Journal of
Proteomics & Bioinformatics

Corzett et al., J Proteomics Bioinform 2013, 6:9
DOI: 10.4172/jpb.1000282

Research Article Open Access

Multivariate Statistical Analysis of Diverse Strains of *Yersinia pestis* by Comparative Proteomics

Todd H Corzett[1], Angela M Eldridge[2], Jennifer S Knaack[3], Christopher H Corzett[1], Sandra L McCutchen-Maloney[1] and Brett A Chromy[1,2]*

[1]*Physical and Life Sciences Directorate Lawrence Livermore National Laboratory, Livermore, CA 94550, USA*
[2]*Department of Pathology and Laboratory Medicine, School of Medicine, University of California Davis, Sacramento, CA 95817, USA*
[3]*Department of Pharmaceutical Sciences, Mercer University, 3001 Mercer University Drive, Atlanta, GA 30341, USA*

Abstract

To address the difficulty in characterizing unusual, engineered or emergent pathogens in clinical and environmental samples, novel methods to discover proteins that differentiate pathogenic strains are needed. Differentially expressed proteins that reveal the function of an uncharacterized strain of bacteria can be considered biomarkers; panels of these can lead to improved pathogen classification and characterization. To this end, the protein expression patterns of differentially virulent isolates of the plague pathogen, *Yersinia pestis*, were studied using two-dimensional difference gel electrophoresis (2-D DIGE). The resulting characterization was used to identify a protein expression panel for the clustering and classification of *Y. pestis* strains. Two different methods were used to produce different biomarker panels based on either experimental- or pattern-based clustering. Each panel is able to successfully classify unknown samples in a blinded fashion, allowing an unbiased discovery of differentially expressed proteins, as well as the rapid classification of protein expression patterns.

Keywords: 2-D DIGE; Plague; *Yersinia pestis;* Biomarkers; Proteomics; Extended data analysis; Strain diversity; DeCyder; Chemometrics

Abbreviations: 2-D DIGE: Two-Dimensional Difference Gel Electrophoresis; IEF: Isoelectric Focusing; PMT: Photo Multiplier Tube; DIA: Differential In-gel Analysis; BVA: Biological Variance Analysis; EDA: Extended Data Analysis; PCA: Principle Component Analysis; PCR: Polymerase Chain Reaction; MALDI-MS: Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry

Introduction

Yersinia pestis, the etiological agent of plague, is endemic to the Southwestern United States, as well as other areas worldwide [1,2], and is the cause of the Justinian, Black Death and Hong Kong major historic pandemics [3]. *Y. pestis* is of current concern to human health because of its past use and future potential as a bioterrorism agent [4]. To gain a better understanding of the virulence mechanism of *Y. pestis*, we previously investigated the proteomic changes associated with the induction of virulence as a function of calcium and temperature on a common, attenuated laboratory strain [5,6].

Here, we characterize and compare the proteome of four *Y. pestis* strains, KIM5D 27, India-195/P, NYC, and PEXU2, with known diversity in origin, virulence level and countermeasure resistance. KIM5 D27, biovar Mediavalis, has a deletion of the *pgm* locus, and is conditionally avirulent [7]. This strain was included for direct comparison to previous proteomic [5,6] and proteogenomic studies [8]. India-195/P, NYC and PEXU2 are virulent strains from the Orientalis biovar. The India-195/P strain was clinically isolated in 1957 from a human bubonic plague patient in India [9], but has become attenuated due to the loss of the pgm locus during passage. The NYC strain was clinically isolated in November 2002 from a patient in New York City who had been exposed to plague in New Mexico, an endemic area [10,11]. The PEXU2 strain, isolated from a Brazilian rodent in 1966, was reported to have an elevated copy number of IS 100 elements, similar to the India-195/P strain [12].

In this study, *Y. pestis* strains were grown under conditions that mimic the induction of virulence and differential protein expression between the strains, and the growth conditions was analyzed by 2-D DIGE as previously reported [5,6]. The DeCyder software package (GE Healthcare, Piscataway, NJ) was used to process 2-D DIGE gel images, detect protein spots, match spots between gels and determine statistical differences in protein abundance levels [13,14]. The analytical complexity of 2-D DIGE necessitates the development of advanced analytical tools to interpret proteomic data. For example, it is possible to analyze proteomic data using statistical packages [15]. These statistical analysis software packages can result in increased confidence for the differential expression data, but require extensive statistical expertise [16-18]. The Extended Data Analysis (EDA) module of DeCyder serves as an alternative analysis tool which can provide multivariate statistics, such as principal component analysis (PCA) [19-22], hierarchical clustering [23-26], K-means clustering [27], and biomarker selection [28-30], and can be evaluated by scientists with less statistical expertise. Further, it is possible to generate a set of protein biomarkers or classifiers, which can be used to designate unknown samples [31,32], based solely on protein expression patterns. Although the sole use of pattern expression has proven problematic for bacterial identification and is not as rapid as PCR or MALDI-MS for bacterial detection, the results obtained from this approach may serve to complement these other detection techniques by further characterizing the particular bacteria under study. Here, we present results from advanced analytical analyses of *Y. pestis* comparative proteomic data and demonstrate correct classification of unknown samples.

***Corresponding author:** Brett A Chromy, Assistant Professor, Department of Pathology and Laboratory Medicine, School of Medicine, University of California Davis, Sacramento, CA 95817, USA, Tel: (530) 752-7229; E-mail: brett.chromy@ucdmc.ucdavis.edu

Received July 24, 2013; **Accepted** September 27, 2013; **Published** September 29, 2013

Citation: Corzett TH, Eldridge AM, Knaack JS, Corzett CH, McCutchen-Maloney SL, et al. (2013) Multivariate Statistical Analysis of Diverse Strains of *Yersinia pestis* by Comparative Proteomics. J Proteomics Bioinform 6: 202-208. doi:10.4172/jpb.1000282

Materials and Methods

Bacterial growths

Y. pestis strains were grown similarly to methods previously described [5]. Four strains of *Y. pestis* (KIM5 D27, India-195/P, NYC, and PEXU2) were grown in bovine calf serum media supplemented with either 0 mM or 4 mM calcium chloride at 37°C. To collect the soluble-cell proteome, cells were lysed using the B-PER reagent (Pierce, Rockford, IL), according to manufacturer's suggested protocols. Protein samples were cleaned using the PlusOne 2-D Clean-Up kit (GE Healthcare, Piscataway, NJ) and resuspended in 100 µL labeling buffer containing 7 M urea, 2 M thiourea, 4% CHAPS and 20 mM Tris (pH 8.5). Protein concentrations were then determined using Advanced Protein Assay reagent (ADV01, Cytoskeleton, Denver, CO).

2-D DIGE

The internal pooled standard, consisting of an equal amount (7.14 µg) of each of the 7 samples was labeled with 200 pmol 3-([4-carboxymethyl] phenylmethyl)-3'-ethyloxacarbocyanine halide N-hydroxysuccinimidyl ester (Cy2). This pooled standard allows for accurate matching and normalized protein abundance measurements across gels. 50 µg of each *Y. pestis* protein sample was labeled with either 200 pmol 1-(5-carboxypentyl)-1'-propylindocarbocyanine halide N-hydroxysuccinimidyl ester (Cy3) or 1-(5-carboxypentyl)-1'-methylindodicarbocyanine halide N-hydroxysuccinimidyl ester (Cy5) dyes, according to the experimental design (Table 1). The pooled standard was labeled with Cy2 and combined with a Cy5 and Cy3 protein sample, and electrophoresed on each gel, as shown in Table 1. "Pick" gels were supplemented with 200 µg of unlabeled pooled standard, in addition to the labeled protein samples to ensure sufficient protein was present for subsequent identification by mass spectrometry. Samples for each gel were adjusted to a total volume of 450 µL, with rehydration buffer consisting of 7 M urea, 2 M thiourea, 4% CHAPS, 1% Pharmalyte and 1.2% DeStreak (GE Healthcare, Piscataway, NJ), and loaded onto 24 cm, pH 4-7 nonlinear, Immobiline IPG DryStrips (GE Healthcare, Piscataway, NJ) for first dimension separation.

IEF separation was carried out using the Ettan IPGphor II (GE Healthcare, Piscataway, NJ), using the following running conditions: 30 V rehydration for 12 h, 500 V for 1 h, 1,000 V for 1 h and 8,000 V for 62,500 Vh. The IPG strips were then reduced for 15 min in equilibration buffer containing 2% SDS, 50 mM Tris-HCl pH 8.8, 6 M urea, 30% glycerol, 0.002% bromophenol blue and 10 mg/mL dithiothreitol. After reduction, the strips were alkylated for 15 minutes with equilibration buffer and 25 mg/mL iodoacetamide. The strips were then loaded onto 26 cm×20 cm precast 12.5% Tris-glycine polyacrylamide gels (Jule, Inc, Milford, CT), and run at 2 W/gel constant power at 22°C using an Ettan DALT 12 (GE Healthcare, Piscataway, NJ), until the bromophenol blue dye-front reached the end of the gels (approximately 16 h).

Gel No.	Cy2	Cy3	Cy5
1	Standard	India-195/P (0mM Ca^{2+})	India-195/P (4mM Ca^{2+})
2	Standard	KIM5 D27 (0mM Ca^{2+})	KIM5 D27 (4mM Ca^{2+})
3	Standard	NYC (0mM Ca^{2+})	NYC (4mM Ca^{2+})
4	Standard	India-195/P (0mM Ca^{2+})	KIM5 D27 (0mM Ca^{2+})
5	Standard	India-195/P (4mM Ca^{2+})	KIM5 D27 (4mM Ca^{2+})
6	Standard	KIM5 D27 (0mM Ca^{2+})	NYC (0mM Ca^{2+})
7	Standard	KIM5 D27 (4mM Ca^{2+})	NYC (4mM Ca^{2+})
8	Standard	NYC (0mM Ca^{2+})	PEXU2 (0mM Ca^{2+})
9	Standard	PEXU2 (0mM Ca^{2+})	India-195/P (0mM Ca^{2+})
10	Standard	NYC (4mM Ca^{2+})	PEXU2 (0mM Ca^{2+})
11[a]	Standard	India-195/P (4mM Ca^{2+})	KIM5 D27 (4mM Ca^{2+})
12[b]	Standard	NYC (0mM Ca^{2+})	KIM5 D27 (0mM Ca^{2+})

[a]Master gel
[b]Additional unlabeled protein added for protein identification

Table 1: 2-D DIGE experimental design.

Gels were imaged using a Typhoon 9410 imager (GE Healthcare, Piscataway, NJ), with a 100 µm resolution. PMT values were adjusted for the optimization of sensitivity and prevention of oversaturation. Cy2 dye was excited at 488 nm and emission spectra obtained at 510 nm; Cy3 dye was excited at 550 nm and emission spectra obtained at 570 nm; Cy5 dye was excited at 650 nm and emission spectra obtained at 670 nm. Unlabeled protein on the pick gel was visualized by post-staining with SYPRO Ruby (Bio-Rad, Hercules, CA), and imaged on the Typhoon 9410. All gel images were cropped to the same size using ImageQuant v5.2 (GE Healthcare, Piscataway, NJ), to remove the edges of the gels while maximizing the number of spots available for analysis.

Gel images, three from each CyDye labeled gel and one additional SYPRO image of the pick gel, were examined using the various modules of the DeCyder v6.5 software package. The Differential In-gel Analysis (DIA) module was used to determine optimal spot detection settings. Images were loaded into the Batch Processor module and spot maps were generated from each gel image, with the estimated number of spots set to 2,500. The automated determination of the master gel, or the gel with the most spots, was bypassed, and the gel displaying the best spot characteristics was labeled the master gel. During batch processing, the Cy2 channel from each gel was used for normalization of spot intensities and for automated matching between gels. After batch processing, the resulting data was further processed using multiple analysis techniques.

Differentially expressed spot determination

To determine which protein spots were differentially expressed between the strains and growth conditions, the 2-D DIGE data was examined using the Biological Variation Analysis (BVA) module of DeCyder. The quality of gel matching was manually verified and landmarks were added when needed to improve match quality. Landmarks are manually validated matched spots that are linked together to help subsequent matching of other adjacent and closely aligned spots. Landmarking joins unmatched spots or fixes incorrectly matched spots, and can be made on both manually matched and computer-derived matched spots. Spot maps from each sample were assembled into individual experimental groups. Spots having a greater than 1.5-fold change in expression between experimental groups, with a P-value ≤ 0.05 and a one-way ANOVA ≤ 0.05, were distinguished as differentially expressed and investigated further. Protein spots of interest were manually verified to be of sufficient quality for mass spectrometry by examining the three-dimensional profile of the protein spot. Artifacts or those spots with volumes close to the background were excluded from additional analyses. The verified spots of interest were then imported into the Extended Data Analysis (EDA) module of DeCyder. A Base Set was created containing spots that were matched on greater than 75% of the spot maps, and that included expression information for all of the experimental groups (critical for classification of unknown samples). Using a 75% or greater value for matching provided a balanced approach. A lower percentage would reduce data quality as poorly matched, or too few n values would result. If a higher percentage of matching were required, less total spots would be available for analysis, which may have made a biomarker selection panel inaccurate. The Base Set of protein spots and spot maps was used

for further EDA analysis. The entire Base Set was then analyzed using PCA and hierarchical clustering to identify protein expression trends.

Two methods of grouping protein spots were employed to find putative biomarkers that distinguish all the experimental groups, and for the categorization of unknown samples. First, the experimental conditions of the samples (i.e. strain and calcium concentration) were used to find the differential spots with similar expression, referred to here as the "experimentally-based method," putative biomarkers specific to the experimental samples were selected. Groups of spots were created for each strain comparison (example, India vs. NYC), and for each growth condition for each strain (e.g., NYC 0 mM calcium vs. NYC 4 mM calcium). Biomarker selection was performed on each group of differentially expressed protein spots and the minimum number of spots required for the greatest discrimination accuracy was selected for each 'experimental' comparison. Second, K-means clustering was used to group spots showing similar patterns of protein expression, a technique referred to here as the "pattern-based method." The spots from each of these patterns were then subjected to biomarker selection, where the minimum number of spots required for the greatest discrimination accuracy, reported as the percent accuracy of class determination [33], were selected. Spots selected as biomarkers from both methods were cross-validated by hierarchical clustering to determine whether selected spots displayed similar trends to the overall dataset, and could be used to accurately classify the samples. Spots can then be used to generate a "pick list" for robotic spot picking and identification by mass spectrometric analysis, as previously described [34].

Rapid identification of "unknown" samples

To demonstrate the ability to correctly categorize "blinded" or "unknown" samples, one spot map (out of four replicates) for three of the experimental samples (KIM5 D27 0 mM calcium, KIM5 D27 4 mM calcium and NYC 0 mM calcium) was randomly chosen and removed from the analyses to function as "unknowns". To reduce the time requirement for this analysis, automated spot matching was used with no manual verification. Spots having greater than 1.5-fold change in expression between experimental groups, with a P-value ≤ 0.05 and a one-way ANOVA ≤ 0.05, were tagged as putative biomarkers for classification of the experimental samples. These spots were then used to create a classifier, or a mapping of the experimental groups to their corresponding expression patterns, to classify "unknown" samples based upon protein expression profiles. By removing the "unknown" spot maps prior to the creation of the classifier, the classifier could be produced independent of the "unknown" spot map and classification was therefore based only on the expression signatures of the known experimental samples analyzed. To test the classifier, the unknown spot maps were then reintegrated into the analysis and classified based on the likeness of expression profiles to the list of putative biomarkers.

Results and Discussion

Four strains of *Y. pestis* (KIM5 D27, NYC, PEXU2, and India-195/P) were grown at 37°C under calcium concentrations known to induce virulence, and to simulate the observed response when *Y. pestis* is transmitted from the flea vector (higher calcium concentration) to the infected host (lower calcium concentration) [35-38]. The four strains were chosen to represent diversity of origin, as well as diversity in virulence level. Our preliminary work in a mouse model indicates that the NYC strain is 1000-fold more virulent than the India-195/P strain, as demonstrated by established ED_{50} levels.

Following co-electrophoresis of the protein samples labeled with three fluorescent dyes, Cy2, Cy3 and Cy5, images were obtained for each experimental sample. Gels were batch-processed and an average of 1918 spots was detected on each gel. The gel displaying the best spot characteristics, containing 1952 spots, was designated the master gel. Automated spot matching resulted in an average of 1182 spots per gel being matched to the master gel. To maximize the quality of matching between the gels, an average of 201 spots were manually landmarked on each gel. Post-landmarking, the matching resulted in an average of 1211 spots being matched to the master gel image and 1409 spots being matched on the SYPRO Ruby image. Manual investigation of the quality of automated spot matching and the addition of landmarks in areas of poor or incorrect automated spot matching prior to further analysis is recommended. Presently, improvements in automated matching algorithms are needed to reduce the extent of landmarking necessary for data analysis.

Differentially expressed spot determination

Of the spots that were matched to the master gel, 679 exhibited differential expression of more than 1.5-fold for at least one comparison between the experimental groups consisting of the strain and growth differences. After manual verification of the differential spots, 446 spots remained for further investigation. The majority of the spots that did not pass manual verification was present at levels close to background, and did not have characteristics of protein spots. The sensitive detection parameters used in this study, while allowing for the detection of low abundant proteins, resulted in increased detection of artifacts that necessitated manual verification. The manually verified spots were then imported into the EDA module of DeCyder and a Base Set was created. After selecting spots that were found on more than 75% of the spotmaps, the remaining Base Set of 401 proteins was used to perform PCA (Figure 1) and hierarchical clustering (Figure 2) to determine trends in the protein expression profiles.

In the PCA, all strains were well spread from one another, with no overlap, suggesting substantial strain-to-strain differences (Figure 1). Some growth conditions, specifically those for the KIM5 D27 and NYC strains, showed large separation, while the India-195/P strain exhibited little separation between the two growth conditions. The lower level of differentiation between the two growth conditions of India-195/P, visualized in the PCA, may be evidence of a decreased low calcium response, and may reflect the lower virulence level of this strain.

Hierarchical clustering corroborated the results found using PCA, as the detected expression trends grouped the samples first by replicate samples of the strain of *Y. pestis*. Clustering further showed that growth condition (0 mM and 4 mM calcium chloride) for each strain was more closely grouped than each different strain to each other, as evidenced by the placement of the tree branches in the clustering heat map (Figure 2). An interesting contrast involves the hierarchical clustering of India-195/P, which resulted in marked reduction in differences between the two growth conditions seen for both PCA (Figure 1) and clustering (Figure 2). Taken together, these data suggest proteome differences are greater between strains than between growth in different calcium concentrations. The greater dissimilarity between each of the four strains was unexpected considering the significant levels of proteomic differences observed during the low calcium response in previous studies [5,6]. The results, however, indicate that there are significant protein expression changes between *Y. pestis* strains that may prove useful for detection and threat characterization of unknown, and/or uncharacterized strains.

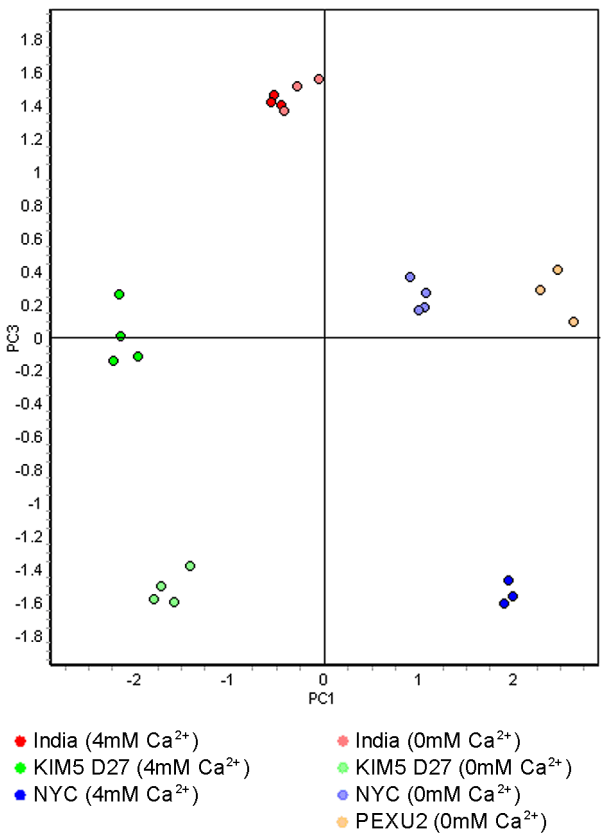

Figure 1: Principle component analysis of the 401 manually verified differentially expressed protein spots. Each dot represents an individual spot map of the gel. Each of the experimental groups is clearly separated from one another, without overlapping of data points, suggesting substantial strain-to-strain differences highlighted in their proteomic spot maps. There is also substantial spread between each of the growth conditions for the KIM5 D27 and NYC strains.

Biomarker selection in EDA was used on the Base Set to select putative biomarkers from the 401 differentially expressed protein spots that could distinguish between the experimental groups. The numbers of spots analyzed for biomarker selection increased the time required to process the data. Another consequence of processing large numbers of spots was the potential for decreased accuracy of the biomarker selection. This was due to the fact that once 100% accuracy was achieved for a set of spots (which only required five spots in the analysis of the entire 401 spot Base Set), the remainder of the spots were added independent of their importance in differentiating the experimental groups. To allow for more accurate and robust biomarker selection, the Base Set was broken into smaller groups prior to biomarker selection.

To investigate the observed protein expression patterns, two different methods, experimental-based and pattern-based, were used for biomarker selection. Spots found to be differential for each of the strain and growth comparisons were assembled into groups using experimental-based method (Table 2) for biomarker studies. Each group was comprised of between 10 to 164 spots, with a total of 323 unique spots being found differential for one or more of the comparisons listed. Marker selection was performed on each group and the number of spots selected, and the classification accuracy for each group is reported. A total of 37 spots were selected that could be used to distinguish the experimental groups (Table 2). Spots showing similar patterns independent of the experimental condition (pattern-based) were assembled into 8 groups using K-means clustering (Table 3), ranging from 6 to 102 spots per group. Biomarker selection was performed on each group and the number of spots selected, along with the classification accuracy of the spots selected for each group is reported. A total of 49 spots were selected that could be used to distinguish the experimental groups using this pattern-based method (Table 3).

The experimentally-based method of biomarker selection identified

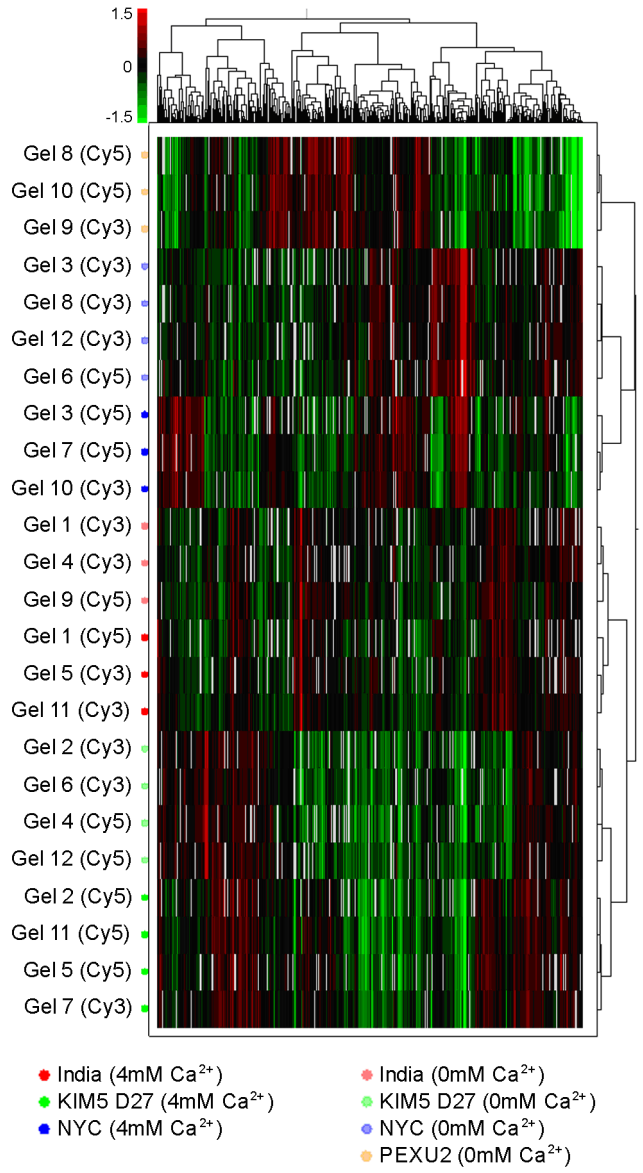

Figure 2: Hierarchical clustering of the 401 differentially expressed protein spots from the Base Set. The spot maps are displayed in rows and the protein spots by column. Differential expression is shown by color, with the red spots being up-regulated and the green being down-regulated relative to the pooled standard. The clustering shows that spot maps are grouped first by the replicate samples of each strain (KIM5 D27, India, NYC, and PEXU2), then by the growth condition (0 mM and 4 mM calcium chloride), and finally between strains, suggesting proteomic spot maps show more differences between strains than they do for different calcium growth conditions.

Group	Comparison	(left vs. right)	Differential Proteins[a]	Markers Selected[a]
1	India-195/P	KIM5 D27	63	6
2	India-195/P	NYC	39	8
3	India-195/P	PEXU2	154	6
4	KIM5 D27	NYC	109	5
5	KIM5 D27	PEXU2	164	6
6	NYC	PEXU2	147	6
7	India-195/P (4mM Ca^{2+})	India-195/P (0mM Ca^{2+})	10	3
8	KIM5 D27 (4mM Ca^{2+})	KIM5 D27 (0mM Ca^{2+})	54	6
9	NYC (4mM Ca^{2+})	NYC (0mM Ca^{2+})	81	9
Total[a]			323	37

[a]Spots may be listed as differential among multiple comparisons, but can only be listed once in the total number of proteins. All marker selection groups resulted in 100% accuracy period

Table 2: The number of differentially expressed protein spots per group for a direct comparison of the experimental conditions (experimentally-based method).

Group	Differential Proteins	Markers Selected	Accuracy (%)
1	78	6	100
2	102	5	100
3	42	5	100
4	81	5	100
5	12	11	83.34
6	74	7	100
7	6	6	80.95
8	6	4	76.19
Total	401	49	

Table 3: The number of differentially expressed protein spots per group for a K-means clustering analysis based on expression patterns (pattern-based method).

37 biomarkers (Figure 3) and requires differentiation of spots based on the strain and growth condition. As a result, the number of groups and the potential selection is heavily biased toward the experimental groups, and is therefore, more applicable to comparative proteomic analyses with multiple experimental conditions. In addition, since the spots are not grouped by the expression patterns determined by the proteomic analysis, the resulting clustering may not resemble that of the entire Base Set (Figure 3 clustering as compared to Figure 2). The pattern-based method identified 49 biomarkers (Figure 4), and is unaffected by experimental conditions since experimental groups were not directly factored into the grouping or selection, so this type of analysis can be applied to all sample types. For example, this approach is well-suited for identifying a particular sample with unknown growth conditions or strain identification. The two selection methods identified 16 biomarkers in common, suggesting that these are particularly important biomarkers for discriminating between strains and growth conditions. Combining the 37 experimentally-based and 49 pattern-based spots resulted in 70 unique spots providing another mechanism for producing a biomarker panel that could be used to distinguish the multiple strains and growth conditions.

Rapid identification of "unknown" samples

Automated spot detection and matching data that was imported into the EDA module for biomarker identification was used to rapidly classify three protein expression patterns that were removed from the analyses to serve as "blinded" or "unknown" samples. Of the 1182 matched spots, 607 exhibited significant differential expression of more than 1.5-fold for one or more of the experimental group comparisons.

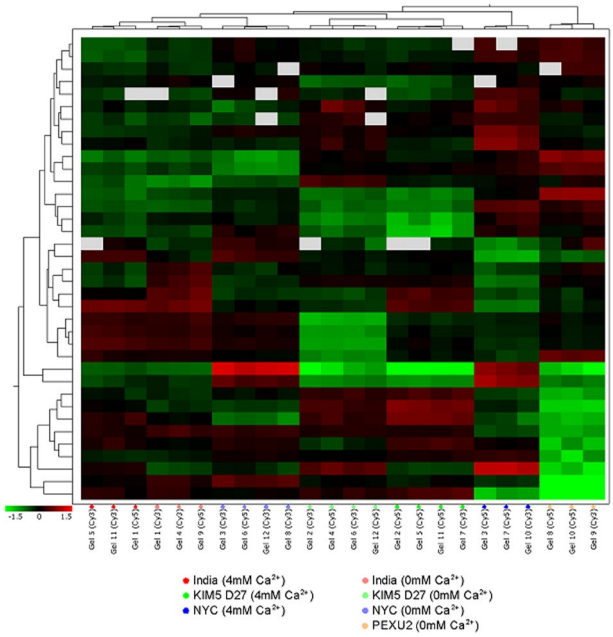

Figure 3: Hierarchical clustering of the 37 differential protein spots selected using the experimentally-based marker selections method. The spot maps are displayed in columns and the protein spots by rows. Differential expression is shown by color, with the red boxes being upregulated and the green being down-regulated relative to the pooled standard. The clustering fails to group the 0 mM Ca^{2+} (light blue) and 4 mM Ca^{2+} (dark blue) growth conditions for the NYC strain together, which differs from the clustering seen with the entire Base Set (see Figure 2) and the pattern-based method (Figure 4).

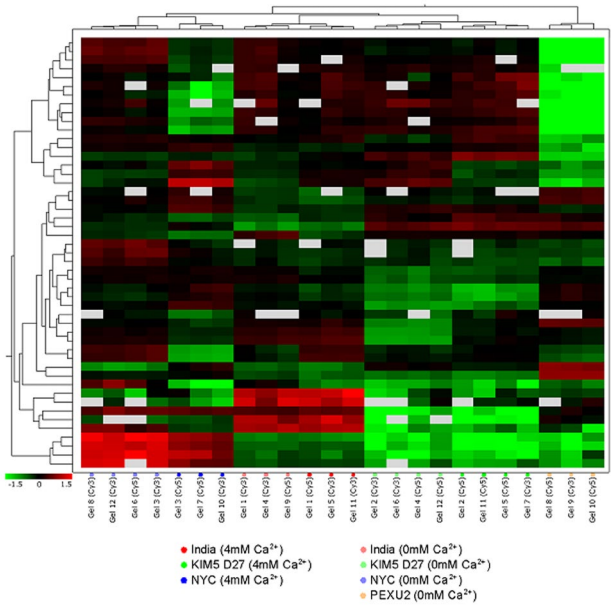

Figure 4: Hierarchical clustering of the 49 differential protein spots selected using the pattern-based marker selections method. The spot maps are displayed in columns and the protein spots by rows. Differential expression is shown by color, with the red boxes being up-regulated and the green being down-regulated relative to the pooled standard. The clustering shows that spot maps are grouped first by the replicates of each strain (KIM5 D27, India, NYC and PEXU2), then by the growth condition (0 mM and 4 mM calcium chloride), and finally between different strains. This hierarchy was the same when using the entire Base Set (see Figure 2).

The differential protein spots were then imported into the EDA module and a Base Set of 424 spots was used to create a classifier, or map of differential proteins necessary for classification. The classifier corresponding to each experimental group was evaluated for its ability to correctly classify each known spot map into the correct experimental group. After the classifier was validated, three "unknown" spot maps were introduced into the analysis (Figure 5), and the classifier was used to match the "unknown" samples with the correct strain and growth condition based on protein expression patterns. The successful classification of the three "unknown" samples into the appropriate experimental group demonstrates that protein expression patterns can be used to identify or characterize an unknown sample. Such protein expression maps can be developed and used to identify strains of infection in various life cycles of *Y. pestis*, such as within the flea vector or human host. These maps could greatly facilitate identification leading to a more rapid outbreak response and aid in data analysis for epidemiologic studies.

Comparison of the two analytical approaches

To find specific biomarkers that are able to differentiate two known parameters, the experimental-based method is more applicable, but to determine general trends in protein expression, the pattern-based approach will provide less experimentally-biased biomarkers. The major differences between the two approaches are the landmarking and manual verification steps required to successfully analyze the data. Substantial effort was required for landmarking and manual verification of the differentially expressed protein spots, when using the experimental-based method. While landmarking and manual verification reduce error associated with the analysis, this study demonstrates that by using EDA after automated spot matching, classification of unknown samples is possible using either experimental- or pattern-based methods. These analyses of 2-D DIGE proteomic data can contribute to a systems-based approach for biomarker discovery and pathogen characterization, adding to current datasets using transcriptomics, mass spectrometry-based proteomics and metabolomics [39]. The analysis described here may not be the best technique for studying bacterial detection, and/or identification, but can help characterize samples that are shown to be quite similar using sensitive, rapid approaches for identification, such as MALDI-MS or PCR. Genomic sequencing is developing into a more rapid, comprehensive approach for bacterial identification, but may require a complementary approach for analysis of the post-translational characterization of a sample, suggesting another potential use for the approach used in this study.

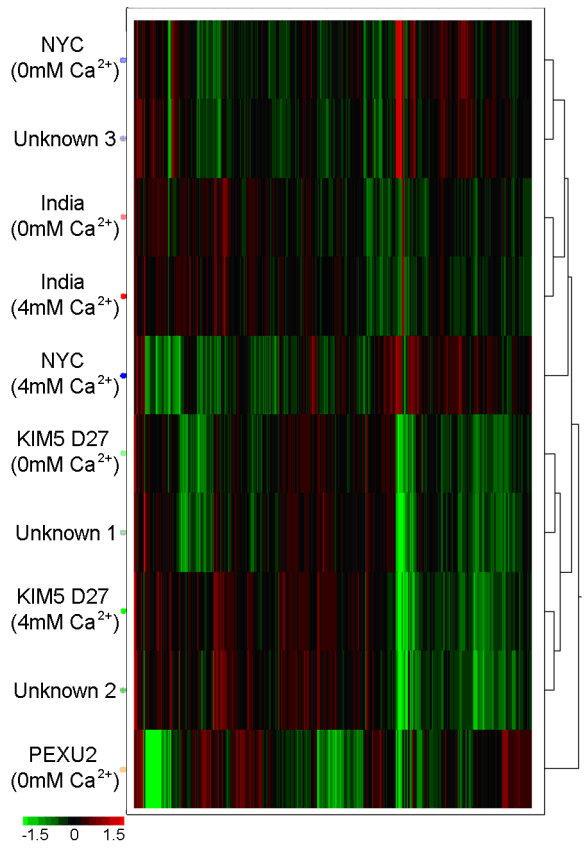

Figure 5: Hierarchical clustering of the 472 differential protein spots used in the classifier to determine clustering of 'unknown' samples. The experimental groups are displayed in rows and the protein spots by column. Differential expression is shown by color, with the red boxes being up-regulated and the green being down-regulated relative to the pooled standard. The unknown samples (Unknown 1, Unknown 2, and Unknown 3) were accurately grouped with their corresponding experimental groups (KIM5 D27 4 mM calcium chloride, KIM5 D27 0 mM calcium chloride, NYC 0 mM calcium chloride), respectively.

Conclusion

Here, we describe multivariate statistical analysis of 2-D DIGE experimental data using DeCyder EDA. The protein expression profiles of four diverse *Y. pestis* strains were compared to determine how proteomic diversity between multiple strains of the same species can enhance current pathogen detection strategies, such as genomic sequencing and MALDI-MS. Two different biomarker panels based on either experimental- or pattern-based methods were obtained. Each panel is able to successfully classify blinded "unknown" samples. The analytical techniques used here allow for the unbiased identification of differentially expressed proteins, as well as the rapid classification of protein expression patterns. Either of these two approaches for biomarker selection can be used to improve the characterization of bacteria or other organisms with similar proteomes using 2-D DIGE.

Acknowledgements

The authors acknowledge funding from the Department of Homeland Security (Biological Countermeasures Program to SMM), Lawrence Livermore National Laboratory, Laboratory Directed Research and Development award (08-ERD-020 to BAC), and the Proteomics Initiative of the Department of Pathology and Laboratory Medicine at U.C. Davis School of Medicine (BAC). The portion of the work carried out at Lawrence Livermore National Laboratory was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

References

1. Adjemian JZ, Foley P, Gage KL, Foley JE (2007) Initiation and spread of traveling waves of plague, *Yersinia pestis*, in the western United States. Am J Trop Med Hyg 76: 365-375.
2. Anisimov AP, Lindler LE, Pier GB (2004) Intraspecific diversity of *Yersinia pestis.* Clin Microbiol Rev 17: 434-464.
3. Perry RD, Fetherston JD (1997) *Yersinia pestis*-Etiologic agent of plague. Clin Microbiol Rev 10: 35-66.
4. Inglesby TV, Dennis DT, Henderson DA, Bartlett JG, Ascher MS, et al. (2000) Plague as a biological weapon: medical and public health management. Working Group on Civilian Biodefense. JAMA 283: 2281-2290.
5. Chromy BA, Choi MW, Murphy GA, Gonzales AD, Corzett CH, et al. (2005) Proteomic characterization of *Yersinia pestis* virulence. J Bacteriol 187: 8172-8180.

6. Hixson KK, Adkins JN, Baker SE, Moore RJ, Chromy BA, et al. (2006) Biomarker candidate identification in *Yersinia pestis* using organism-wide semiquantitative proteomics. J Proteome Res 5: 3008-3017.

7. Buchrieser C, Rusniok C, Frangeul L, Couve E, Billault A, et al. (1999) The 102-kilobase pgm locus of *Yersinia pestis*: sequence analysis and comparison of selected regions among different *Yersinia pestis* and *Yersinia pseudotuberculosis* strains. Infect Immun 67: 4851-4861.

8. Payne SH, Huang ST, Pieper R (2010) A proteogenomic update to *Yersinia:* Enhancing genome annotation. BMC Genomics 11: 460.

9. Chen TH, Foster LE, Meyer KF (1961) Experimental comparison of the immunogenicity of antigens in the residue of ultrasonated avirulent *Pasteurella pestis* with a vaccine prepared with killed virulent whole organisms. J Immunol 87: 64-71.

10. Centers for Disease Control and Prevention (CDC) (2003) Imported plague-New York City, 2002. MMWR Morb Mortal Wkly Rep 52: 725-728.

11. Guarner J, Shieh WJ, Chu M, Perlman DC, Kool J, et al. (2005) Persistent *Yersinia pestis* antigens in ischemic tissues of a patient with septicemic plague. Hum Pathol 36: 850-853.

12. Hinchliffe SJ, Isherwood KE, Stabler RA, Prentice MB, Rakin A, et al. (2003) Application of DNA microarrays to study the evolutionary genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*. Genome Res 13: 2018-2029.

13. Gharbi S, Gaffney P, Yang A, Zvelebil MJ, Cramer R, et al. (2002) Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system. Mol Cell Proteomics 1: 91-98.

14. Bech-Serra JJ, Borthwick A, Colomé N, ProteoRed Consortium, Albar JP, et al. (2009) A multi-laboratory study assessing reproducibility of a 2D-DIGE differential proteomic experiment. J Biomol Tech 20: 293-296.

15. Marengo E, Robotti E, Bobba M (2008) 2D-PAGE maps analysis. Methods Mol Biol 428: 291-325.

16. Corzett TH, Fodor IK, Choi MW, Walsworth VL, Chromy BA, et al. (2006) Statistical analysis of the experimental variation in the proteomic characterization of human plasma by two-dimensional difference gel electrophoresis. J Proteome Res 5: 2611-2619.

17. Fodor IK, Nelson DO, Alegria-Hartman M, Robbins K, Langlois RG, et al. (2005) Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using DeCyder. Bioinformatics 21: 3733-3740.

18. Carpentier SC, Panis B, Swennen R, Lammertyn J (2008) Finding the significant markers: Statistical analysis of proteomic data. Methods Mol Biol 428: 327-347.

19. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (2001) Multi-and megavariate data analysis.

20. Umetrics (2013) SIMCA, Sweden, UK.

21. Wold H (1966) Estimation of principal components and related models by iterative least squares in Multivariate Analysis. Academic Press, NY, USA.

22. Lamoureux L, Simon SL, Plews M, Ruddat V, Brunet S, et al. (2013) Urine proteins identified by two-dimensional differential gel electrophoresis facilitate the differential diagnoses of scrapie. PLoS One 8: e64044.

23. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868.

24. Everitt BS, Landau S, Leese M (2001) Cluster Analysis. (4th Ed), Wiley, USA.

25. Sokal R, Mitchener C (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38: 1409-1438.

26. Tibshirani RG, Walther, Hastie T (2001) Estimating the number of clusters in a dataset via the gap statistic. J R Statist Soc B 63: 411-423.

27. Lloyd S (1982) Least squares quantization in pcm. IEEE Trans Inf Theory 28: 128-137.

28. Webb AR (2002) Statistical pattern recognition. (2nd Ed), Wiley, USA.

29. Witten IH, Frank E (2000) Data Mining. Morgan Kaufman, Massachusetts, USA.

30. Kim H, Watkinson J, Anastassiou D (2011) Biomarker discovery using statistically significant gene sets. J Comput Biol 18: 1329-1338.

31. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537.

32. Lee JW (2009) Method validation and application of protein biomarkers: basic similarities and differences from biotherapeutics. Bioanalysis 1: 1461-1474.

33. DeCyder 2D V6.5 (2005) EDA User Manual: GE Healthcare 330.

34. Mahnke RC, Corzett TH, McCutchen-Maloney SL, Chromy BA (2006) An integrated proteomic workflow for two-dimensional differential gel electrophoresis and robotic spot picking. J Proteome Res 5: 2093-2097.

35. Bölin I, Forsberg A, Norlander L, Skurnik M, Wolf-Watz H (1988) Identification and mapping of the temperature-inducible, plasmid-encoded proteins of *Yersinia* spp. Infect Immun 56: 343-348.

36. Cornelis GR (2002) *Yersinia* type III secretion: Send in the effectors. J Cell Biol 158: 401-408.

37. Heesemann J, Gross U, Schmidt N, Laufs R (1986) Immunochemical analysis of plasmid-encoded proteins released by enteropathogenic *Yersinia* sp. grown in calcium-deficient media. Infect Immun 54: 561-567.

38. Ramamurthi KS, Schneewind O (2002) Type iii protein secretion in *Yersinia* species. Annu Rev Cell Dev Biol 18: 107-133.

39. Ansong C, Schrimpe-Rutledge AC, Mitchell HD, Chauhan S, Jones MB, et al. (2013) A multi-omic systems approach to elucidating *Yersinia* virulence mechanisms. Mol Biosyst 9: 44-54.