Computational and Experimental Simulations for Mass Spectrometry based identification of Biological Weapons.

Jane Razumovskaya[2,3], W. Judson Hervey[2], Nathan C. VerBerkmoes[1,2], Miriam Land[3], Loren Hauser[2,3], Frank W. Larimer[2,3], Ying Xu [2,3], Gary J. Van Berkel[1,2], Douglas E. Goeringer[1]

Chemical Sciences Division, Oak Ridge National Laboratory[1]; Graduate School of Genome Science and Technology (GST), University of Tennessee/Oak Ridge National Laboratory[2]; Life Science Division, Oak Ridge National Laboratory[3]

Overview:

Purpose: To assess the possibility of employing mass spectrometry in the field of bioweapon detection using computational and experimental techniques.

➢ Using a set of computer simulations, approximate what mass spectrometric technique and approach would be best for the task: Top Down or Bottom Up approach (Ver Berkmoes *et. al* 2002) .

➢ Using the experimental Bottom Up technique, to perform a sets of tests for detection and identification of an organism within a complex mixture.

Introduction:

➢ Current political events and acts of terrorism have elevated the demand for suitable instrumentation to detect and identify potentially threatening agents, such as bacteria, viruses, toxins, and chemical agents.

➢ This heightened demand for a robust instrument with the capability to simultaneously identify all possible threats within a narrow timeframe exceeds current technology.

➢ In order to develop novel instrumentation with such capabilities, it is necessary to probe the threshold of current instrumentation using both experimental and computation simulations.

➢ Computational simulations were run comparing a subset of abundant genes from the target organism against the background complex mixture database at both the tryptic peptide level (Bottom Up) and the whole protein level (Top Down) to determine the optimal mass spectrometric approach for the bioweapon detection as well as what level of mass accuracy needed to make confident identifications.

➢ Experimental simulations were run by combining the whole cell lysate protein fraction of five organisms under two scenarios: a Battlefield scenario and a Homeland Security scenario.

Computational Simulations:

➢ A "complex mixture" of microbial organisms was generated from annotated genomes, containing ~80,000 unique entries from 12 organisms (Figure3).

➢An approach for detecting a target organism within a complex mixture of organisms was designed (Figure2):

- ➢ An experimentally detected set of proteins from the target organism ("representative set") is used as the database for the identification.
- ➢An experimental library of fragmentation patterns is composed from each of the "representative set" proteins.
- ➢The "representative set" database is compared to the complex mixture, in order to identify the presence or absence of target organism in the mixture.

➢*E. coli* was selected as the "target organism" which is to be detected within the background "complex mixture".

➢A "representative set" database was created for the target organism:

- ➢ An *E. coli* set of proteins, homologous to the proteins experimentally detected in both *S. oneidensis* and *R. palustris (and therefore concidered housekeeping)* were accepted as the "representative set" for the target organism . This database contains 376 number of unique entries.

➢In order determine the most advantageous approach for organism detection in a background complex mixture, theoretical simulations of both the Top Down and Bottom Up techniques (Figure1) were applied to detect the target organism (Figures 4, 6,7 ,8).

➢A normalized scoring scheme, based on SEQUEST (Eng *et al.* 1994) X correlation score was used for comparison of fragmentation patterns (Formulas 1, 2).

➢A normalized organism scoring scheme was developed, in order to assess the reliability of organism detection in the background complex mixture based on the identified proteins from the "representative set" database (Formulas 3, 4).

➢The simulations involve comparison between the mixture of 80,000 proteins and the database of "representative set" from the target organism, in an attempt to show that using the described method with a selected set of cutoffs, the target organism will not be falsely detected within the mixture of 12 other organisms (Figures 8, 9).

Figure 1.

Organism Detection Approach (Top Down and Bottom Up) :

Figure 2.

Organisms used in the simulations:

Figure 3.

➢The Computational Biology Group (Larimer *et al.*) at ORNL annotated the genomic sequences of 5 organisms sequenced by the JGI using a pipeline consisting of **three common utilities: Critica, Generation, and GLIMMER and numerous database queries**.

➢ These 12 organisms, containing ~80,000 proteins, are frequently found within the environment and, therefore, were considered a reasonable background "complex mixture" for the study.

➢Within the computational simulation, *E. coli* was selected to be the target organism, with 376 proteins as the "representative set."

Comparison of Top Down and Bottom Up approaches based on intact masses analysis:

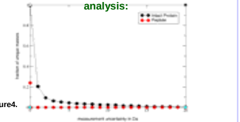


Figure4.

Fraction of unique masses of proteins and peptides from "representative set" as compared to the background "complex mixture".

Protein masses of Representative Set and Background Set:

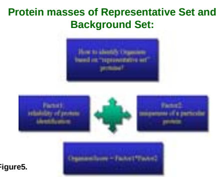


Figure5.

Protein/peptide Identification Based on MS/MS:

The protein/peptide scoring scheme is based on the current SEQUEST X correlation score:

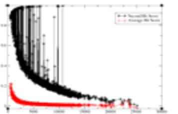
$$F(\tau) = \sum_{i=1}^{n} x[i] * y[i+\tau]$$ where τ is the displacement (**Formula1**)

and Score is defined as:

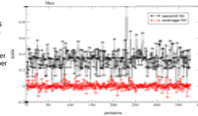
$$Score(f1, f2) = (F(0) - \sum_{i=-10}^{10} F(i))/Score(f1, f1)$$ (**Formula2**)

Comparison of Top Down and Bottom Up approaches based intact mass and MS/MS analysis:

All the corresponding proteins/peptides (by parent mass) from the complex background are scored against the database of proteins/peptides from the "representative set". The comparison between fragmentation patterns of identical sequences produce Score=1.



Identifying peptides with MS/MS. Each peptide has complete MS/MS fragmentation.



Identifying 376 representative proteins with MS/MS. The allowed number of fragments per protein is 10.

Figure6.

The second hit score shows the top score of proteins in the "complex mixture" background vs the "representative set". The number of the proteins that are unique in the "representative set" far exceeds the number of unique peptides in the "complex mixture" background if the accepted score cutoff is greater than 0.4. The average score, shows the average of scores of all the matched proteins/peptides in the "complex mixture" background set.

Top Down Protein identification with varying amounts of MS/MS fragmentation (1-150 MS/MS ions):

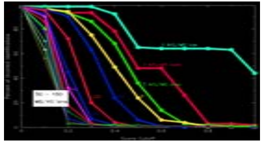


Figure 7.

Number of MS/MS ions is crucial for correct protein identification. These plots present the dependency among the number of MS/MS ions, the identification score and the percent of incorrect identifications made. At 20-30 MS/MS fragments per protein, the identification becomes reliable as incorrect identifications are less than at 10% at score 0.4 (Score range [0:1]).

Comparison between Top Down and Bottom Up approaches for Bioweapon Identification:

Percent of Incorrect Identifications vs Protein Score:

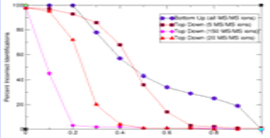


Figure 8.

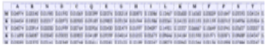
Performances of Bottom Up peptide assignment and Top Down protein assignment are shown. Bottom Up approach appears with complete MS/MS fragmentation. Three Top Down experiments are shown, with 5, 150 and 20 MS/MS fragments per protein. With increasing "Score cutoff", the "Percent of Incorrect Identifications" decreases. It is shown that at Score = 0.6, Bottom Up shows 38% of incorrect identifications, while Top Down with 5 MS/MS fragments, shows 18% of incorrect identifications, Top Down with 150 shows 2%, and Top Down with 20, 4%. The figure shows that 20 fragments per protein is enough for reliable identification. The results illustrate the advantage of Top Down method over the Bottom Up for this type of identification, as peptides are more likely to yield a false identification.

Protein Uniqueness considerations:

➢It is impossible to assess whether a protein is unique from a simple database search, since there is no specific background database for the problem. Thus the protein uniqueness must be approximated based on the current state of genetic information.

➢Solution: Family based profiles.

➢Using multiple sequence alignment between members of the same family, it is possible to compute the likelihood of a protein to be completely conserved between different organisms.



Family based profile matrix example. The column describes the amino acid sequence of a protein, while the row corresponds to the 20 possible amino acids. An element of matrix i,j, describes the frequency of amino acid i in position j of the sequence.

u_i is the normalized measure of uniqueness of the protein in nature:

$$u_i = 1 - \prod_{k=1}^{m} FM(a_k, k)$$ (**Formula 3**)

where, **FM** is frequency matrix, **m** is the number of amino acids in the protein sequence and **(a_k,k)** is the frequency of occurrence of **k**th amino acid in **k**th position within the protein family.

Reliability of Organism Identification in an unknown complex mixture:

$$OrganismScore = \sum_{i=1}^{N} u_i P_i \text{ and } \sum u_i = 1$$ (**Formula 4**)

where **N** is the cardinality of "representative set, u_i is the normalized measure of uniqueness of the protein in nature, P_i is a binary indicator whether protein is present or not, (in the following studies will be substituted by protein reliability score). u_i is normalized, so that **Organism Score** always equals to one (scores of different organisms can be compared)

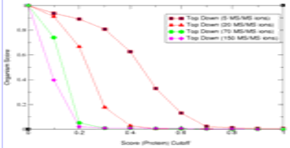


Figure 9.

As the protein Score cutoff is lowered (as it will be necessary for real experimental data), the number of false positives increases, thus increasing the likelihood of falsely identifying the target organism in the complex background. When cutoff is low and all the proteins are accepted as identified, the target organism is falsely detected with the score of 1, while when it is high, there are no false identifications and the OrganismScore is close to 0. The OrganismScore scheme adds only a few percent improvement to the detection (Figure8), however in this study, there is only one protein that is shared between the "representative set" and the complex mixture background (it's weight was calculated to be the least among all the other proteins in the "representative set").

Conclusions:

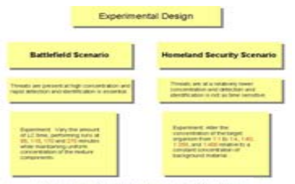
Computational analysis of the bioweapon detection in the complex mixture showed that:

➢Top Down would be more desirable approach to bioweapon identification as opposed to the bottom up approach. The reduction in complexity of the database analysis contributes to the top down method as the more appealing method.

➢The mass accuracy of the experiment needed for such identification are modest. provided the accuracy of the mass spectrometer ± 10Da, MS/MS fragmentation efficiency starting from 10-20 fragments per protein).

➢It is likely that using the described schemes it is not only possible to identify a target organism within a complex mixture background, but to avoid false positive identification of target organisms.

Experimental Simulations:



Components of Mixture for Experimental Simulations

	Organism	Database ID Code	Function within Mixture	[Lysate]
	A. thaliana	ath	Background	1.8 mg/mL
	S. cerevisiae	sce	Background	3.0 mg/mL
	R. palustris	rpal	Background	2.5 mg/mL
	S. oneidensis	spul	Background	2.4 mg/mL
	E. coli	eco	Target	Variable

For both the battlefield and homeland security scenarios, the complex mixture of organisms was searched against the database file containing the components of the above mixture, and eight other organisms. The intention was to simulate analysis of a mixture whose concentration and components were unknown.

Methods:

Cell Growth and Production of Protein Fractions:

➢Organisms in the complex mixture were grown under their corresponding optimal growth conditions and harvested for total protein content.

➢Cells were harvested, washed twice with buffer, and disrupted with sonication. *A. thaliana* samples were disrupted with liquid nitrogen. One crude protein fractions was created by centrifugation (10,000g for 20 minutes).

➢Protein fractions were mixed, denatured, reduced and digested with sequencing-grade trypsin (Promega).

LC-MS/MS Analysis and Database Searching:

➢One-dimensional LC-MS/MS experiments were performed with an Ultimate HPLC (LC Packings, a division of Dionex, San Francisco, CA) coupled to an LCQ-DECA or LCQ-DECA XP ion trap mass spectrometer (Thermo Finnigan, San Jose, CA) equipped with an electrospray source. Injections were made with a Famos (LC Packings) autosampler onto a 50ul loop. Flow rate was ~4ul/min with a 240min gradient for each run.

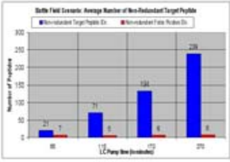
➢A VYDAC 218MS5.325 (Grace-Vydac, Hesperia, CA) C18 column (300µm id x 15cm, 300Å with 5µm particles) or a VYDAC 238EV5.325 monomeric C18 (300µm id x 15cm, 300Å with 5µm particles) was directly connected to the Finnigan electrospray source with 100µm id fused silica.

➢For data acquisition, the LCQ was operated in the data dependent mode with dynamic exclusion enabled, where the top four peaks in every full MS scan were subjected to MS/MS analysis.

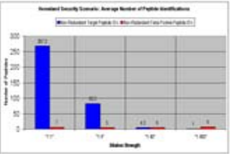
➢To increase dynamic range separate injections were made with a total of 4 or 5 separate m/z ranges scanned.

➢The resultant MS/MS spectra from each fraction were searched with SEQUEST against the database previously described. The raw output files were filtered and sorted with DTASelect and were compared with Contrast (Tabb, D.L.; et al, Journal of Proteome Research, 2003). Perl scripts were written to parse the Contrast output and select the presence or absence of the target organism and false positive peptides.

Results:



➢**Battlefield Scenario:** At all LC time intervals, the presence of the target organism, *E. coli*, was detected.



➢ **Homeland Security Scenario:** At 1:1 and 1:4 dilutions, the presence of the target organism, *E. coli*, was detected. At the 1:40 and 1:400 dilutions, the target cannot conclusively be identified. An additional experiment, a 1:250 dilution, was performed with increased instrument time, but only yielded 7 non-redundant peptides of the target organism.

Conclusions/Future Work:

➢Alternative methods of data analysis are currently underway in order to determine the optimal statistical method of identification.

➢Both MASCOT (Matrix Science) and Protein and Peptide Prophet (ISB) are candidates to find reliable peptides for confirmation of presence or absence of the target organism.

➢Explore the threshold of detection within the 1:4 and 1:40 dilutions to characterize the limit at the 270 minute analysis.

➢Discover the threshold of detection independent of time.

Acknowledgements:

- Tema Fridman, Ed Uberbacher, Inna Vokler for invaluable discussions.
- Grace Vydac for HPLC columns
- Barry Bruce, David McWilliams, M. Brad Strader, Patricia Lankford, and Dale Pelletier for assistance with organism growth and harvest.

References:

1. VerBerkmoes, N.C; Bundy, J.L.; Hauser, L.; Asano, K.G.; Razumovskaya, J.; Larimer, F.; Hettich, R. L.; Stephenson, J.L. Jr. *JPR.* **2002,** *1*,(3) 239-252
2. Tabb, D. L.; McDonald, W. H.; Yates, J.R. III. *JPR.* **2002,** *1*, 21-26
3. Eng, J. K., McCormack, A. L.; Yates, J. R. III. *JASMS.* **1994**, 5, 976-989.

Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.