the only way to generate energy. Many proteins in this category were found highly expressed under the defined fermentation conditions. It is interesting to mention that the key enzyme for the Entner-Doudoroff pathway, KHG/KDPG aldolase, was unexpectedly also identified. Its function for anaerobic glycerol metabolism is unknown and in fact has not been studied so far.

The *dha* regulon includes 15 ORFs which encode 5 metabolic enzymes, namely dihydroxyacetone kinase I and II (DHAK I and II), glycerol dehydrogenase (GDH), glycerol dehydratase (GDHt) and 1,3-propanediol dehydrogenase (PDOR), 1 regulatory protein, 1 activator for GDHt, 1 transport facilitator and 2 proteins of unknown functions [26]. As shown in Table 1 [see Additional file 1], we have identified all the 5 metabolic enzymes or their subunits with the help of strain-specific database ProtKpn. Of particular interest is the identification of both DHAK I and II

Table 2: Compare cross-species identification using public protein databases with identification using the specific protein database ProtKpn for the identification of protein of *dha* regulon of *Klebsiella pneumoniae*

Spot No[a]	Annotation and/or Homologue Proteins	KPN[b]	Knexus NCBInr	PeptIdent Swiss-prot/ TrEMBL		Mascot NCBInr Score[d]		Mascot kpn Score[d]	
			Z(%)[c]	Score[d]	SC[e] (%)	Score[d]	SC[e] (%)	Score[d]	SC[e] (%)
5	Glycerol dehydrogenase (EC 1.1.1.6) (GLDH) P45511_GLDA_CITFR	862	n.i.	n.i.	-	n.i.	-	121	39
39			n.i.	n.i.	-	n.i.	-	136	34
97	Dihydroxyacetone kinase (EC 2.7.1.29) sp| P45510| DAK_CITFR	332	n.i.	0.19	9.4	n.i.	-	131	31
42	Dehydroxyacetone kinase II, subunit 1 (EC 2.7.1.121) P76015_YCGT_ECOLI	863	n.i.	n.i.	-	n.i.	-	102	42
114			n.i.	n.i.	-	n.i.	-	89	29
18	Dihydroxyacetone kinase II, subunit 2 (EC 2.7.1.121) P76014_YCGS_ECOLI	864	n.i.	n.i.	-	n.i.	-	152	77
153			n.i.	n.i.	-	n.i.	-	103	60
33	Glycerol dehydratase beta subunit (EC 4.2.1.30) tr| O08505 [Klebsiella pneumoniae]	855	91	n.i.	-	n.i.	-	71	46
62	glycerol dehydratase small subunit (EC 4.2.1.30) tr| Q59475	854	98	n.i.	-	94	52	96	52
12	1,3-propanediol dehydrogenase (EC 1.1.1.202) Q59477_DHAT_KLEPN	858	n.i	0.31	34.1	94	34	105	36
35			99	0.35	54.0	182	54	175	48
36			91	0.18	37.0	75	36	63	26
103			n.i.	0.25	31.8	114	38	114	38
107			99	0.42	35.9	158	41	184	41
6	Hypothetical oxidoreductase yqhD (EC 1.1.-.-) Q46856_YQHD_ECOLI	3405	n.i.	0.19[16]	15.5	n.i.	-	199	40
40			n.i.	n.i.	-	n.i.	-	130	29
117			n.i.	n.i.	-	n.i.	-	119	37
129	Putative glycerol dehydrogenasetr| Q8ZR27	2195						74	42
151	orfY, unknown function gi| 940439| (U30903) [Klebsiella pneumoniae]	857	84	n.i	-	81	48		

[a]Refers to the proteins labelled in Figure 1 [b] Protein access numbers in the ProtKpn database [c]Knexus uses ProFound as search program. Profound calculates the probability that a candidate in a database search is the protein being analysed., A Z score is estimated as an indicator of the quality of the search result, when the search result is compared against an estimated random match population. Z score is the distance to the population mean in unit of standard deviation. It also corresponds to the percentile of the search in the random match population. [d] Using Peptident score is the number of peptides that match the theoretical peptides from a database entry divided by the total number of peptide masses specified for the search. Using Mascot score is -10*Log(P), where P is the probability that the observed match is a random event. If there is also a superscript number beside the score, it represents the position of this protein in the protein candidate list. Otherwise, it is the top one. [e]SC: Sequence coverage, defined as the ratio of the portion of protein sequence covered by matched peptides to the whole length of protein sequence.

of the *dha* regulon. The expression level of DHAK II was much higher than DHAK I. DHAK II was recently found by us as a second dha kinase and its expression well explains some peculiar observations of the fermentation process [23,33]. The expression of OrfY, which is a common component in the *dha* regulons of different organisms with unknown function, was identified as well. We found that the amino acid sequence of this protein (orfY) is slightly different from the one in the public database, which belongs to another *K. pneumoniae* strain ATCC 25655. The difference of several amino acids obviously affected the scoring, since this protein is quite small (15.4 kDa). A putative glycerol dehydrogenase (spot 129, KPN 2195) was also found beside the well-known glycerol dehydrogenase (GDH) of *dha* regulon. This protein is of 79% identical to the putative glycerol dehydrogenase of *S. typhimurium* (TrEMBL:Q8ZR27) and 29% to GDH of *K. pneumoniae*. The function of this protein deserves further study.

Many of the enzymes of the *dha* regulon were found to appear as different protein isoforms on the 2-DE gels because of post-translational modification (Table 1 [see Additional file 1]). The identification of these different isoforms of the enzymes is of special interest for understanding both their expression regulations and activity

controls through covalent modifications and will be further studied.

By applying the method presented in this work several new enzymatic and regulatory proteins were identified that have large impacts for understanding and optimizing the microbial production of 1,3-propanediol. The expression patterns of some of these proteins were discussed in term of metabolic pathway analysis of this emerging important industrial bioprocess elsewhere [22,23,33]. The identified protein spots are being used for comparison of protein expression profiles of *K. pneumoniae* to elucidate metabolic pathway regulation associated with gene overexpression or knockout experiments aimed at development of more efficient bioprocess for 1,3-propanediol production. The protein database and the method for protein identification can be also used to study other important biological processes and phenomena such as biofilm formation, nitrogen fixation and antibiotic resistance in *K. pneumoniae*.

Conclusion

The combined use of high-resolution 2-DE separation, high-throughput MS analysis and raw genome sequences for an extensive and reliable identification of proteins has been shown in this work to significantly accelerate the proteomic and functional-genomic studies of *K. pneumoniae* anaerobically grown on glycerol. In particular, the establishment of a strain-specific protein database from unannotated genome sequences simplifies and improves the protein identification to a large extent. With this approach, identification of a large portion of the expressed protein spots from 2-DE analysis can be achieved for this organisms with high confidence simply by peptide mass fingerprinting using MALDI-TOF MS data.

Material and Methods

Organism and cultivation

Klebsiella pneumoniae DSM 2026 obtained from the German Collection of Microorganisms (DSMZ) was used in this study. Cultivation medium and conditions in a fed-batch bioreactor were reported in detail by Wang et al. (*23*).

Two-dimensional gel electrophoresis

Intracellular proteins of *K. pneumoniae* from whole cell extraction were separated by two dimensional gel electrophoresis mainly according to the manual published by A. Görg et al. http://www.weihenstephan.de/blm/deg/manual and described in detail by Wang et al. [23].

Protein preparation for Mass Spectrometric analysis

Protein spots were excised from 2-DE gels, washed several times with 200 μl water, dehydrated in 50 μl acetonitrile, and dried in a vacuum concentrator (Eppendorf, Concentrator 5301). The gel pieces were treated with 100 mM NH_4HCO_3 containing 20 mM DTT at 56°C for 30 min and then with 100 mM NH_4HCO_3 containing 55 mM Iodoacetamide at room temperature in the dark for 30 min. Acetonitrile was added between the treatments to dehydrate the gel pieces. Subsequently, the gel pieces were washed twice with 100 mM NH_4HCO_3 and dehydrated with acetonitrile, then dried completely in the vacuum concentrator. For sequence-specific digestion the gel pieces were re-swollen in minimal volumes of 50 mM NH_4HCO_3containing 2 ng/μl Trypsin (sequencing grade modified, Promega Corp.) and incubated at 37°C overnight (<15 h). The resulting peptides were obtained by successive extraction of the digested mixtures with 25 mM NH_4HCO_3 and acetonitrile and then 5% formic acid (HCOOH) and acetonitrile. The extracts were pooled together, dried in the vacuum concentrator and reconstituted in 20 μl of an aqueous solution containing 0.5% HCOOH and 5% methanol (MeOH). The peptide extracts were purified on reversed-phased C_{18} ZipTip pipette tips (Millipore Corp.). Briefly, ZipTips were washed with an aqueous solution containing 0.5% HCOOH and 65% MeOH and equilibrated with a solution containing 0.5% HCOOH and 5% MeOH. Peptide extracts were then applied to the ZipTips. After washing with the solution containing 0.5% HCOOH and 5% MeOH, the purified and concentrated peptide extracts were eluted with 4–5 μl of a solvent containing 1.0% HCOOH and 65% MeOH.

MALDI/TOF-MS analysis of tryptic peptides

Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) was employed to obtain peptide mass fingerprint of a given protein. 0.5 – 1 μl of each concentrated peptide solution was mixed with the same volume of a saturated matrix solution of α-cyano-4-hydoxycinnamic acid (Bruker Daltonics) in 0.5% HCOOH-65%MeOH and spotted onto a 384 MTP target and dried at room temperature. In the case of very small protein spots (low expressions) peptides were directly eluted onto the target with 1–2 μl matrix solution after the treatment with ZipTip. The molecular masses of the tryptic peptides were determined in the positive-ion mode on a Bruker Ultraflex time-of-flight mass spectrometer (Bruker Daltonics GmbH, Germany) using an acceleration voltage of 20 kV.

ESI-MS/MS sequencing of selected peptides

Electrospray ionization quadrupole-time-of-flight tadem mass spectrometry (ESI-QqTOF MS/MS) was performed to acquire partial amino acid sequences of a protein. 3 μl of a concentrated peptide solution were filled into gold-coated nanospray glass capillaries and placed orthogonally in front of the entrance hole of a Q-TOF 2 mass spectrometer (Micromass, Manchester, England) equipped

with a nanospray ion source. A voltage of approximately 700–1000 V was applied to the capillary. For collision-induced dissociation, parent ions were selectively transmitted from the qudrupole mass analyzer into the collision cell. Argon was used as the collision gas and the kinetic energy was set at around -20 -40 V for optimal fragmentation. Daughter ions acquired were then separated by the orthogonal time-of-flight mass analyzer.

Protein identification using public protein sequence databases

Peptide mass fingerprints (PMFs) obtained from MALDI-TOF MS analysis were used for cross-species protein identification in public protein primary sequence databases. **Mascot** (Matrix Science Ltd., UK, http://www.matrixscience.com), **PeptIdent** (Swiss Institute of Bioinformatics, http://www.expasy.ch/tools/peptident.html) and **Knexus**™ (Proteometrics Inc., http://www.proteometrics.com) were employed for analysis of the Maldi data using the public databases NCBInr and SWISS-PROT/TrEMBL. Trypsin was given as the digestion enzyme, 2 missed cleavage sites were allowed, Cysteine was modified by iodoacetamide and methionine was assumed to be partially oxidized. All peptide mass values are monoisotopic and the mass tolerance was set at 200 ppm, but the observed mass accuracy was usually better than 50 ppm for identified peptides. Using **PeptIdent** *Mr* and *pI* values observed from the 2-D electrophoresis were also used as search parameters with *pI* range set at 0.5 and *Mr* range at 20%.

Selected peptides of a protein were fragmented by ESI-QqTOF MS/MS. MS/MS spectra were enhanced using the Max Ent 3 software (Micromass), followed by automatic or manual sequencing using the PepSeq program of the software package Masslynx™ Version 3.5 (Micromass). The partial amino acid sequences obtained were used for similarity searching of amino acid sequences against the SWALL Non-Redundant Protein Sequence Database using FASTA3 http://www.ebi.ac.uk/fasta33/ on the internet.

Protein identification using genome sequences of K. pneumoniae strain MGH 78578

The genome of the strain used in this study, *Klebsiella pneumoniae* DSMZ 2026, is not yet sequenced. However, another *Klebsiella* strain (*K. pneumoniae* MGH 78578) that is very similar to *K. pneumoniae* DSMZ 2026, was sequenced by the Genome Sequencing Center in the Medical School of Washington University http://genome.wustl.edu/projects/bacterial/. A whole genome shotgun approach was used to generate the 7.9 time coverage of genome data given as 341 contigs (state of January 2002). Until now there is no annotation publicly available for this organism. The contigs of the *K. pneumoniae* strain MGH 78578 were downloaded as a local database. Open reading frames (ORFs) were predicted from the contigs and translated to protein sequences by using the web version of the program GeneMarkS [34]. The functions of these proteins were assigned by comparing their sequences to public protein database SWISS-PROT and TrEMBL. Isoelectric point (*pI*) and molecular weight (*Mr*) of the proteins were calculated by using Vector NTI Suite 7.1 (InforMax, USA). Both genome sequences and protein sequences were formatted as local databases of BLAST (Basic Local Alignment Search Tool) [35].

After the development of a strain-specific protein sequence database (ProtKpn) for *K. pneumoniae*, it was formatted and installed on our local **Mascot** server http://genome.gbf.de/bioinformatics/index.html. PMFs from MALDI-TOF MS analysis were compared to the predicted peptide masses in this specific protein database using **Mascot** as a search program. Additionally, partial amino acid sequences from ESI-QqTOF MS/MS analysis were searched in the same database using the NCBI local **BLAST** function of the program BioEdit (downloaded from http://www.mbio.ncsu.edu/BioEdit/bioedit.html.

Additional material

Additional File 1

Table 1. Overview of K. pneumoniae *proteins identified after in-gel digestion with trypsin and MALDI-TOF MS and/or ESI-QqTOF MS/MS analysis [see Additional file 1].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1477-5956-1-6-S1.doc]

Acknowledgements

This work was financially supported by the European Committee Fifth Framework Project QLK5-1999-01360, the Bundesministrium für Bildung und Forschung (BMBF), Germany (Grant No. 031U110A) and by the project A5 in the Sonderforschungsbereich 578 der Deutschen Forschungsgemeinschaft (DFG). One of the authors (J. Sun) greatly acknowledges the Ph.D Scholarship of German Academic Exchange Service (DAAD). M. Probst and H. Zhao are thanked for their technical assistance. We are grateful to the GBF MS analytic group, especially A. Abrahamik, U. Beutling and J. Majewski for their help in the MALDI-TOF MS and ESI-QqTOF MS/MS analysis.

References

1. Hochstrasser DF, Sanchez J-C, Appel RD: **Proteomics and its trends facing nature's complexity.** *Proteomics* 2002, **2:**807-812.
2. Mann M, Hendrickson RC, Pandey A: **Analysis of proteins and proteomics by mass spectrometry.** *Annu Rev Biochem* 2001, **70:**437-473.
3. Gevaert K, Vandekerckhove J: **Protein identification methods in proteomics.** *Electrophoresis* 2000, **21:**1145-1154.
4. Chalmers MJ, Gaskell SJ: **Anvances in mass spectrometry for proteome analysis.** *Curr Opin Microbiol* 2000, **11:**384-390.
5. Yate JR III: **Mass spectrometry from genomics to proteomics.** *Trends Genet* 2000, **16:**5-8.

6. Yates JR III: **Database searching using mass spectrometry data.** *Electrophoresis* 1998, **19:**893-900.
7. Hancock WS, Wu S-L, Shieh P: **The challenges of developing a sound proteomics strategy.** *Proteomics* 2002, **2:**352-359.
8. Griffin TJ, Aebersold R: **Advances in proteome analysis by mass spectrometry.** *J Biol Chem* 2001, **276:**45497-45500.
9. Pandey A, Mann M: **Proteomics to study genes and genomes.** *Nature* 2000, **405:**837-846.
10. Lahm H-W, Langen H: **Mass spectrometry: A tool for the identification of proteins serarated by gels.** *Electrophoresis* 2000, **21:**2105-2114.
11. Lopez MF: **Better approaches to finding the needle in a haystack: Optimizing proteome analysis through automation.** *Electrophoresis* 2000, **21:**1082-1093.
12. Washburn MP, Yate JR III: **Analysis of the microbial proteome.** *Curr Opin Microbiol* 2000, **3:**292-297.
13. O'Connor CD, Adams P, Alefounder P, Farris M, Kinsella N, Li Y, Payot S, Skipp P: **The analysis of microbial proteomes: Strategies and data exploitation.** *Electrophoresis* 2000, **21:**1178-1186.
14. Lopez MF: **Proteome analysis I. Gene products are where the biological action is.** *J Chromatogr B* 1999, **722:**191-202.
15. Quadroni M, James P: **Proteomics and automation.** *Electrophoresis* 1999, **20:**664-677.
16. Blackstock WP, Weir MP: **Proteomics: quantitative and physical mapping of cellular proteins.** *Trends-Biotechnol* 1999, **7:**121-127.
17. Mathesius U, Imin N, Chen HC, Djordjevic MA, Weinman JJ, Natera SHA, Morris AC, Kerim T, Paul S, Menzel C, Weiler GF, Rolfe BG: **Evaluation of proteome reference maps for cross-species identification of proteins by peptide mass fingerprinting.** *Proteomics* 2002, **2:**1288-1303.
18. Lester PJ, Hubbard SJ: **Comparative bioinformatic analysis of complete proteomes and protein parameters for cross-species identification in proteomics.** *Proteomics* 2002, **2:**1392-1405.
19. Wasinger VC, Urquhart BL, Humphery-Smith I: **Cross-species characterization of abundantly expressed *Ochrobactrum anthropi* gene products.** *Electrophoresis* 1999, **20:**2196-2203.
20. Biebl H, Menzel K, Zeng AP, Deckwer WD: **Microbial production of 1,3-propanediol.** *App Microbiol Biotechnol* 1999, **52:**289-297.
21. Zeng AP, Biebl H: **Bulk-Chemicals from Biotechnology: the case of microbial production of 1,3-propanediol and the new trends.** In *Adv Biochem Eng Biotechnol Volume 74*. Edited by: *Schügerl K, Zeng AP. Berlin, Springer*; 2002:237-257.
22. Zeng AP, Sun J, Wang W, Hartlep M, Deckwer W-D: **Use of genomic, proteomic and metabolic data and mathematic modeling for a system analysis of glycerol bioconversion to 1,3-propanediol,.** In *Proceedings of 3rd International Conference on Systems Biology* Edited by: *Aurell E, Elf J, Jeppsson J. Karolinska Institutet, Stockholm*; 2002:54-55.
23. Wang W, Sun J, Hartlep M, Deckwer WD, Zeng AP: **Combined use of proteomic analysis and enzyme activity assays for metabolic pathway analysis of glycerol fermentation by *Klebsiella pneumoniae*.** *Biotechnol Bioeng* 2003, **83:**525-536.
24. Harry JL, Wilkins MR, Herbert BR, Packer NH, Gooley AA, Williams KL: **Proteomics: Capacity *versus* utility.** *Electrophoresis* 2000, **21:**1071-1081.
25. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20:**3551-67.
26. Liska AJ, Shevchenko : **Expanding the organismal scope of proteomics: Cross-species protein identification by Maβ spectrometry and its implications.** *Proteomics* 2003, **3:**19-28.
27. Wilkins MR, Williams K: **Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation.** *J Theor Biol* 1997, **186:**7-15.
28. Chernushevich IV, Loboda AV, Thomson BA: **An introduction to quadrupole-time-of-flight mass spectrometry.** *J Mass Spectrom* 2001, **36:**849-865.
29. Shevchenko A, Chernushevich IV, Ens W, Standing KG, Thomson B, Wilm M, Mann M: **Rapid' de Novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer.** *Rapid Commun Mass Spectrom* 1997, **11:**1015-1024.
30. Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, Standing KG: **Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching.** *Anal Chem* 2001, **73:**1917-1926.
31. Bienvenut WV, Déon C, Pasquarello C, Campbell JM, Sanchez J-C, Vestal ML, Hochstrasser DF: **Matrix-assisted laser desorption/ionisation-tadem mass spectrometry with high resolution and sensitivity for identification and characterization of proteins.** *Proteomics* 2002, **2:**868-876.
32. Medzihradszky KF, Campbell JM, Baldwin MA, Falick AM, Juhasz P, Vestal ML, Burlingame AL: **The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer.** *Anal Chem* 2000, **72:**552-558.
33. Sun J, van den Heuvel J, Soucaille P, Zeng AP: **Comparative genomic analysis of *dha* regulon and related genes for anaerobic glycerol metabolism in microorganisms.** *Biotechnol Prog* 2003, **19:**263-272.
34. Besemer J, Lomsadze A, Borodovsky M: **GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions.** *Nucleic Acids Research* 2001, **29:**2607-2618.
35. Altschul SF, Madden TL, Schäffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997, **25:**3389-3402.