Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 363 (2007) 822–827
BBRC
www.elsevier.com/locate/ybbrc

2DBase: 2D-PAGE database of *Escherichia coli*

Chandran Vijayendran [a,b,*,1], Sebastian Burgemeister [b,1], Karl Friehs [b], Karsten Niehaus [c], Erwin Flaschel [b]

[a] *International NRW Graduate School in Bioinformatics and Genome Research, University of Bielefeld, D-33594 Bielefeld, Germany*
[b] *Fermentation Engineering Group, University of Bielefeld, D-33594 Bielefeld, Germany*
[c] *Faculty of Biology, University of Bielefeld, D-33594 Bielefeld, Germany*

Received 7 September 2007
Available online 21 September 2007

Abstract

We present a web-based integrated proteome database, termed 2DBase of *Escherichia coli* which was designed to store, compare, analyse, and retrieve various information obtained by 2D polyacrylamide gel electrophoresis and mass spectrometry. The main objectives of this database are (1) to provide the features for query and data-mining applications to access the stored proteomics data (2) to efficiently compare the specific protein spots present in the comparable proteome maps and (3) to analyse the data with the integrated classification for cellular functions of gene products of *E. coli*. This database currently contains 12 gels consisting of 1185 protein spots information in which 723 proteins were identified and annotated. Individual protein spots in the existing gels can be displayed, queried, analyzed, and compared in a tabular format based on various functional categories enabling quick and subsequent analyses. Our database satisfies the requirement to be a federated 2-DE database by accomplishing various tasks through a web interface providing access to a relational database system. The 2DBase of *E. coli* database can be accessed at http://2dbase.techfak.uni-bielefeld.de/.

Keywords: 2D-PAGE; 2D gels; *E. coli*; Proteomics; Database; MG1655

Ever since the term ''proteome'' was introduced [1], conventional 2-DE gel electrophoresis has remained the major method for proteome analysis [2,3]. High throughput proteomic data generated by 2-DE gel experiments require elaborate data handling to ensure comprehensive analyses. Increasing amount of data increases the complexity of comparing maps present in any existing database. Several 2-DE gel electrophoresis databases have been published in recent times which contain large amounts of experimental proteomics data generated by various high-throughput methodologies (http://expasy.org/world-2dpage/). With the rapid increase in the raw proteomics data within or between laboratories, it has becoming increasingly challenging to meaningfully compare the results from such large datasets containing numerous 2-DE maps. Database management systems, along with proficient methods of map comparison, would enhance the analyses.

Here we report a proteomics database of *Escherichia coli* which currently consists of 1185 protein spots information in which 723 protein spots were identified and annotated from 12 gels. Among them, 10 gels were generated during microbial evolutionary experiments (unpublished) and the remaining two gels are discussed later on in this report. The database is a relational database system (Fig. 1A) and supports extensive search functions according to several fields (accession number, gene id, description, author, spot id, p*I*/MW) (Fig. 1B). We have applied an extensive, quick, efficient and easy approach to compare various gels by classifying each protein spot utilizing a previously published classification system for cellular functions of gene products of *E. coli* [4,5]. We used a scoring function generated from the peak value and normalized quantity of the protein spots,

* Corresponding author. Address: International NRW Graduate School in Bioinformatics and Genome Research, University of Bielefeld, D-33594 Bielefeld, Germany. Fax: +49 0521 1066475.
E-mail address: cvijayen@cebitec.uni-bielefeld.de (C. Vijayendran).
[1] Equal contributors.

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.09.050

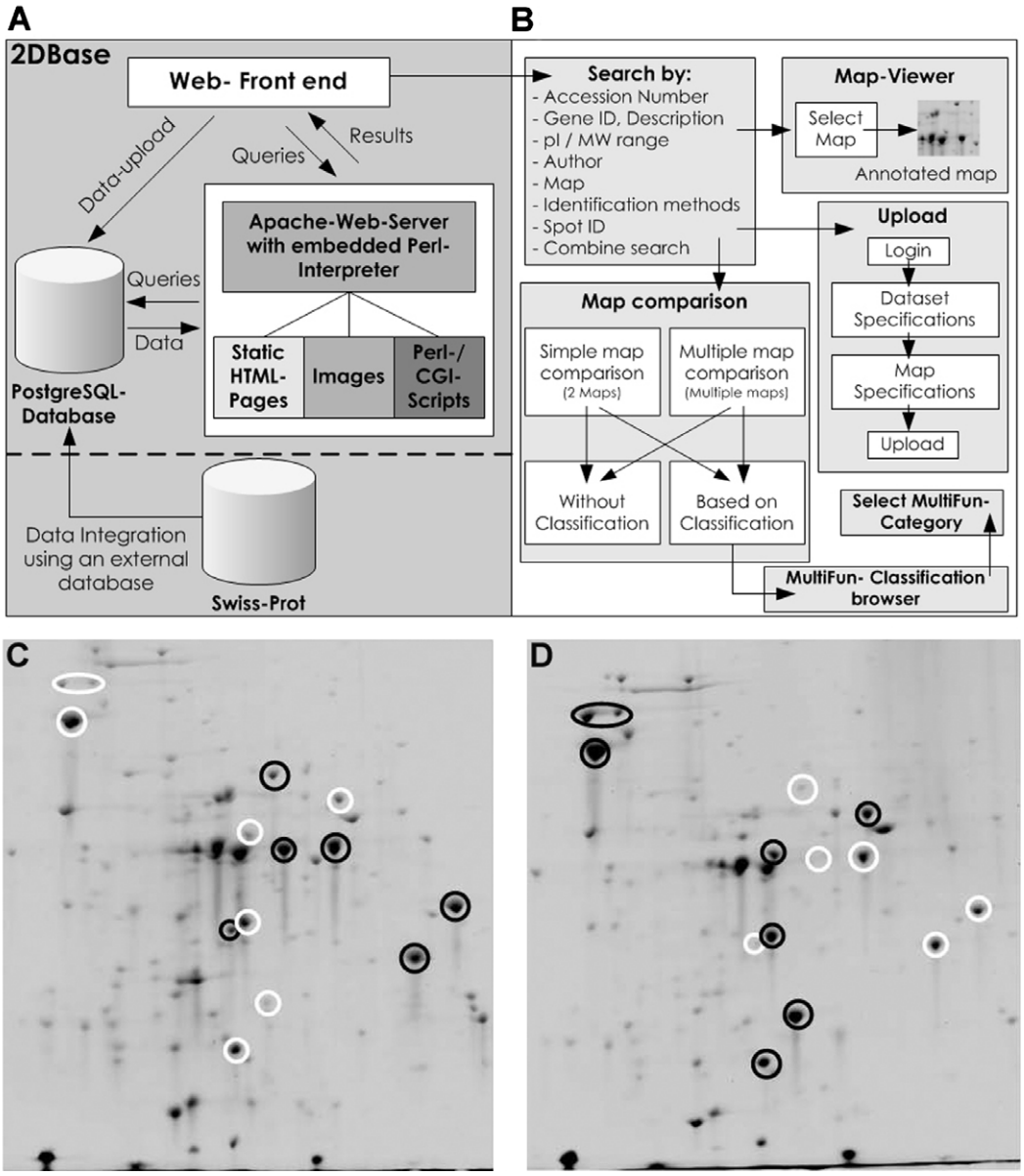

Fig. 1. 2DBase of *Escherichia coli* (A) Data flow diagram of 2DBase of *Escherichia coli*. (B) The database is a relational database system, showing its extensive search, comparison and classification options. (C,D) CAS-PLUS and CAS-MINUS proteome gels. The proteome gels from a strain (MG1655) which was grown in the presence (C) and absence (D) of Casamino-acids, respectively. Encircled black and white symbols represent protein over- and under-expression in the corresponding gel.

and this information was utilized in a sortable table format. We also made a functional classification of the protein spots. All of this enabled us to quickly analyze and compare the gels in an efficient manner. To our knowledge, there have been no previous reports of a proteome database which has the option of analyzing and comparing 2-DE gels at a single protein comparison level across all the gels aided by functional category classification.

Materials and methods

Strain and culture conditions. The bacterial strain, MG1655 was used in this study which is a derivative of *E. coli* K-12. All the experiments were conducted in 250 mL of M9 minimal medium with or without 2 gL^{-1} of Casamino-acids (CAS) supplemented with 4 gL^{-1} of glucose in covered 1 L Erlenmeyer flasks at 37 °C.

Two-dimensional SDS–PAGE. Approximately, 4.1×10^{10} cells were harvested from the exponential phase in all experiments. The cells were pelleted down at 6000 rpm at 4 °C for 10 min. Harvested cells were washed with low salt solution (3 mM KCl, 1.5 mM KH_2PO_4, 68 mM NaCl, 9 mM NaH_2PO_4) trice. The pellet was then resuspended in rehydration buffer (9 M urea, 4% 3-([3-chloramidopropyl]dimethylammonio)-1-propane-sulfonate (CHAPS), 85 mM dithiothreitol (DTT), 0.5 mM pefabloc SC and stored at −20 °C until next use. The cells were lysed by ultrasonication for 10 s, five times at 10% of maximum output (Branson sonifier 450). After 30 min of incubation at 37 °C with DNase and RNase, eventually the debris was pelleted down and the proteins present in the supernatant were precipitated with acetone at −20 °C overnight. The precipitated proteins were then resuspended in rehydration buffer and 300 μg of protein sample was loaded to 24 cm, pH 4–7 Immobiline dry strips (Amersham Biosciences) along with 1.5 μl (IPG)-buffer pH 4–7 dissolved in it for each strip.

The strips were focused on an IPG-phor (Amersham Biosciences) for 1 h at 0 V, 12 h at 30 V, 2 h at 60 V, 1 h at 1000 V, and at 8000 V until approximately 75,000 V h was reached. The strips were equilibrated in 5 ml of a solution containing 6 M urea, 50 mM Tris (pH 8.8), 30% (v/v) glycerol, 20 gL^{-1} SDS and 20 gL^{-1} DTT on a tilt table for 15 min. The solution was discarded and 5 ml of a second solution was added for 15 min containing 6 M urea, 50 mM Tris (pH 8.8), 30% (v/v) glycerol, 20 gL^{-1} SDS and 25 gL^{-1} iodoacetamide. The second dimension was performed on an EttanDalt (Amersham Biosciences) electrophoresis unit. The strips were placed on a 1.5-mm thick, 12.5% poly-acrylamide gel and sealed with 0.1% agarose in SDS-electrophoresis buffer containing 0.01% Bromphenol blue. The gel electrophoresis was performed for 30 min at 3 W per gel followed by a further run at 20 W per gel until the end. For comparative analysis, gels were stained with Coomassie blue stain.

In-gel tryptic digestion and mass spectrometry. Protein spots were excised from 2-DE gels with a spot picker and placed into 96-well microtiter plates, which were washed twice with TFA:acetonitrile:water (0.1:60:40). The tryptic digest was performed as reported previously with slight modifications [6]. The samples containing the tryptic-digested proteins were mixed at a 1:1 ratio with a solution of water:acetonitrile:TFA (67:33:0.1) saturated with α-cyano-cinnamic acid. The mass spectrum was obtained on a Ultraflex MALDI-TOF/TOF (Bruker). The annotation of the peptide mass fingerprints was performed by the MASCOT search engine (Matrix Science). The search was done against our local *E. coli* database. The parameters used were, Taxonomy: All entries; Enzyme: Trypsin; Missed cleavages: 1; ppm: 100; Database: *E. coli*.

Analysis of two-dimensional protein gels. For the comparison of the global proteome profiles of MG1655 in both the conditions (CAS-MINUS and CAS-PLUS), the gels were scanned and digitized. The protein spot densities between both the conditions were utilized for gel comparison. Image smoothing, spot detection, spot quantification, image alignment, spot matching, spot annotation, molecular weight and p*I* calculation, and variation analysis of the protein gels was performed using PDQuest software (Bio-Rad). For each protein spot, the annotated information along with the peak area and normalized quantity values were obtained. Along with these exported annotations, the protein spots were analyzed by grouping them into various functional categories based on MultiFun and Gene Ontology terms, the classification system for cellular functions of gene products of *E. coli* [4,5] consisting of 10 major functional categories.

Results and discussion

Construction and content

The database was designed to systematically input, store, compare, analyze, and output all the information related to an experiment. Proteome maps generated from 2-DE gel electrophoreses were scanned and the protein spots were subjected to tryptic digestion and identified by comparing the peptide masses, which were obtained on a Ultraflex MALDI-TOF/TOF (Bruker). The digitized gel images were normalized, and annotated using the PDQUEST program (version 6.2; Bio-Rad). For each protein spot, the annotated information along with the obtained *X* and *Y* coordinates, the peak area and normalized quantity values were stored in the database. Internally these data, along with the gel images (Fig. 1C and D) were stored by an upload function option in the database.

This database was created using the Make2D-DB II Package/version: 2.50.1 (http://expasy.org/ch2d/make2ddb/) [7] with additional characteristics (map comparison, map upload, protein spot comparison with functional classification, etc.) based on our modifications. The identified protein spots with the SWISS-PROT accession number were stored in a relational database that was made accessible online via a common gateway interface (cgi) script on a linux web server (Fig. 1A). Information pertaining to each protein spot can be accessed via a clickable gel image (Fig. 1B).

Expression levels of proteins can be compared one-by-one by viewing the corresponding protein spot image in a tabular format from the selected gels based on a scoring function, calculated based on the peak and normalized quantity values (see below). For each identified protein spot, the spot ID, SWISSPROT accession number, B number, name, function, peak value, normalized quantity, protein spot image, and score are available in a tabular format. Moreover, additional information can be accessed via links to the SWISS-PROT database.

For further analysis of the proteome data, we utilized MultiFun—the classification system for cellular functions of gene products of *E. coli* [4,5] consisting of 10 major functional categories. These major categories are further sub-divided into a hierarchical scheme. The complete hierarchical structure of the MultiFun classification can be screened for the expression of various proteins involved in a particular functional category in question from the selected gels. As a result, all the individual protein spots are summarized in a table consisting of vital information with the spot image along with the classification based on functional category which enhances the analysis.

Scoring function

To enhance the significance of the spot comparing feature we included a simple scoring function. The computed score rates the difference among two spots. It enables the user to efficiently compare and concentrate on specific proteins, which exhibit significant differential expression within the performed proteomic experiments. Furthermore, it offers the possibility to compare all the proteins spot-by-spot.

The computation of the score was based on the two available (from PDQuest) spot-specific values, Normalized Quantity (NQ) and Peak Value (PV). These values enabled us to characterize the correlative protein amount of each spot. Thus, the score is based on the difference of Normalized Quantity and Peak Value of two spots. But due to basic limitations of the 2D gel electrophoresis these values are not reliable when applied to the faint spots. This limitation made it necessary to include a rating (in terms of scoring function) of the basic intensity and size of the spot. Using this, we adjusted the score accordingly. Furthermore, we normalized NQ, PV and the computed differences to values between 1 and 100 to receive a better comparability of the scores.

To compute the scoring function, first we calculated the highest measured Peak Value (maxPVGel) and the highest measured Normalized Quantity (maxNQGel) of the two gels G_1 and G_2 which we want to compare. These values enabled us to rate the relative size and intensity of a spot. They would be computed as follows:

$$\mathrm{maxPVGel}_{G_1,G_2} = \max\left\{\bigcup_{j=1}^{n} \mathrm{PV}_{G_1(j)}, \mathrm{PV}_{G_2(j)}\right\} \quad (1)$$

In this formula the variable n is related to the number of spots on a gel and the variable j is used to represent the corresponding spots on both gels. Below, only the computations of the PV-related values are shown to ensure a better overview. The computation of the NQ-related values was performed in an equivalent way.

Additionally, while normalization we calculated the maximal PV-difference (maxΔPVGel) and the maximal NQ-difference (maxΔNQGel) of two corresponding spots on the gels which the user selects to compare. The related equation is showed here:

$$\max \Delta\mathrm{PVGel}_{G_1,G_2} = \max_{1\leqslant j\leqslant n}\left\{\left|\mathrm{PV}_{G_1(j)} - \mathrm{PV}_{G_2(j)}\right|\right\} \quad (2)$$

These gel and spot spanning values were combined with the specific values of each concrete spot comparison. Therefore it is necessary to compute the maximal PV-value (maxPVSpot) or the maximal NQ-value (maxNQSpot) of the viewed spot j, respectively, as follows:

$$\mathrm{maxPVSpot}_{G_1(j),G_2(j)} = \max\{\mathrm{PV}_{G_1(j)}, \mathrm{PV}_{G_2(j)}\} \quad \text{with} \quad 1 \leqslant j \leqslant n \quad (3)$$

As discussed before, the differences between the measured values of two faint spots are not reliable. Hence, we used the PV- and NQ-values to compute the score, respectively. For computing the score where only one of the two spots of the comparison is faint are expected to have low PV- and NQ-values which are of highly significance and this type of protein expression changes are not due to limitations of the 2D gel electrophoresis. These types of expression changes are of strong indication for significant changes in the proteome. Designating this, the score should represent these expression changes in a suitable way.

By using the maximal PV- and NQ-value of a spot comparison we could accomplish the requirement reliably. For example, a spot which is faint due to gel electrophoresis limitations would have similar low PV-values and NQ-values at both gels. Using the maximal PV- and NQ-value of these spots would certainly have a low score in our computation. In the case if a spot is faint on one gel but thick and dark on the other gel, the maximal PV- and NQ-value of these spots would certainly have a high score, which ignores the PV- and NQ-value of the faint spot. As a result, we obtain a score which are highly reliable based on the PV- and NQ-values of the thick and dark spots which is not influenced by the faint spots.

The following computation of the PV-difference (ΔPVSpot) and the NQ-difference (ΔNQSpot) of a pair of spots represents the base of the scoring function:

$$\Delta\mathrm{PVSpot}_{G_1(j),G_2(j)} = \left|\mathrm{PV}_{G_1(j)} - \mathrm{PV}_{G_2(j)}\right| \quad \text{with} \quad 1 \leqslant j \leqslant n \quad (4)$$

Utilizing the aforementioned derived values enabled us to compute intermediate scores for the concrete comparison of two spots. PVNQScore rates the relative size and intensity, respectively. It is computed as follows:

$$\mathrm{PVNQScore}_{G_1(j),G_2(j)} = \left(\frac{\mathrm{maxPVSpot}_{G_1(j),G_2(j)}}{\mathrm{maxPVGel}_{G_1,G_2}} + \frac{\mathrm{maxNQSpot}_{G_1(j),G_2(j)}}{\mathrm{maxNQGel}_{G_1,G_2}}\right) \times 50 \quad (5)$$

All scores were normalized to values between 1 and 100. PV- and NQ-related values of a spot were treated as equal characterisations for the amount of protein expressed. The computation of the relative PV-difference score (ΔPVScore) is shown in the following equation:

$$\Delta\mathrm{PVScore}_{G_1(j),G_2(j)} = \frac{\Delta\mathrm{PVSpot}_{G_1(j),G_2(j)}}{\max \Delta\mathrm{PVGel}_{G_1,G_2}} \times 100 \quad (6)$$

The relative NQ-difference (ΔNQScore) was determine in an equivalent way:

$$\Delta\mathrm{NQScore}_{G_1(j),G_2(j)} = \frac{\Delta\mathrm{NQSpot}_{G_1(j),G_2(j)}}{\max \Delta\mathrm{NQGel}_{G_1,G_2}} \times 100 \quad (7)$$

Based on several empirical tests we concluded that ΔPVScore as well as ΔNQScore provides good indications to rate the differences between two spots. Mostly the obtained high score was related to significantly differential expressed proteins. It was necessary to combine both the scores which would enhance the analysis of monitoring the differential expressed proteins. This combination was necessary because ΔPVScore and ΔNQScore does not correlate with each other in every case.

The overall score (Score) to rate the difference of spot j on gel G_1 in comparison with its corresponding spot j on gel G_2 results from the average of the aforementioned intermediate scores. It was computed as follows:

$$\mathrm{Score}_{G_1(j),G_2(j)} = \frac{\mathrm{PVNQScore}_{G_1(j),G_2(j)} + \Delta\mathrm{PVScore}_{G_1(j),G_2,(j)} + \Delta\mathrm{NQScore}_{G_1(j),G_2(j)}}{3} \quad (8)$$

By combining the spot-specific values (PV and NQ) along with the PVNQ intermediate score with the weighting ratio of 2:1 we were able to ensure two important things, on the one hand that the computation of the overall score is based on the differential expression of the proteins. On the other hand, the score ensures that the faint spots attains lower score than the thick and dark spots (higher scores) harboring similar relative differences between them.

Utility

To demonstrate the performance and the utility of the method used for comparing the proteome maps in this database, we applied this method to the proteome gels derived from a strain (MG1655) which was grown in the presence and absence of Casamino-acids (CAS), respectively (Fig. 1C and D). For this experiment approximately 4.1×10^{10} cells were harvested from M9 minimal medium culture supplemented with or without Casamino-acids from the exponential phase of growth. A protein sample of 300 μg was loaded on an Immobiline dry strip (Amersham Biosciences) with a length of 24 cm, p*I*-range 4–7. The first dimension was developed on an electrophoresis apparatus IPG-phor (Amersham Biosciences) until 75,000 V h was reached. The second dimension was developed by means of a 12.5% poly-acrylamide gel. For comparative analysis the gels were stained with coomassie blue. The excised protein spots were tryptic digested and the mass spectra were obtained on a Ultraflex MALDI-TOF/TOF (Bruker). Annotations of the peptide mass fingerprints were carried out by the MASCOT search engine (Matrix Science). The parameters used were: Taxonomy: all entries; Enzyme: trypsin; Missed cleavages: 1; ppm.: 100; Database: *E. coli*. 110 protein spots were analyzed in each gel, as a result 99 proteins could be annotated in both the gels. Many proteins were differentially expressed (Fig. 1C and D). Comprehensive screening of the functional category "metabolism" (1-MultiFun class) revealed substantial over-expression of the enzymes involved in the tryptophan amino-acid biosynthesis (1.5.1.15-MultiFun class) in the CAS-MINUS proteome map (Fig. 2A and Supplementary Fig. 1). Similarly by browsing the table of "transport" functional category (4-MultiFun class) revealed that proteins involved in the amino-acid transport were over-expressed in CAS-PLUS proteome map (Fig. 2B). Micro-organisms are known to utilize the available nutrient resources present in the environment rather than to synthesize the nutrients by themselves. Amino-acids present in the medium have to be transported into the cytoplasm for utilization. In line with these known phenomena, the proteins involved in the transport function were over-expressed in the sample grown in the presence of Casamino-acids (Fig. 2B) and on the other hand the protein involved in the amino-acid biosynthesis were over-

MG1655_CAS_MINUS : MG1655_CAS_PLUS

A) **1: Metabolism**

(Browse functional category Metabolism based on MultiFun classification)

Spot	Swissprot ID	B-Number	Gene	Function	Peak Value	Norm. Quantity	Gel 1	Peak Value	Norm. Quantity	Gel 2	Score ↑
6201	P0A879	b1261	trpB	tryptophan synthase, beta subunit	1.647	61093.8		0.121	3647.65		58
8303	P0A879	b1261	trpB	tryptophan synthase, beta subunit	1.263	38329.6		0.007	95.8		42
7701	P0A6P9	b2779	eno	enolase	1.926	30097.7		0.422	12680.7		37
6603	P0AD96	b3460	livJ	leucine/isoleucine/valine transporter subunit -I- periplasmic-binding component of ABC superfamily	0.013	1089.63		1.231	31312.3		36
6601	P23843	b1243	oppA	oligopeptide transporter subunit -I- periplasmic-binding component of ABC superfamily	0.707	39313		0.257	8765.98		28

B) **4: Transport**

(Browse functional category Transport based on MultiFun classification)

Spot	Swissprot ID	B-Number	Gene	Function	Peak Value	Norm. Quantity	Gel 1	Peak Value	Norm. Quantity	Gel 2	Score ↑
6603	P0AD96	b3460	livJ	leucine/isoleucine/valine transporter subunit -I- periplasmic-binding component of ABC superfamily	0.013	1089.63		1.231	31312.3		36
6601	P23843	b1243	oppA	oligopeptide transporter subunit -I- periplasmic-binding component of ABC superfamily	0.707	39313		0.257	8765.98		28
8309	P39172	b1857	znuA	zinc transporter subunit -I- periplasmic-binding component of ABC superfamily	0	3.72		0.692	23989.5		25
5102	P0AEM9	b1920	fliY	cystine transporter subunit -I- periplasmic-binding component of ABC superfamily	1.136	37768.4		0.797	32574.4		17
1	P69783	b2417	crr	glucose-specific enzyme IIA component of PTS	1.373	38663.2		1.256	33229.3		16

Fig. 2. MultiFun-metabolism and transport functional categories. (A) MultiFun-metabolism-functional category. Comprehensive screening of the functional category metabolism (1-MultiFun class) showing substantial over-expression of the enzymes involved in the tryptophan amino-acid biosynthesis in the CAS-MINUS (Gel 1) proteome map. (B) MultiFun- transport- functional category. The transport functional category (4- MultiFun class) showing the proteins involved in the amino-acid transport being over-expressed in CAS-PLUS (Gel 2) proteome map.

expressed in the sample grown in the absence of Casaminoacids (Fig. 2A and Supplementary Fig. 1). Examining the 2-DE gel protein spots one-by-one along with the functional classification enhanced the efficiency of the analyses enormously. This approach, when applied for multiple gels emerges as a valuable approach to analyze the available resources at once (Supplementary Fig. 2).

In our database, by examining the proteins in the proteome maps spot-by-spot individually aided with the alignment obtained form 2-DE gels analysis software we were able to combine and utilize the functional classification to enhance the analysis. Similar approaches of protein spot analysis would certainly improve the efficiency of comparison of the data generated from different conditions and from various laboratories. Furthermore, the approaches outlined here could be applied to the analysis of proteomic databases of other organisms.

Availability

The 2DBase of *E. coli* database can be accessed at http://2dbase.techfak.uni-bielefeld.de/.

Acknowledgment

The work was supported by a scholarship from the NRW International Graduate School in Bioinformatics and Genome Research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.09.050.

References

[1] V.C. Wasinger, S.J. Cordwell, A. Cerpa-Poljak, J.X. Yan, A.A. Gooley, M.R. Wilkins, M.W. Duncan, R. Harris, K.L. Williams, I. Humphery-Smith, Progress with gene-product mapping of the mollicutes: *Mycoplasma genitalium*, Electrophoresis 16 (1995) 1090–1094.

[2] J.E. Celis, H.H. Rasmussen, P. Gromov, E. Olsen, P. Madsen, H. Leffers, B. Honore, K. Dejgaard, H. Vorum, D.B. Kristensen, et al., The human keratinocyte two dimensional gel protein database (update 1995): mapping components of signal transduction pathways, Electrophoresis 16 (1995) 2177–2240.

[3] T. Rabilloud, Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains, Proteomics 2 (2002) 3–10.

[4] M.H. Serres, S. Goswami, M. Riley, Genprotec: an updated and improved analysis of functions of escherichia coli k-12 proteins, Nucleic Acids Res. 32 (2004) D300–D302.

[5] M.H. Serres, M. Riley, Multifun, a multifunctional classification scheme for *Escherichia coli* k-12 gene products, Microb. Comp. Genomics 5 (2000) 205–222.

[6] W.J. Henzel, T.M. Billeci, J.T. Stults, S.C. Wong, C. Grimley, C. Watanabe, Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases, Proc. Natl. Acad. Sci. USA 90 (1993) 5011–5015.

[7] K. Mostaguir, C. Hoogland, P.A. Binz, R.D. Appel, The Make 2D-DB II package: conversion of federated two-dimensional gel electrophoresis databases into a relational format and interconnection of distributed databases, Proteomics 3 (2003) 1441–1444.