Chromatography
Separation Techniques

Deshpande et al. J Chromatograph Separat Techniq 2011, S5
DOI: 10.4172/2157-7064.S5-001

Research Article Open Access

ABOid: A Software for Automated Identification and Phyloproteomics Classification of Tandem Mass Spectrometric Data

Samir V. Deshpande[1]*, Rabih E. Jabbour[2], Peter A. Snyder[2], Michael Stanford[2], Charles H. Wick[2] and Alan W. Zulich[2]

[1]*Science and Technology Corporation, 500 Edgewood Road, Suite 205, Edgewood, MD 21040, USA*
[2]*U.S. Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD 21010, USA*

Abstract

We have developed suite of bioinformatics algorithms for automated identification and classification of microbes based on comparative analysis of protein sequences. This application uses sequence information of microbial proteins revealed by mass spectrometry-based proteomics for identification and phyloproteomics classification. The algorithms transforms results of searching product ion spectra of peptide ions against a protein database, performed by commercially available software (e.g. SEQUEST), into a taxonomically meaningful and easy to interpret output. To achieve this goal we constructed a custom protein database composed of theoretical proteomes derived from all fully sequenced bacterial genomes (1204 microorganisms as of August 25th, 2010) in a FASTA format. Each protein sequence in the database is supplemented with information on a source organism and chromosomal position of each protein coding open reading frame (ORF) is embedded into the protein sequence header. In addition this information is linked with a taxonomic position of each database bacterium.

ABOid analyzes SEQUEST search results files to provide the probabilities that peptide sequence assignments to a product ion mass spectrum (MS/MS) are correct and uses the accepted spectrum–to-sequence matches to generate a sequence-to-organism (STO) matrix of assignments. Because peptide sequences are differentially present or absent in various strains being compared this allows for the classification of bacterial species in a high throughput manner. For this purpose, STO matrices of assignments, viewed as assignment bitmaps, are next analyzed by a ABOid module that uses phylogenetic relationships between bacterial species as a part of decision tree process, and by applying multivariate statistical techniques (principal component and cluster analysis), to reveal relationship of the analyzed unknown sample to the database microorganisms. Our bacterial classification and identification algorithm uses assignments of an analyzed organism to taxonomic groups based on an organized scheme that begins at the phylum level and follows through classes, orders, families and genus down to strain level.

Introduction

It is a challenging task to translate the raw proteomics data, generated during high throughput MS experiments, into biologically meaningful results that are suitable for the classification and identification of agents of biological origin (ABO).

Automated detection and identification of pathogenic microorganisms is highly important in many areas of public health. Currently, more than 1200 bacteria have been fully sequenced and more 541 sequencing projects are in progress as of August 25th, 2010. Completely sequenced genomes provide amino acid sequence information of every protein potentially expressed by these organisms. Hence, the combination of this resource with mass spectrometry (MS) technologies capable of identifying amino acid sequences of proteins [1] enables one to design new procedures for the identification and classification of bacteria based on querying proteomic sequences.

To this end, gel-free proteomic procedures based on coupling liquid chromatography (LC) electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI) with tandem mass spectrometry (MS/MS) analysis of peptides generated from cellular proteins is being developed as an attractive technological platform for identification and classification of microorganisms [2-6].

Although the MS/MS based sequencing of peptides by using database search engines or by *de novo* sequencing of peptides are common practices [7], it is still a challenging task how to translate the raw data generated from MS/MS experiments into biologically meaningful and easy to interpret results suitable for identification and classification of microorganisms with high confidence.

We present a suite of bioinformatics algorithms for an automated identification and classification of bacteria based on the peptide sequence information generated from LC ESI MS/MS analysis of tryptic digests of bacterial protein extracts and profiling of the sequenced peptides to create a matrix of sequence-to-organism (STO) assignments. Using database bacteria proteomes as a reference, we have developed an unsupervised approach to reveal the relatedness between the test and database microorganisms. This binary matrix is analyzed using diverse visualization and multivariate statistical techniques for bacterial classification and identification.

The relevance and consistency of this suite of algorithms, named ABOid, for the identification of bacteria is demonstrated by using an illustrative example of processing MS/MS spectra of peptide ions obtained during LC-MS analysis of a model bacterial mixture.

***Corresponding author:** Samir V. Deshpande, Science and Technology Corporation, 500 Edgewood Road, Ste 205, Edgewood, MD 21040, USA, Tel: 410-436-4348; Fax :410-436-1912 ; E-mail: samir.v.deshpande.ctr@mail.mil

Received June 03, 2011; **Accepted** July 20, 2011; **Published** July 22, 2011

Citation: Deshpande SV, Jabbour RE, Snyder PA, Stanford M, Wick CH, et al. (2011) ABOid: A Software for Automated Identification and Phyloproteomics Classification of Tandem Mass Spectrometric Data. J Chromatograph Separat Techniq S5:001. doi:10.4172/2157-7064.S5-001

Experimental Section

Sample processing

Bacterial cells (*Escherichia coli* K-12 and *Bacillus cereus* ATCC 14579) were grown and processed before LC-MS analysis as described previously [6]. Bacteria were lysed by sonication and proteins extracted from ruptured cells were denatured with urea, reduced with dithiothreitol, alkylated with iodoacetamide, and digested by trypsin.

A mixture of peptides obtained by trypsin digestion of their cellular proteins were separated on a C18 column (100 μm i.d. × 100 mm) and analyses of electrosprayed peptide ions were carried out using an ion trap mass spectrometer (Thermo Scientific, San Jose, CA). Each MS data acquisition cycle consisted of a full-scan MS over the mass range *m/z* 400-1400, followed by data-dependent MS/MS scans over *m/z* 200-2000 on the five most intense precursor ions from the survey scan.

Data processing

MS/MS spectra were searched with TurboSEQUEST(Bioworks 3.1; Thermo Scientific, San Jose, CA) against a database constructed from FASTA formatted theoretical proteomes downloaded from an NCBI server, which are predicted from fully sequenced bacterial genomes. SEQUEST output files were processed with ABOid.

Protein database and database search engine

A protein database was constructed in a FASTA format using the annotated bacterial proteome sequences derived from fully sequenced chromosomes of 1125 bacteria, including their sequenced plasmids (as of 25th August, 2010). A PERL program was written to automatically download these sequences from the National Institutes of Health National Center for Biotechnology (NCBI) site (http://www.ncbi.nlm.nih.gov). Each database protein sequence was supplemented with information about a source organism and a genomic position of the respective ORF embedded into a header line. The database of bacterial proteomes was constructed by translating 8,323,020 putative protein-coding genes and consists of 2,521,160,789 amino acid sequences of potential tryptic peptides obtained by the *in silico* digestion of all proteins (assuming up to two missed cleavages).

The experimental MS/MS spectral data of bacterial peptides were searched by a SEQUEST [8] algorithm against this protein database.

Program operation

ABOid is a suite of algorithms developed in-house using Microsoft Visual Basic (VB).NET and PERL to analyze bacterial similarities and their identification from a virtual array of STO matrix of assignments. A data flow chart is shown in Figure 1.

dbCurator

The first module (dbCurator) written in Perl downloads the microorganism sequences and edits the header information of each protein. This new theoretical proteome of a microorganism is appended in the existing flat file that is saved as FASTA format. dbCurator also updates the in-house microorganism relational database (MyABOid) created using MySQL (http://www.mysql.com/) with the microorganism information like name, strain, sequencing Center, and other available data related to each bacterium. ABOid utilizes two databases: flat file FASTA format database used as the reference and MyABOid, which is a central repository database.

BacDigger

The second module (BacDigger) is designed to analyze the

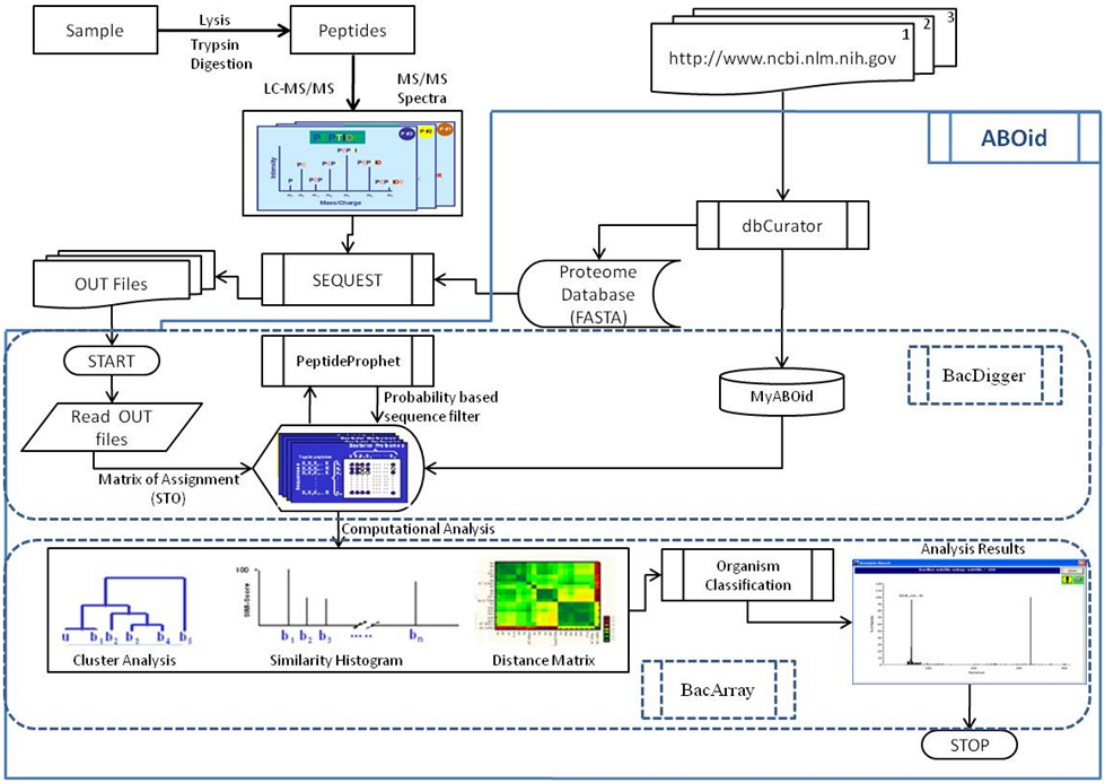


Figure 1: Flow chart of the ABOID application.

SEQUEST output files. The function of this module is to retrieve sequence matches to the in-house reference bacterial proteomes based on the identity of the peptides determined by SEQUEST and to obtain values for the matching parameters [8] like X_{corr}, ΔC_n, S_p, RS_p, ΔM, and number of amino acids in the peptide sequence identified. Assigning each peptide sequence a probability score determined by running PeptideProphet [9] algorithm to validate the SEQUEST peptide sequence assignments. Using the information contained in the reference of each output file, a STO binary matrix of assignments is created. This matrix of assignments, generated using raw results, is archived in a comma separated file format (CSV) for audit. Based on probability values, determined by a PeptideProphet algorithm that a sequence was correctly identified; a user specified threshold is applied to elements of the STO matrix of assignments to filter out low probability matches. This new'extracted' STO matrix of assignments is also saved in a CSV format. In addition, BacDigger removes duplicate sequences from the data set and retains only a unique set of peptides.

BacArray

The third module (BacArray) takes the STO matrices of assignments and displays numerical values in the form of color bitmaps ('virtual arrays') as shown in Figure 3. During this process, BacArray rearranges assignments by grouping database bacteria according to their taxonomic positions by using the relevant information stored in MyABOid. This allows for interactive browsing of sequence assignments, which can be further validated as they are dynamically linked with NCBI protein databases for blasting the sequences of interest.

The BacArray communicates with external statistical libraries to apply multivariate statistical techniques like principal component and cluster analysis to the STO matrices of assignments. In addition, it generate combined reports of such analyses, thus enabling the module to display the most probable taxonomic position of studied microorganisms and provides a user friendly display of results.

The application's graphical user interface (GUI) is developed in Microsoft Visual Basic.NET (http://msdn.microsoft.com/vbasic), data processing algorithms are developed using Microsoft C++ and PERL (http://www.activestate.com/Products/ActivePerl). Statistical analysis is performed using R 2.4.1 and Statistica (Statsoft, Inc., Tulsa, OK) MySQL Server is used to archive the data and results.

Results

MS/MS spectra of peptide ions generated during the electrospray ionization process of tryptic peptides derived from bacterial proteins were searched against the protein database with SEQUEST and the output files were processed by ABOid (Figure 1). Amino acid sequences of peptides were validated using probability scores generated by PeptideProphet and a set of 289 accepted peptide sequences (P> 0.98) were considered as elements of a row vector b_u that represents the peptide profile of unknown (*u*). Accordingly, sequence-to-organism (STO) assignments a_{1i} are elements of a row vector b_i that represents a peptide profile of a database bacterium assigned as number 1, and in general assignments a_{ij} are elements of row vectors b_i, where *i* represents the theoretical proteome of a i^{th} bacterium in the database (i=1, 2, 3, …, 203). All these row vectors form a matrix of assignment $A_{(m+1)x\,n}$ that is visualized in Figure 2A as a virtual array of *n=289* peptide sequences assigned to *m*=202 theoretical proteomes of database bacteria and an unknown microorganism (or their mixture). Conversely, each column vector represents a phylogenetic profile $\mathbf{s}_j$ of a peptide sequence. Thus for each MS/MS analysis, a binary matrix of assignments ***A*** is created with entries representing the presence or absence of a given sequence in each theoretical proteome of database microorganism. Similar b profiles indicate a correlated pattern of relatedness that in the majority of cases reflects the presence of identical sequences among orthologs

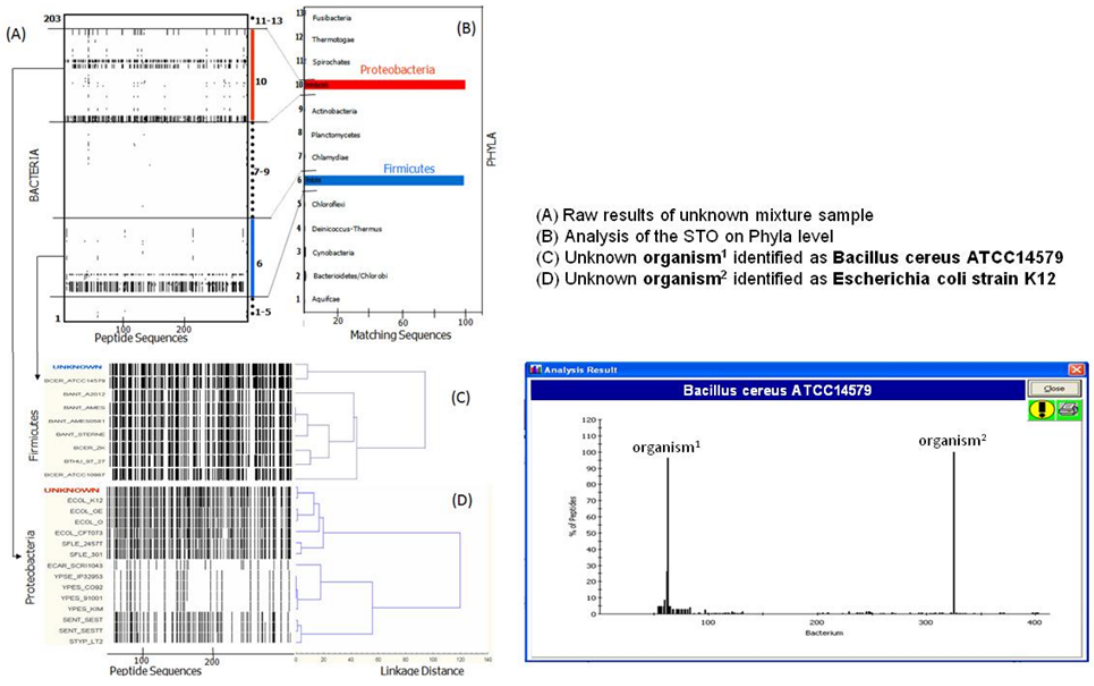


Figure 2: Data Analysis Pathway. (A) Virtual Array of 289 peptide sequences assigned to proteomes of 209 bacteria; (B) Histogram of the number of matching sequences assigned to 'super-proteomes' of 13 phyla obtained by merging database bacteria according to their taxonomic position; (C) cluster analysis of a sequence-to-organism assignment matrix for Firmicutes; (D) cluster analysis of a sequence-to-organism assignment matrix for Proteobacteria.

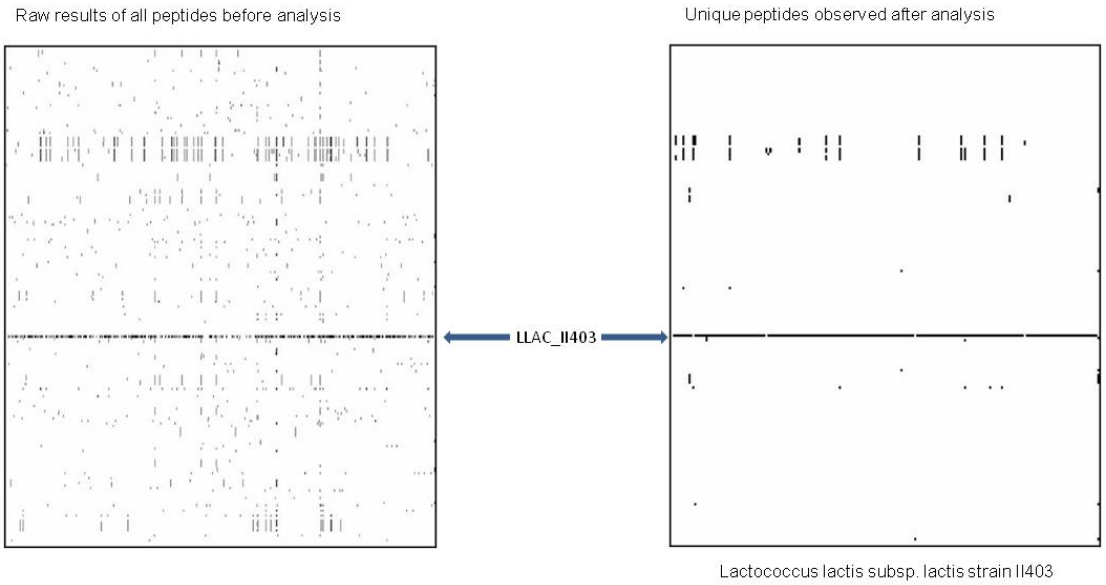
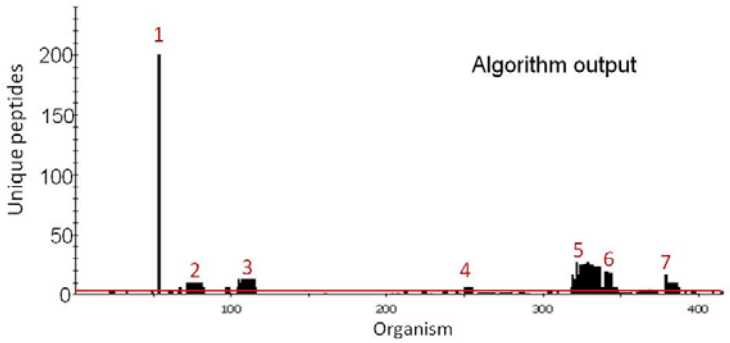


Figure 3: Virtual Array Bitmap.



Unique Peptides Analyses:
(1) *Bacillus cereus* ATCC14579, (2) *Staphylococcus Aureus* , (3) *Streptococcus pyogenes*,
(4) *Burkholderia Thailandensis*, (5) *Escherichia coli* strain K12, (6) *Salmonella enterica*
(7) *Pseudomonas aeruginosa* strain PA01

Figure 4: Results of blinded study showing results of seven organism mixture.

or other functional gene segments. The method predicts that peptide sequences $\mathbf{b}_u$, derived from cellular proteins of an unknown bacterium *u*, are most likely to be similar or even identical with a reference database bacterial strain b_i represented by a vector $\mathbf{b}_i$ with a highest number of non-zero elements.

The STO matrix of assignments was next analyzed by computing sequence assignments to merged proteomes that comprise bacteria grouped into 'super-proteomes' of 13 phyla represented in the database. The results shown in Figure 2B indicate that 98 unique sequences were assigned to the phylum *Proteobacteria* while 99 were assigned to *Firmicutes*. Thus, confirming the presence of a mixture of bacteria and allowing the classification of these organisms on the phylum level. The STO assignment sub matrices were further analyzed separately and the results obtained are shown in Figure 2C and Figure 2D as dendrograms representing results of cluster analyses that are accompanied by bar graphs that visualize peptide profiles of the test samples ('unknown') and the most similar database strains revealed by cluster analysis.

Cluster analysis provides a way to determine groups based on similarity within a data set, and perform Principal Component Analysis (PCA) to derive parsimony and reduce the dimensionality to measure the variation among the proteins identified from different strains of same species. Hierarchical clustering was performed using furthest neighbor (complete) linkage with squared Euclidean distances as the similarity metric and was used as an exploratory tool to examine relationships of a test microorganism with the database bacteria. Cluster analyses were performed automatically by linking STA Cluster library from Statistica to the BacArray module.

In Figure 2C one can observe two main clusters, the first one groups the 'unknown' with a database strain *Bacillus. cereus* ATCC 14579,

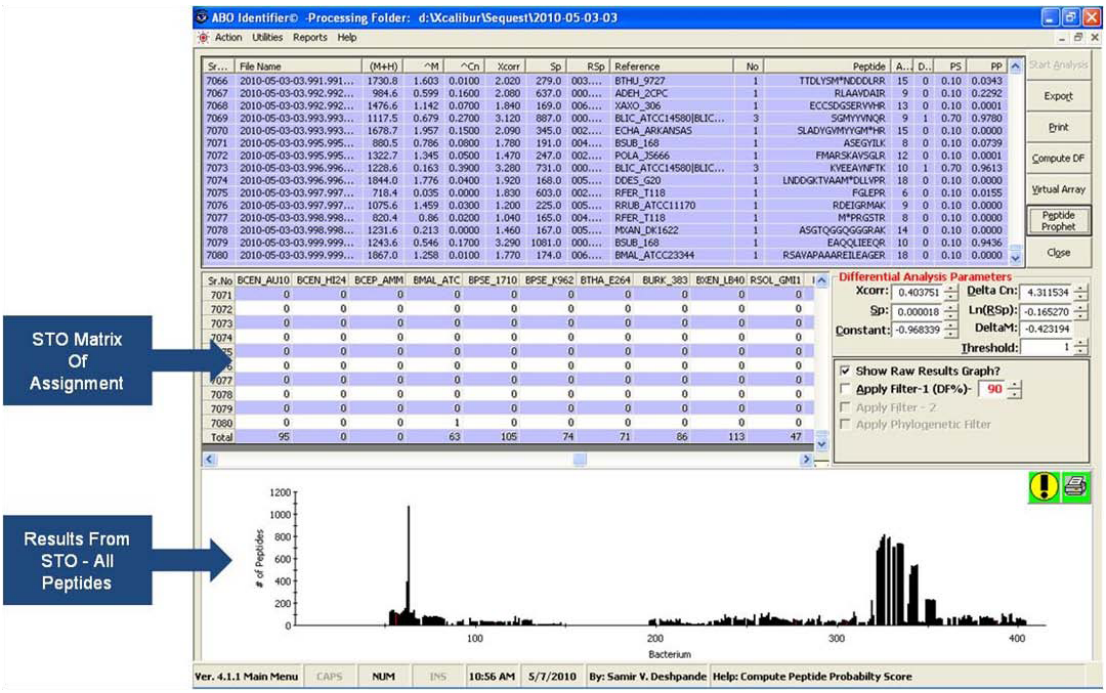


Figure 5: Screen shot of the application.

while the second cluster comprises diverse *Bacillus* strains classified as *Bacillus cereus*, *Bacillus anthracis* and *Bacillus thuringiensis*. In Figure 2D the test sample forms a subcluster with a database *E. coli* K-12 strain because they differ by only two peptide sequences. This subcluster is grouped with other *E. coli* and *Shigella flexneri* strains into a cluster that is substantially different in comparison to the next closest cluster that comprises *Salmonella* and *Yersinia* strains.

Figure 3 shows a virtual array plot of a single organism identified as *Lactococcus lactis* subsp. lactis strain II403 from the raw data before processing and unique peptides observed after data analysis. Figure 4 shows the algorithm generated results for seven organism mixture identified as (1) *Bacillus cereus* ATCC14579, (2) *Staphylococcus aureus*, (3) *Streptococcus pyogenes*, (4) *Burkholderia thailandensis*, (5) *Escherichia coli* strain K12, (6) *Salmonella enterica* and (7) *Pseudomonas aeruginosa* strain PA01. All the samples used in the analysis had been double blind bacterial samples. Further utility of ABOid on double blind bacterial samples can be found in an Applied Environmental Microbiology publication [10]. Figure 5 is the screenshot of the application

Conclusion

The results of applying ABOid for analysis of a bacterial sample composed of a mixture of *E. coli* K-12 and *B. cereus* ATCC 14579 strains demonstrate that mass spectrometry based proteomic approach, combined with ABOid for analysis SEQUEST output files, allows for automated assignment of analyzed organisms to taxonomic groups. Moreover, ABOid reveals genome-traced relatedness between bacteria that is suitable for fast and reliable classification and even identification of bacteria up to the strain level. Therefore, the application of this algorithm for analyses of proteomics data constitutes a new method that may function as a strong complement to DNA based approaches of comparing bacterial genomes.

In summary, ABOid is capable of revealing the identification of microbial mixture contents. The software application does not require prior knowledge of the sample and can be applied to pure cultures and mixtures. It allows for strain level identification based on comparative analysis of protein sequences and the un-sequenced bacterial strains not in our database are identified to their close-neighbor species. Statistical and visualization tools of the software allow its utilization by non-specialists end user.

Acknowledgement

This work is supported by the Defense Threat Reduction Agency (DTRA) # BRCALL07-N-2-0026.

References

1. Ruedi Aebersold, David R. Goodlett (2001) Mass Spectrometry in Proteomics. Chemical Reviews 101: 269-296.
2. Chen W, Laidig KE, Park Y, Park K, Yates Jr et al. (2001) Searching the Porphyromonas gingivalis genome with peptide fragmentation mass spectra. Royal Society of Chemistry: Cambridge, Royaume-Uni 126.
3. Warscheid B, Fenselau C(2003) Characterization of Bacillus Spore Species and Their Mixtures Using Postsource Decay with a Curved-Field Reflectron. Analytical Chemistry 7*:* 5618-5627.
4. Washburn MP,Wolters D,Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technolog. Nat Biotech 19: 242-247.
5. Wolters DA,Washburn MP,Yates JR (2001)An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. Analytical Chemistry 73: 5683-5690.
6. Dworzanski JP, Snyder AP, Chen R, Zhang H, Wishart D, et al. (2004) Identification of Bacteria Using Tandem Mass Spectrometry Combined with a Proteome Database and Statistical Scoring. Analytical Chemistry 76: 2355-2366.
7. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422: 198-207.

8. Jimmy K Eng, Ashley L McCormack, John R Yates III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry 5: 976-989.

9. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. Analytical Chemistry 74: 5383-5392.

10. Jabbour RE, Deshpande SV, Wade MM, Stanford MF, Wick CH et al. (2010) Double-Blind Characterization of Non-Genome-Sequenced Bacteria by Mass Spectrometry-Based Proteomics. Applied and Environmental Microbiology 76: 3637-3644.