

## Regis University ePublications at Regis University

---

All Regis University Theses

---

Fall 2011

# Exploring Information Technologies to Support Shotgun Proteomics

Alexander M. Mendoza

*Regis University*

Follow this and additional works at: <https://epublications.regis.edu/theses>



Part of the [Computer Sciences Commons](#)

---

### Recommended Citation

Mendoza, Alexander M., "Exploring Information Technologies to Support Shotgun Proteomics" (2011). *All Regis University Theses*. 627.

<https://epublications.regis.edu/theses/627>

This Thesis - Open Access is brought to you for free and open access by ePublications at Regis University. It has been accepted for inclusion in All Regis University Theses by an authorized administrator of ePublications at Regis University. For more information, please contact [epublications@regis.edu](mailto:epublications@regis.edu).

**Regis University**  
College for Professional Studies Graduate Programs  
**Final Project/Thesis**

# **Disclaimer**

Use of the materials available in the Regis University Thesis Collection ("Collection") is limited and restricted to those users who agree to comply with the following terms of use. Regis University reserves the right to deny access to the Collection to any person who violates these terms of use or who seeks to or does alter, avoid or supersede the functional conditions, restrictions and limitations of the Collection.

The site may be used only for lawful purposes. The user is solely responsible for knowing and adhering to any and all applicable laws, rules, and regulations relating or pertaining to use of the Collection.

All content in this Collection is owned by and subject to the exclusive control of Regis University and the authors of the materials. It is available only for research purposes and may not be used in violation of copyright laws or for unlawful purposes. The materials may not be downloaded in whole or in part without permission of the copyright holder or as otherwise authorized in the "fair use" standards of the U.S. copyright laws and regulations.

EXPLORING INFORMATION TECHNOLOGIES TO SUPPORT SHOTGUN PROTEOMICS

A THESIS

SUBMITTED ON THE FIRST OF NOVEMBER, 2011

TO THE DEPARTMENT OF DATABASE TECHNOLOGY

OF THE SCHOOL OF COMPUTER & INFORMATION SCIENCES

AT REGIS UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF MASTER OF SCIENCE IN

DATABASE TECHNOLOGIES

BY



---

Alexander M. Mendoza

APPROVALS



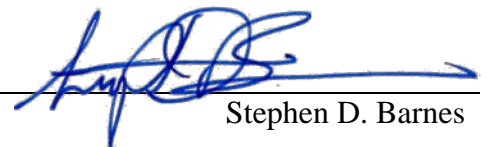
---

Darl Kuhn, Thesis Advisor



---

Robert T. Mason



---

Stephen D. Barnes

## SHOTGUN PROTEOMICS

### **Abstract**

Shotgun proteomics refers to the direct analysis of complex protein mixtures to create a profile of the proteins present in the cell. These profiles can be used to study the underlying biological basis for cancer development. Closely studying the profiles as the cancer proliferates reveals the molecular interactions in the cell. They provide clues to researchers on potential drug targets to treat the disease. A little more than a decade old, shotgun proteomics is a relatively new form of discovery, one that is data intensive and requires complex data analysis. Early studies indicated a gap between the ability to analyze biological samples with a mass spectrometer and the information systems available to process and analyze this data. This thesis reflects on an automated proteomic information system at the University of Colorado Central Analytical Facility.

Investigators there are using cutting edge proteomic techniques to analyze melanoma cell lines responsible for skin cancer in patients. The paper will provide insight on key design processes in the development of an Oracle relational database and automation system to support high-throughput shotgun proteomics in the facility. It will also discuss significant contributions, technologies, software, a data standard, and leaders in the field developing solutions and products in proteomics.

*Keywords:* high-throughput shotgun proteomics, Oracle, relational database

## SHOTGUN PROTEOMICS

### **Acknowledgements**

I would like to acknowledge my mentor during my time at the Colorado Central Analytical Facility in Boulder. The late Dr. Katheryn Resing was an extraordinary innovator in shotgun proteomics and contributed to my experience with data systems tremendously. Not too many graduate students get to develop high-throughput proteomic systems with multiple terabyte Oracle databases. A dream position for me, I consider myself fortunate for the time we had together working in the lab and I will always remember Dr. Resing's passion for science and discovery.

In addition to Dr. Resing, I'd also like to thank all of the lab members that helped make this project a success. Dr. Natalie Ahn, Dr. William Old, Dr. Karen Meyer-Arendt, Lauren Aveline-Wolf, and Kevin Pierce are greatly appreciated for their support and dedication to the advancement of proteomics.

## SHOTGUN PROTEOMICS

**Table of Content**

Abstract .....	2
Acknowledgements .....	3
Table of Content .....	4
List of Figures .....	6
Chapter One: Background of Proteomics .....	7
Chapter Two: Introduction to Shotgun Proteomics .....	9
Sample Preparation in Shotgun Proteomics.....	9
Shotgun Proteomics Data Propagation .....	11
MS/MS Data Explained .....	12
Data Processing After the Instrument .....	12
Using Protein Profiles to Understand Cancer .....	13
Sequence Database Repositories.....	14
Sequence Database Format .....	15
Sequence Data Standards.....	16
Chapter Three: Challenges in Human Proteomics Studies .....	16
Chapter Four: Data Processing In Shotgun Proteomics.....	18
Data at the Instrument .....	19
Protein Inference Variables.....	21
Entity Relationship Diagrams (ERD) .....	21
Using Design to Create Structure .....	22
From Flat File to Relational Database .....	24
Automation and SmartJobs .....	25
Chapter Five: Database Design for Proteomics .....	27
A Peptide-centric Homogeneous Database.....	27
TurboSequest Tables Structure .....	29
The RANKED_HITS Relationship.....	30
Chapter Six: Proteomics Laboratory Infrastructure .....	31
Vulnerability at the Acquisition Workstations .....	31
Hardware Resource Utilization and Cost.....	36
Chapter Seven: Proteomics Software Development .....	35
Proteomic Data Management and Analysis Software .....	36
Proteomic Data Standards.....	37
Open Data Standards for Proteomics .....	40
Software Explosion .....	41
Proteomic Software Market Value.....	41
Third Generation Products .....	42
Mascot Integra .....	43
Trans-Proteomic Pipeline (TPP) .....	44
TPP Dataflow Explained.....	45
Discussion of Mascot Integra vs. Tans-Proteomic Pipeline (TPP) Data Storage .....	46
Native XML Databases (NXD) .....	47

SHOTGUN PROTEOMICS

Chapter Eight: Proteomic Information Future ..... 48

    Microarray Technology ..... 48

    Mass Spectrometry Imaging ..... 49

    Closing Thoughts ..... 50

References ..... 51

## SHOTGUN PROTEOMICS

**List of Figures**

Figure 1. Visualization of a proteome.....	8
Figure 2. A complete breakdown of the biological sample processing in the laboratory. ....	10
Figure 3. Data file propagation due to multidimensional chromatography and gas phase fractionation. ....	11
Figure 4. MS/MS spectra and peptide sequence mapping with a protein identification. .	13
Figure 5. Organizations maintaining publicly available protein sequence database for proteomic studies. ....	15
Figure 6. An example of a protein sequence in FASTA format. ....	16
Figure 7. Industry leaders in protein search engines.....	21
Figure 8. A complete view on the dataflow in TurboSequest process in an Entity Relationship Diagram with GUI forms and metadata.....	22
Figure 9. Perl code to parse an .OUT file to relational database tables.....	24
Figure 10. A SmartJob or batch file DOS (Windows), created by a Perl script to process data. ....	26
Figure 11. Database relationships schema of the TurboSequest Processing. ....	29
Figure 12. The Central Analytical Faculty's VPN diagram circa October 2002.....	32
Figure 13. Presentation slide of proteomic infrastructure circa 2000. ....	35
Figure 14. Leading proteomic applications. ....	37
Figure 15. An example of MIAPE-compliant data management dataflow .....	39
Figure 16. The mzXML msInstrument element). ....	41
Figure 17. <i>Mascot Integra</i> architecture diagram.....	43
Figure 18. Schematic overview of the TPP workflow.....	45
Figure 19. Mass spectrometric images of a mouse brain section. ....	49



## SHOTGUN PROTEOMICS

**Chapter One: Background of Proteomics**

The word *proteome* is a term that describes the complex composition of all the proteins in a cell. These proteins created by genes through the processes of transcription and translation are essential for life as they carry out the cell's routine functions. A cell's proteome is extremely dynamic. Proteins can change structures or modify with particular stresses or in response to signals in the cell. For example, nearly all-human cancers involve abnormal changes in the structure of key regulatory proteins induced by the addition of a phosphate group. These changes or regulatory mechanisms can be characterized using proteomics. The pressing need to characterize cellular interactions, the proteins involved in these processes, and the underlying mechanisms have led to extensive studies of the proteome. Advancements in high-resolution mass spectrometry drive *proteomics*, a field focused on the presence, abundance, structure, function and localization of a cell's proteins.

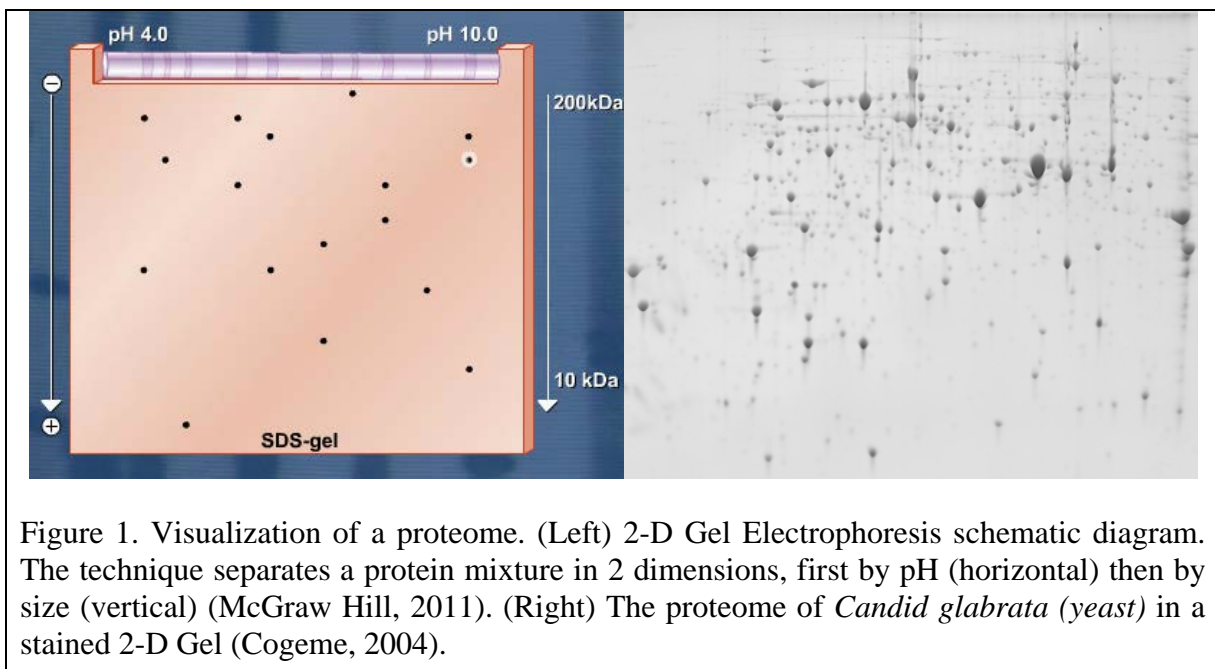
The term proteome first appeared in the nineties as biochemists developed new technologies for investigating proteins at this scale (N.L. Anderson, N.G. Anderson, 1998). Akin to the term *genomics*, Marc Wilkins at the University of New South Wales, Sydney, Australia in 1994 coined the word, a combination of the words *protein* and *genome* (University of New South Wales. 2010). Researchers conducting proteomic studies have a wide variety of interests including protein composition, location, quantity, structure and function of proteins (Blackstock & Weir, 1999).

Proteomics has proven to be an effective tool of many areas of traditional protein chemistry, but has also helped pave the way forward in numerous other areas. Biomedical

## SHOTGUN PROTEOMICS

research has particularly benefited from the technology, leveraging it to better understand therapeutic targets in biological systems. Proteomics has shown promising initiatives in identifying novel biomarkers of various diseases (Ahram & Petricoin, 2008). Proteins constitute the main bulk of therapeutic targets in the cell, accounting for more than 98% of drug targets (Drews, 2000).

A proteome can be visualized by using a 2-dimensional polyacrylamide gel that separates the proteins in the mixture by their isoelectric point and molecular size. Early efforts in complete proteome analysis focused on gel-based separation of all soluble proteins expressed in a cell. The result is a gel with pools of proteins that can be stained and visualized. Figure 1 illustrates the physical properties 2-D gel uses and an example of the yeast proteome on a stained polyacrylamide gel.



## SHOTGUN PROTEOMICS

### **Chapter Two: Introduction to Shotgun Proteomics**

Understanding shotgun proteomic sample processing is critical for developing an information system to support it. Samples are prepared in a variety of ways, but normally the steps include extracting a complex mixture of proteins from cells grown in culture, enzymatic digestion of this mixture into peptides, followed by multidimensional chromatography and tandem mass spectrometer analysis.

The initial biological samples for these studies are collected from patients receiving treatment for melanoma. The clinics and hospitals performing the excisional biopsies use the tissue to diagnose the medical condition and qualifying samples participate in the proteomic studies. In the laboratory these cells are grown in cell culture and eventually prepared for analysis with a mass spectrometer.

#### **Sample Preparation in Shotgun Proteomics**

A sample prepared for analysis undergoes a series of events, which are relevant for the database and automation system. The steps include multidimensional chromatography separation of the sample, prior to the instrument, which simplifies the complex cellular mixture based on the physical properties of the mixture's content (Kajdan, Cortes, Kuppanan, & Young, 2008). This simplification is critical for acquiring the maximum amount of usable spectrum from the instrument.

First, the proteins in the mixture are separated by their mass via size exclusion gel filtration. This process separates the initial sample into as many as forty fractions. These protein fractions are still too complex for efficient identification of constituent proteins.

## SHOTGUN PROTEOMICS

The proteins in each fraction undergo a process called trypsinization. They are denatured and digested with the protease trypsin into their constituent peptides. Next these peptides are then subjected to strong cation exchange chromatography (SCX), which further separates the peptides on the property of *basicity*, producing on average 16 fractions for each SCX fraction. These steps simplify the mixture into fractions with similar physical properties and improve the ability of the mass spectrometer to analyze them.

The instrument analyzes sample fractions eluted on-line or directly from the chromatographer over a period of time. The sample is ionized into a *gas-phase* for yet one more separation, and then analyzed by the instrument. This generates MS/MS fragmentation spectra data. Figure 2 describes the steps involved in processing a sample (from left to right). This series of events generates all the proteins in the study and initiates the data lifecycle. Details of the laboratory preparation and experimental steps are an essential part of the database and can link any findings to the biological procedures in the lab. The following sections will discuss the initial protein sample processing in shotgun proteomics and how data is created in detail.

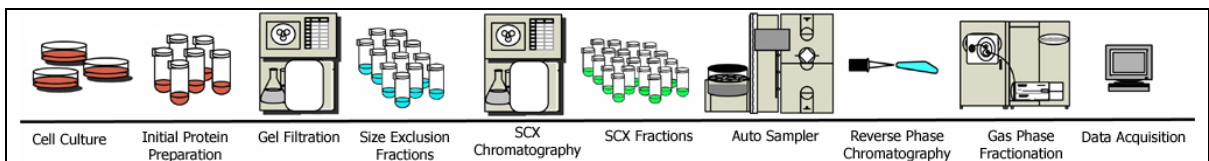


Figure 2. A complete breakdown of the biological sample processing in the laboratory.

1. Cell cultures grown from tissue and prepared for shotgun proteomic processing.
2. Samples are separated by mass with size exclusion gel filtration. One sample yields 13 size exclusion fractions of itself.
3. Trypsinization of the size exclusion fractions are then subjected to strong cation exchange chromatography (SCX) and yields 16 SCX fractions each.
4. An auto-sampler will submit the SCX fractions to the Mass Spectrometer for the

## SHOTGUN PROTEOMICS

final separation that happens while peptide is a gas.

### Shotgun Proteomics Data Propagation

Each fraction is analyzed over seven overlapping mass ranges in the instrument. Called gas phase fractionation, this improves the ability of the mass spectrometer to analyze more of the peptides in the sample. The lower intensity ions, typically more difficult to detect, are sequenced more effectively with this method. The instrument will produce a file for each mass range analyzed. The grand total, now over 1450 binary data files, are created from one initial protein sample in shotgun proteomics (Resing, Meyer-Arendt, Mendoza, et al., 2004). Figure 3 is a schematic breakdown of how the initial protein sample creates over 1450 data files.

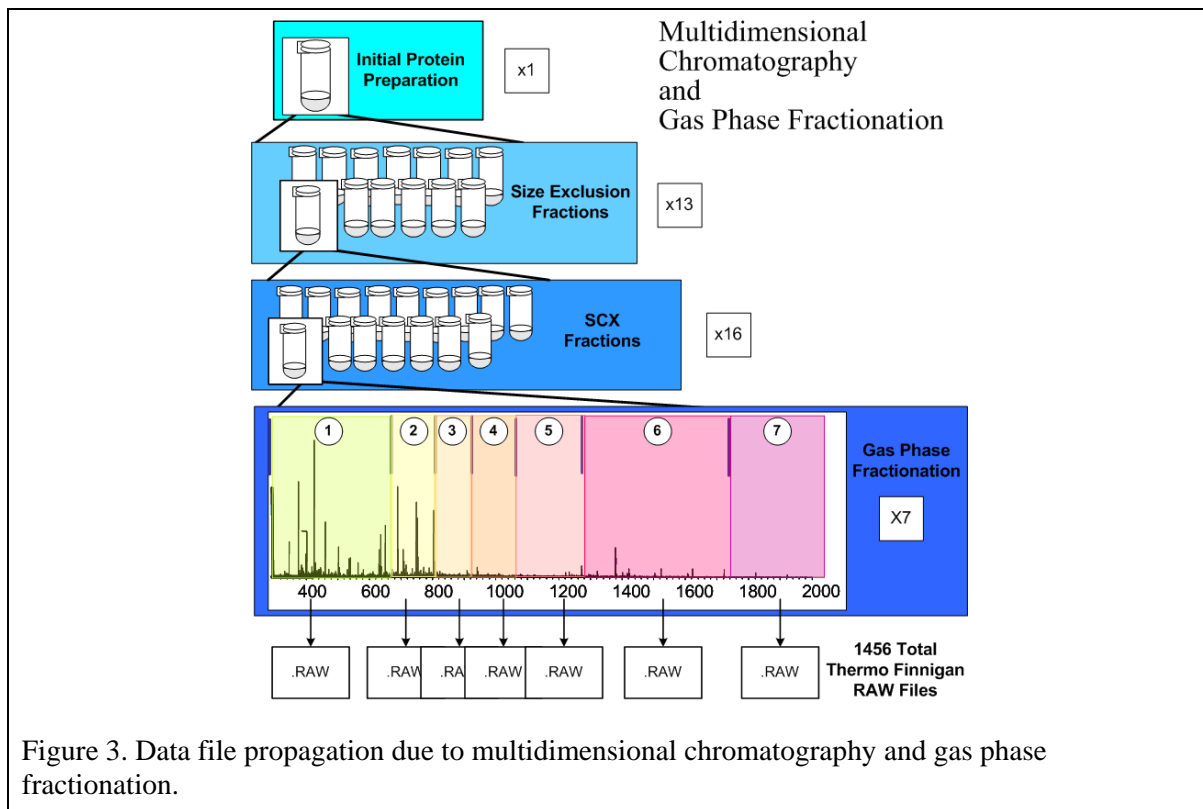


Figure 3. Data file propagation due to multidimensional chromatography and gas phase fractionation.

## SHOTGUN PROTEOMICS

**MS/MS Data Explained**

The term MS/MS is specific to tandem mass spectrometers and refers to the use of two mass analyzers in tandem to measure both a peptide's unfragmented mass (mass-to-charge ratio) and those peptides fragmented in the gas-phase. The resulting MS/MS serves as a unique fingerprint of the peptide, and can be used to identify the peptide in an unknown complex mixture. A highly sensitive technique, these analyzers measure the different physical characteristics of the sample material, in this case peptide fragment ions generate MS/MS spectrum data, plotted as it relates to the time (seconds) the peptide fragments appeared in the instrument's analyzer.

**Data Processing After the Instrument**

This data file or spectrum is the instrument's raw output and must be interpreted carefully to gather information about the samples protein content. Proteomic mass spectrum can be correlated with known peptide repositories and information about the sample's proteins can be gathered. Using sophisticated search algorithms, the raw datasets are interrogated against a catalog of known protein sequences to determine the proteins present in the sample. This search strategy maps protein sequences to MS/MS spectra. Figure 4 is a diagram of a single peptide fragment undergoing a process called protein inference. This is a process where software will assign the most likely protein assignment for the peptides observed in the data. The *protein-centric* approach matches the peptides directly to protein database entries and reports peptides within the context of proteins (Meyer-Arendt, 2011). In a simple instance, the process is basically estimating

## SHOTGUN PROTEOMICS

the protein content in sample fraction by constructing a probable list of amino acid and peptide sequences based on ion intensity detected by the instrument.

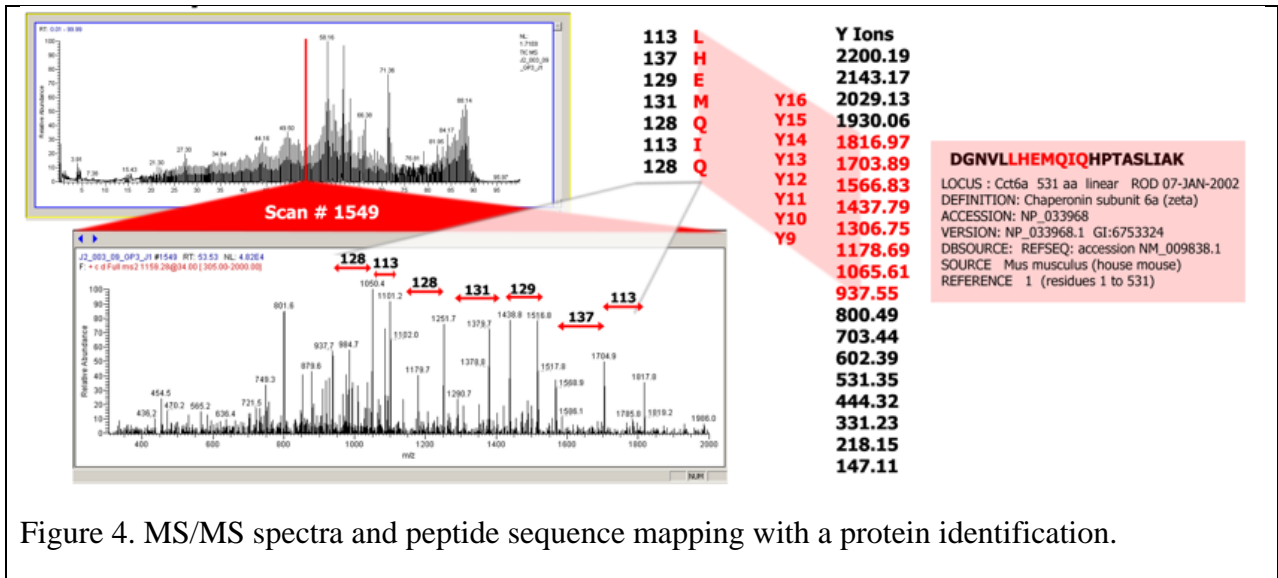


Figure 4. MS/MS spectra and peptide sequence mapping with a protein identification.

An alternative strategy for this process uses direct spectrum-to-spectrum matching against a reference library of previously observed MS/MS (Yen, 2008). This approach is limited by the small sizes of the available peptide MS/MS libraries and the inability to evaluate the rate of false assignments (Yen, 2008), but can be used to enhance analytics by estimating false positive rates.

### Using Protein Profiles to Understand Cancer





Comparing diseased profiles to normal or less affected cell profiles and looking for differences in protein levels as the cancer grows can reveal clues about how the cancer progresses into more pervasive and dangerous forms. Researchers can characterize a stage of progression by the cell's proteome. Using these profiles, the investigators are developing an unparalleled understanding of how cancer progresses at the cellular level

## SHOTGUN PROTEOMICS

and mapping possible drug treatment targets to stop or inhibit it. The technology has proved critical in identifying new targets for therapeutic treatments and markers for early cancer detection (Resing, Meyer-Arendt, Mendoza, et al., 2004).





### Sequence Database Repositories

Proteomic data processing workflows using software provided by the instrument's manufacturer is based on traditional utilizations of that instrument. Users can correlate uninterrupted tandem mass spectra of peptides with amino acid sequences from known protein and nucleotide databases. These peptide sequence databases are available through various academic institutions, international scientific communities and government agencies. Figure 5 details a list of contributors maintaining popular sequence database repositories. Users can download peptide databases that are species specific or revisions of curated datasets maintained by bioinformatics institutes.

	<p><b>The Gene Ontology Project</b> A major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases (GO, 2011).</p>
	<p><b>NCBI Protein Databases</b> The Protein databases are a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB (BLAST, 2011).</p>
	<p><b>UniProt</b> A merger of the information contained in Swiss-Prot, TrEMBL and PIR to produce a comprehensive database. All entries are highly annotated, some manually (Swiss-Prot and PIR) whilst other in an automated fashion using sequence similarity to previously annotated proteins (Uniprot, 2011).</p>
	<p><b>UniProtKB/Swiss-Prot</b></p>



## SHOTGUN PROTEOMICS

	A high quality annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions. UniProtKB/Swiss-Prot is the manually annotated and reviewed section of the UniProt Knowledgebase (Boutet. E., 2007).
	<b>The PRoteomics IDentifications Database</b> A centralized, standards compliant, public data repository for proteomics data. Developed to provide the proteomics community with a public repository for protein and peptide identifications together with the evidence supporting these identifications (EMBL-EBI PRIDE, 2011).
	<b>The Ensembl Group</b> Consists of between 40 and 50 people, divided into a number of teams producing genome databases for vertebrates and other eukaryotic species, and making them freely available online (Ensembl, 2011).
	<b>KEGG Pathway</b> A collection of manually drawn pathway maps developed at the Kanehisa Laboratories at Kyoto University and the University of Tokyo (KEGG, 2011).
	<b>OWL</b> A non-redundant composite of 4 publicly-available primary sources: SWISS-PROT, PIR (1-3), GenBank (translation) and NRL-3D (DbBrowser, 2011).
Figure 5. Organizations maintaining publicly available protein sequence database for proteomic studies.	

**Sequence Database Format**

The FASTA file format is a common representation of the protein sequences database. Essentially a text file, they are considered known sequence repositories. The largest FASTA file now exceeds 40 GB. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data (Blast & FASTA, 2011). The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. All lines in the file, description and sequences, must be shorter than 80

## SHOTGUN PROTEOMICS

characters in length. Figure 6 is the TVFV2E envelope protein sequence in FASTA format.

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCNSVSVVHCTNLMNTTVTTGLLNG
SYSENRTQIWQKHRTSNDALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHS
QKYNLRLRQAWCHFSPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWG
DPETANLWFNCHGEFFYCKMDWFLNYLNNLTVDADHNECKNTSGTKSGNKRA
PGPCVQRTYVACHIRSVIIWLETISKKTYAPPREGHLECTSTVTGMTVELNYIPKN
RTNVTLS PQIESIWAAELDRYKLVEITPIGFAPTEVRRYTG GHERQKRVPFVXXXX
XXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

Figure 6. An example of a protein sequence in FASTA format (NCBI Blast, 2011).

### Sequence Data Standards

Sequence databases represent the amino acid and nucleic acid codes according to the International Union of Biochemistry (IUB) and the International Union of Pure and Applied Chemistry's (IUPAC) standards. Based on the best evidence at the time, these protein sequence lists are in a constant state of flux. As proteomic discoveries are made and published, the FASTA files are revised, typically in versions. This fact adds another layer of complexity to proteomic datasets logically, as it relates to the FASTA file versions and time. Experiment results are often reanalyzed, as new releases of the FASTA files are available. Proteomic databases thus rely heavily on version annotation.

## Chapter Three: Challenges in Human Proteomics Studies

Researchers at the University of Colorado Central Analytical Facility have developed novel techniques for studying the mammalian proteome and processing the resulting data. One of their goals is the global protein characterization of melanoma cell

## SHOTGUN PROTEOMICS

lines to determine the molecular mechanisms of cancer progression and metastasis. They have discovered, like many others creating proteomic data, that the available software and conventional workflows in data analysis do not provide adequate throughput and data analysis sophistication needed for their studies (Resing, Meyer-Arendt, Mendoza, et al., 2004). Most of the analysis and sample processing software commercially available is not designed to handle the volume and complexity of this proteomic data.

In addition, the sheer magnitude of mammalian proteomes presents a difficult problem for proteomic profiling when compared to that of a simple mixture. The human proteome contains more than 12,000 proteins compared to 5,000 proteins of baker's yeast, *Saccharomyces cerevesiae*. Furthermore, more of the proteome contains homologous or similar proteins, which exist in concentrations varying over at least 12 orders of magnitude. This presents many more challenges for detecting and quantifying the proteome.

Shotgun proteomic techniques coupled with advancements in instrumentation have created a critical demand for robust and sophisticated information systems. Not only to manage the data throughout the experiment lifecycle, but also to facilitate adequate analysis of the data sets. Proteomic datasets are accompanied by the metadata that described details about the logical production and processing of the data itself. To understand an analysis, perform comparisons between datasets, or derive statistics from their aggregation, it is crucial to understand both the biological and the methodological contexts (MIAPE, 2011), much of which is captured in the metadata.

## SHOTGUN PROTEOMICS

The following chapters will include a discussion about a database system approach for bridging the gap between proteomic data production and the comprehensive analysis developed at the Central Analytical Facility in Colorado. The sample processing, design, planning, and implementation of their automated system and Oracle database will be discussed. Additionally key technologies from off the shelf solutions will reflect leading industry interests and product developments.

### **Chapter Four: Data Processing In Shotgun Proteomics**

Understanding the series of events that creates the data in the system leads to the comprehensive understanding of the laboratory's internal processes. Mass Spectrometers dedicated to proteomics and peptide analysis go through different configurations as each project or experiment is analyzed. Each instrument technique may require slightly different, to completely different settings. Scheduling machine time so projects with the same instrument configuration proceed in sequence is ideal. This adds consistency to instrument sensitivity and usually leads to better data sets. The instrument reconfiguration typically requires device calibrations and in some cases special instrument interface hardware to be installed. This process takes time as the mass spectrometer is reconfigured, conditioned and tuned for the different methods. The system's fine-tuning and adjustment to achieve the highest sensitivity is an art, time consuming, and normally minimized.


## SHOTGUN PROTEOMICS

**Data at the Instrument**





Capturing details about the instrument itself is where the data lifecycle actually starts. Data about the instrument's configuration and operating environment add value to the proteomic dataset. It enables an understanding of how the data was acquired and is the first important piece in the database design. Often over looked, this information is key to understanding laboratory processing logic and developing management strategies for scheduling instrument time.

Data products from the instruments are processed with software that interprets the file's spectra, sequences the peptides present, then searches known protein sequence databases for those peptides. The algorithmic matching of observed peptides with known peptides results in probabilistically scored protein identifications. Protein identifications using two or more search engines are preferred as this strengthens data evidence. Figure 7 is a list of the major protein search algorithm contributors and their product's description.


Detailing the design and data system in the following chapters will focus on ThermoElectrons, TurboSequest and Matrix Science's Mascot products. These products were some of the first on the market and considered by many as the unofficial standard.

	<b>TurboSequest</b> SEQUEST is the most widely used software tool for identifying proteins in complex mixtures. It is a mature, robust program that identifies peptides directly from uninterrupted tandem mass spectra. TurboSequest provides a Windows-based graphical user interface for running SEQUEST and interpreting results (Lundgren DH et al., 2005).
---	---

## SHOTGUN PROTEOMICS

	<p><b>Mascot</b></p> <p>Mascot is a powerful search engine that uses mass spectrometry data to identify proteins from primary sequence databases. The experimental mass values are compared with calculated peptide mass or fragment ion mass values, obtained by applying cleavage rules to the entries in a comprehensive primary sequence database. By using an appropriate scoring algorithm, the closest match or matches can be identified (Matrix Science., 2011).</p>
	<p><b>X! Tandem</b></p> <p>X! Tandem open source is software that can match tandem mass spectra with peptide sequences. This software has a very simple, sophisticated application programming interface (API): it simply takes an XML file of instructions on its command line, and outputs the results into an XML file. This format is used for the entire X! series search engines, as well as the GPM and GPMDB (X! Tandem, 2011).</p>
	<p><b>OMSSA</b></p> <p>The Open Mass Spectrometry Search Algorithm [OMSSA] is an efficient search engine for identifying MS/MS peptide spectra by searching libraries of known protein sequences. OMSSA scores significant hits with a probability score developed using classical hypothesis testing. The same statistical method used in BLAST (OMSSA, 2011).</p>
	<p><b>Andromeda</b></p> <p>A peptide search engine using a probabilistic scoring model. On proteome data, Andromeda performs as well as Mascot, a widely used commercial search engine, as judged by sensitivity and specificity analysis based on target decoy searches. It can handle data with arbitrarily high fragment mass accuracy. It is able to assign and score complex patterns of post-translational modifications, such as highly phosphorylated peptides, and accommodates extremely large databases (Cox, J. et al., 2011).</p>

## SHOTGUN PROTEOMICS

	<b>BiblioSpec</b> A suite of software tools for creating and searching MS/MS peptide spectrum libraries BiblioSpec 2.0 stores spectrum libraries as sqlite3 files and freely available under the BSD license (BiblioSpec 2011).
Figure 7. Industry leaders in protein search engines.	

**Protein Inference Variables**

Protein inference uses different software parameters and variables configured depending on the experiment and search engine. Each dataset has a set of attributes that specifically describes the data itself. The *metadata* (data about data), also called *metacontent*, is critical for performing comparisons between datasets. In an automated system, metadata is collected about every process event as the data flows from the instrument to the database.

**Entity Relationship Diagrams (ERD)**

Creating a data flow diagram and entity relationship diagram assists the Database developer to design normalized tables for both the result data and metadata produced in the system. The diagrams are tools to help conceptualize the process and code the database structures to efficiently accommodate the data. Figure 8 illustrates the TurboSequest dataflow Entity Relationship Diagram. The original Graphic User Interface (GUI) from the manufacturer's software parameters are used to identify the critical program variables. The parameters for invoking this package can be passed directly to the software with an OS batch file at the command line (discussed later in this chapter).



## SHOTGUN PROTEOMICS

Capturing this information insures consistent processing and a complete history of the data. Information about the biological sample, processing workflow and the analysis methods used to create the experiment's results are critical pieces in the system. A logical mapping of these data elements and an idea of how the database table attributes should be defined are drawn from the diagram.

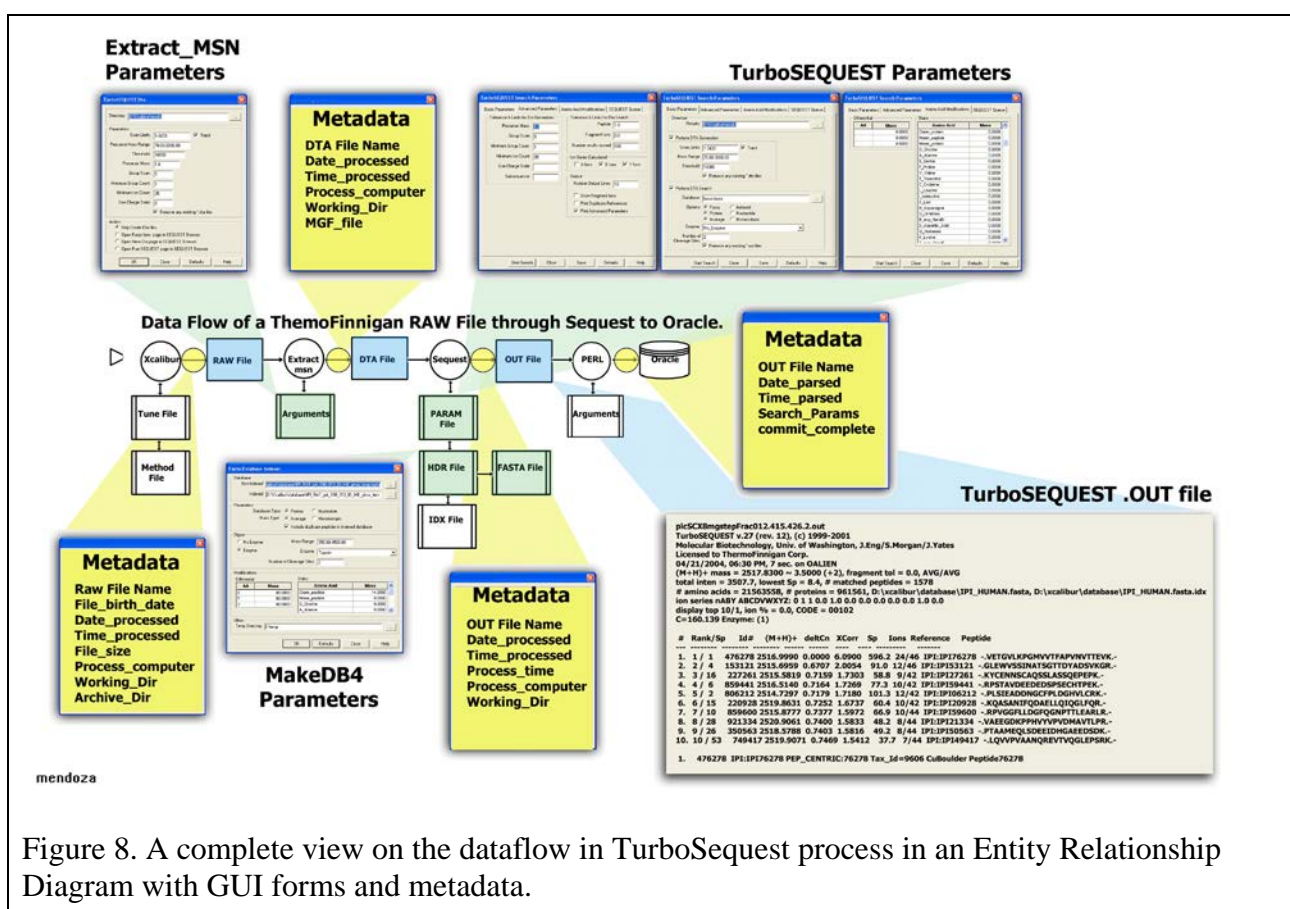


Figure 8. A complete view on the dataflow in TurboSequest process in an Entity Relationship Diagram with GUI forms and metadata.

## Using Design to Create Structure

Defining the specific entities in the dataflow conceptually helps create the database structure. Each entity has attributes that describe it and relationships connecting it in the system. In figure 8, the Metadata appears in the yellow windows and occurs after



## SHOTGUN PROTEOMICS

every process. Metadata is an important component in proteomic information systems. It assists in creating data management strategies, quality control, and projecting storage utilization.

Storing protein identification data in a relational database leverages time tested Relational Database Management System (RDBMS) technologies as well as enhances informatics and collaborative computing opportunities. The Structured Query Language (SQL), management features like backup recovery tools, SQL based software, and direct connectivity options make relational database systems ideal for proteomic data warehouses. Software developers can leverage the database logic, packages, views and functions to quickly prototype novel code or algorithm ideas. This architecture minimizes file handling by enabling users to connect directly to filtered datasets in the database. Mapping to the original files stored in a archived location and having only significant data, the best data statistically on hand in the database, conserves expensive *fast-read* disk resources and adds a level of efficiency to the informatics.

Creating different file formats like mzXML or mzML (discussed later), and producing subsets or super sets of the data, statistically filtered data sets, can be done using simple queries over multiple experiments. Database direct port connection via ODBC or JDBC will connect other analytical applications like SpotFire or R, and even Spreadsheets like Excel or CALC for users in the lab. Most of the protein search engine packages available do not support direct database connectivity. Those that do support direct connections take on more of a data *pipeline* characteristic; a custom tailored data processing for the specific method or analysis, covered later in the paper.

## SHOTGUN PROTEOMICS

## From Flat File to Relational Database

Creating software that will convert output data into rows in a database has many labs writing custom scripts in Perl or Python to process data from the search engines output to a database. These scripts are used to capture both the information about the result file for example, the file's creation date, file size, file location, and the pertinent result data (experiment data) in the file. Commonly called *file parsers*, these scripts ingest the file's information into relational tables. Figure 9 highlights conceptually the data taken from each file. Text is ingested into tables by the file parser that converts lines from a ThermoElectron .OUT file into rows in an Oracle database while reporting the file's metadata.

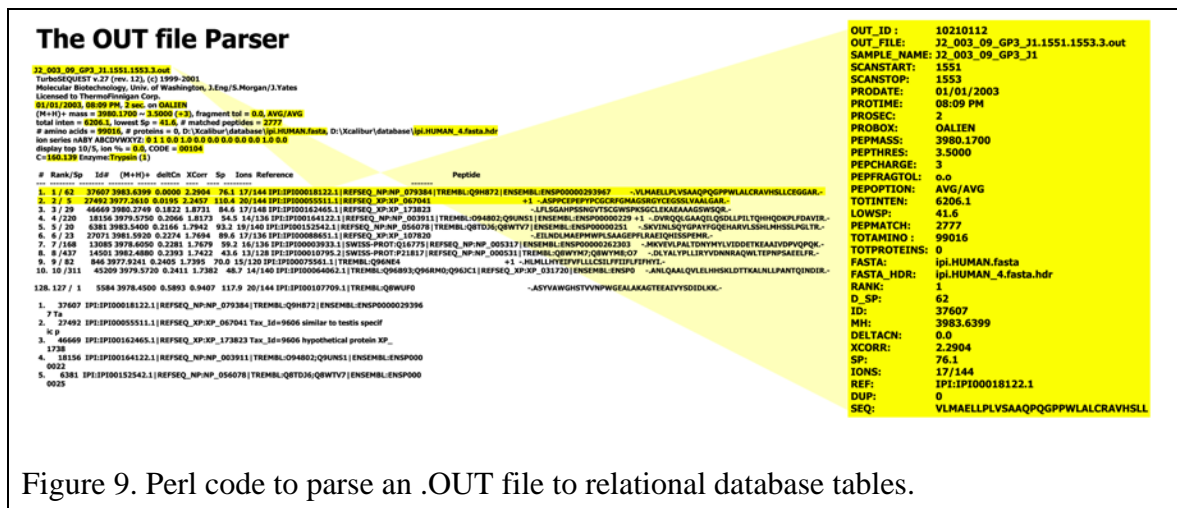


Figure 9. Perl code to parse an .OUT file to relational database tables.

Each output file format requires its own logic and parser code. This particular parser takes the top two ranked matches in the TurboSequest identifications. The Perl script utilizes the Perl database interface (DBI) to input the data directly into the database

## SHOTGUN PROTEOMICS

tables. From here the .OUT can be compressed and archived to slow media, its filtered contents available via the database which resides on faster disk resources.

Each instrument vendor has a proprietary data format and analysis method. File parsers are created for the specific data types, but must stay agile, as they require updates when the output formats change with different versions of the vendor instrument's software. Comparing the search results and correlating the top hits with other software packages is difficult to impossible with vendor analysis tools alone. Collaboration on datasets, comparing between search engines, and comprehensive analysis requires additional 3rd party software or custom information systems.

### **Automation and SmartJobs**

Enabling automatic proteomic data processing simply stated is, orchestrating the essential technologies of the dataflow at the command line. This logic sidesteps the traditional workflow. Only utilizing the required underlying proprietary code, such as the protein search algorithms in the dataflow, has facilitated the freedom to create a novel automated high-throughput approach. Representing the complete proteomic dataset in relational tables enables process optimization and enhances informatics by minimizing file handling.

Jobs in the form of batch files are created based on specific parameters assigned by the database with Perl scripts. These Perl scripts direct file traffic and assigned search engine jobs to the processing nodes. The jobs are preconfigured, called SmartJobs; they essentially represent the automation logic and mechanism. Although Perl scripts create

## SHOTGUN PROTEOMICS

the SmartJobs, their configuration is dictated by the database. That is, the instructions needed to form the smart jobs are not hard-coded in the scripts; rather the Perl scripts reference the parameters from the database for each job. This enables the logic that dictates sample processing and resource allocation to be managed at the database level.

Figure 10 is an example of a SmartJob used to process TurboSequest data on processing nodes running the Windows OS.

```

REM# SmartJob_ID: 1941
REM# Data_file: plcSCX8mgstepFrac005_20020408
REM# @_Node: BELLES.colorado.edu

REM# Make dir and copy data file from archive directory.
mkdir I:\Xcalibur_I\results\plcSCX8mgstepFrac005_20020408
COPY \\bluesky\DTA\20020408\plcSCX8mgstepFrac005_20020408.mgf
I:\TurboSequest\results\plcSCX8mgstepFrac005_20020408

REM# Expand .MGF format to .DTA format for TurboSequest.
I:
cd I:\TurboSequest\results\plcSCX8mgstepFrac005_20020408
D:\Automation\expand_mgf.pl plcSCX8mgstepFrac005_20020408.mgf

REM# Run TurboSequest
C:\Xcalibur\system\programs\BioworksBrowser\sequest27.exe -
DD:\Automation\ipi_HUMAN\ipi_HUMAN_v3_00_2.fasta -
ID:\Automation\ipi_HUMAN\ipi_HUMAN_v3_00_2.fasta.idx -
PD:\Automation\ipi_HUMAN\ipi_HUMAN_v3_00_2.params *.dta

REM# Copy result to archive directory
COPY *.OUT \\bluesky\calibur\results\plcSCX8mgstepFrac005_20020408\
ipi_HUMAN_v3_00_2

```

Figure 10. A SmartJob or batch file DOS (Windows), created by a Perl script to process data. This Smart job is configured to make a directory and pull (copy) data from a centralized repository to a processing node (BELLES), then run TurboSequest with a specific FASTA file and parameters. After the search engine completes it then copy the output (.OUT) to an archive location. Comments are annotated with REM##.

## SHOTGUN PROTEOMICS

The lab's logic requires a high level of confidence with two or more search engine results in the protein identifications to publish their findings. Essentially these search engines generate the same type of data in different formats using their own flavor of technologies (Java, C#). This requires database tables and automation scripts specifically designed for processing, storing and comparing results between the different search engines.

### **Chapter Five: Database Design for Proteomics**

Metadata contributes to the sample processing management logic. Load balancing, quality control and scheduling of the instruments are all measurable using this repository. The experiment result data also resides in the database. Working copies of the production data is delivered with the speed and efficiency of a RDBMS. The next sections discuss how database technology can benefit proteomic data processing and details the design of the Oracle database to support TurboSequest results.

#### **A Peptide-centric Homogeneous Database**

The file parsers capture flat files information into relational tables from the different search engines. Once in the data repository the data takes on homogeneous characteristics. This enables code development in analytics and data management capacities to increase. Optimizing protein search algorithms, writing software for reducing the false positive and false negative frequencies, maximizing reproducibility, and generating statistically filtered datasets in multiple formats are common uses of this database. A peptide-centric database provides a concise data repository for analytical

## SHOTGUN PROTEOMICS

software development. Having a normalized, indexed repository of proteomic data is the return on the investment and the top requirement for the system developer.

Design of the database is done by analyzing the relationships that exist in the dataflow between the major processes then defining a table structure for each entity. Creating a series of scripts that populate the tables and defining logic to systematically storing the instrument data files in a data warehouse completes the automation steps. The dissemination of this data enables the development of analysis logic that supersedes a single flavor of protein search engine or file format. The entire industry is moving towards using multiple protein search engines for each dataset to identify validate and compare results.

Bioinformatics and search algorithm development benefits tremendously from having well studied datasets to compare baseline results. Standard datasets in an annotated repository are a strong foundation for comprehensive analytics software development. Using known mixtures of protein samples to test instrument sensitivity and having well studied datasets to test an algorithm's effectiveness enhances the lab's analytics. Developers can leverage this platform with application frameworks to construct experiment datasets for algorithm development. Data management user interfaces and software languages like Perl, PHP, Ruby, PL/SQL, and Python are just a handful of the tools one can utilize to connect directly to the database. A database platform provides a high level of accommodations for creating management and analytical source code.

## SHOTGUN PROTEOMICS

**TurboSequest Tables Structure**

The schema in Figure 11 represents the working TurboSequest database tables used in 2003 for proteomic processing. The data in the database is normalized and organized into tables to minimize redundancy. The schema was developed for processing ThermoElectron LCQ™ Classic, Quadrupole Ion-Trap Mass Spectrometer data with their TurboSequest protein search engine. Most of the table specific attributes have been omitted from the diagram so that we can focus on the relationships and logic of the design based on the Entity Relationship Diagram (ERD) of the process (see Figure 8).

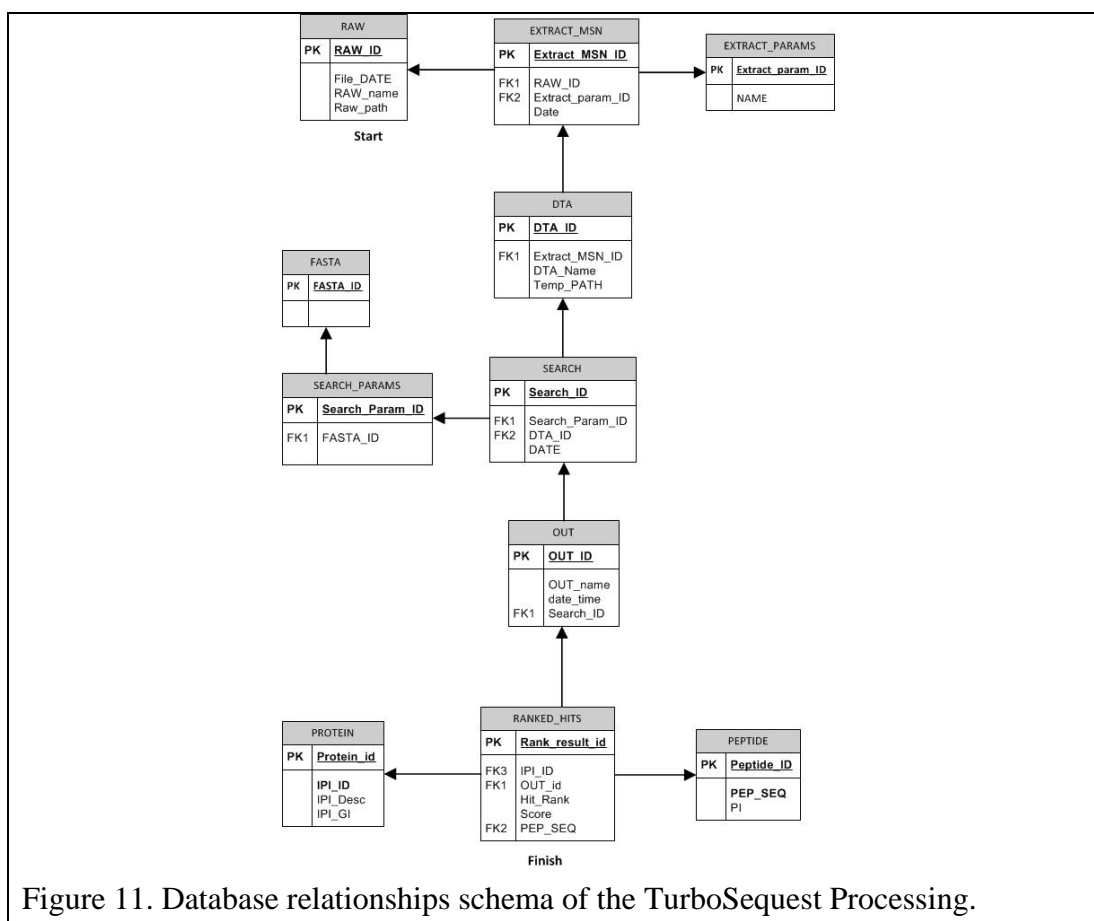


Figure 11. Database relationships schema of the TurboSequest Processing.

## SHOTGUN PROTEOMICS

Start by examining the RAW table and its relationship with the EXTRACT\_MSN table. This join represents the event of a RAW file being processed by *extract\_msn.exe*, a utility that uses a list of parameters (EXTRACT\_PARAMS) to convert the instrument's .RAW file into a set of MS/MS values called .DTA files. The DTA table has a one-to-many relationship with the RAW table, through EXTRACT\_MSN. The .DTA files are not captured in the database or archived in the data warehouse; rather they are generated on-demand in seconds as necessary.

Identifying the peptides in the mass spectra, now in .DTA format, with the protein search algorithm TurboSequest is represented by the SEARCH table. This table links the .DTA file, to the TurboSequest search and the parameters (SEARCH\_PARAMS) used. A FASTA sequence database is part of the search parameters. The search software creates an output called an .OUT file. The protein search algorithm will return one .OUT file for every .DTA file, but not every .OUT file will contain a valid identification or *ranked hit*. The peptide and protein relationships relates to the original .DTA file as a ranked hit, represented in the schema in the RANKED\_HITS table. All of the ranked hits in the search are related to the International Protein Index (IPI) via their IPI\_ID. The International Protein Index contains a number of non-redundant proteome sets from higher eukaryotic organisms and is the standard in the industry.

**The RANKED\_HITS Relationship**

The table relationships are important for normalization. RANKED\_HITS relates the FASTA sequence database via the result (.OUT) files and the search method used. As



## SHOTGUN PROTEOMICS

new FASTA files are released and a given dataset is researched, a new instance of RANKED\_HITS is created. This relationship makes it possible to track results for a given identification with each FASTA version. It creates a one-to-many relation with the original MS/MS (.DTA file) and the different FASTA searches it may undergo. A comparison of this data can quantify the significance of the FASTA updates.

### **Chapter Six: Proteomics Laboratory Infrastructure**

High throughput proteomic data analysis requires the computational power and infrastructure like that of a small Internet startup. One of the major challenges in beginning proteomic studies is coupling a biochemistry laboratory focused on mass spectrometry to a data center for the informatics. This has traditional biochemist scrambling to learn, implement and manage the computational resources.

#### **Vulnerability at the Acquisition Workstations**

The University of Colorado's Core facility operates its proteomic informatics on a secured private network. The most valuable hardware in the system, the mass spectrometers, must be in a network with limited to no outside access from the Internet. This is due to the vulnerable nature of the acquisition workstations that directly interface the instruments. The instrument vendor often will not support operating system security patches fast enough to keep up with Internet threats. This is a serious vulnerability of a critical component, one that carries considerable risk to core processing, and the lab's most expensive hardware. Figure 12 is a schematic diagram of the network topology in the facility. Stretched over two buildings, the Cristol Chemistry and Ekeley Sciences

## SHOTGUN PROTEOMICS

buildings on the CU campus, the virtual private network creates an encrypted virtual tunnel over the Internet to ensure secured computing. These two buildings are connected via a Cisco VPN concentrator. The laboratory's bench workstations connect to the Internet via a DHCP offered by the university's network service outside of the private network.

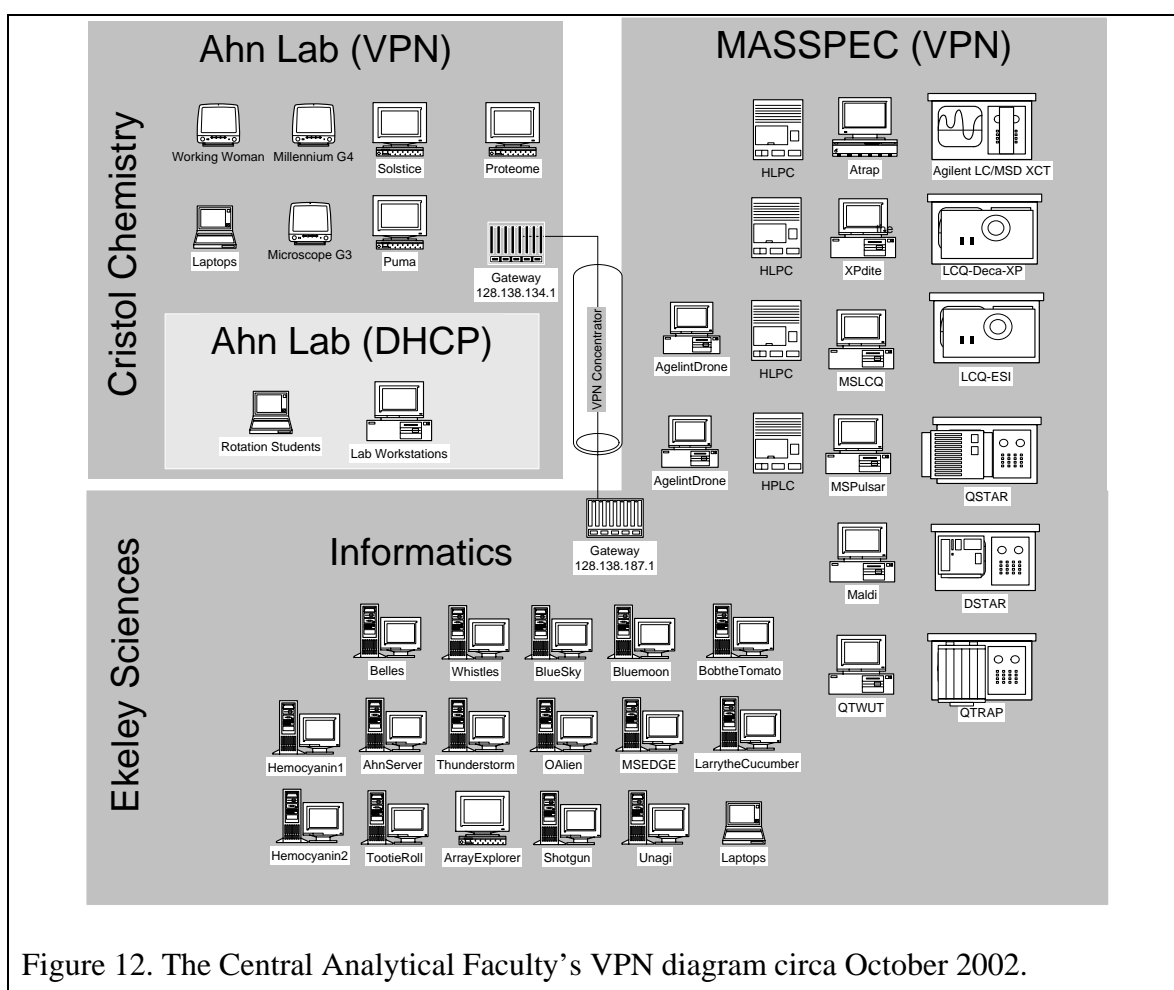


Figure 12. The Central Analytical Faculty's VPN diagram circa October 2002.

The laboratory's core processing and informatics are accommodated on a virtual private network with limited to no connectivity to the Internet. A secured intranet

## SHOTGUN PROTEOMICS

provides a safe and relatively cheap way of connecting to web servers, and data portals with role based access control. This protects the instruments, computing assets and automated processes by funneling network traffic to manageable entities which minimizes risk to critical processes.

The Central analytical facility's network spans multiple buildings on the CU campus and allows end users to connect to the data repositories via hardened Oracle web servers, SFTP, and Mascot (Apache) services. The out-facing machines have the latest OS security patches and managed services. Extending connectivity via remote access VPN enables secure data exchange of proteomic results to authorized users only, while completely isolating the data production and informatics. Instrument workstations, processing nodes, and database servers are shielded from direct connectivity, maintaining patch versions still compatible with the required production software.

### **Hardware Resource Utilization and Cost**

A quick breakdown of computing power and equipment needed for the standard automation system starts at the mass spectrometer. The instruments alone are about \$500,000 each, so instead of listing a complete expense list for shotgun proteomics studies (it is expensive), this section will focus on the minimum computer hardware needed for constructing an automated shotgun proteomic system.

The data from the instrument is collected by the acquisition workstations running vendor software that controls its electronics. Vendors recommend a particular class of desktop computer for each system and these workstations usually ship with the mass spectrometer. Configured by service technicians during the initial setup and collaboration

## SHOTGUN PROTEOMICS

of the instrument they mostly run instrument software. The average cost of these workstations is about \$1300 with licensing and support agreements accompanying the instrument purchase.

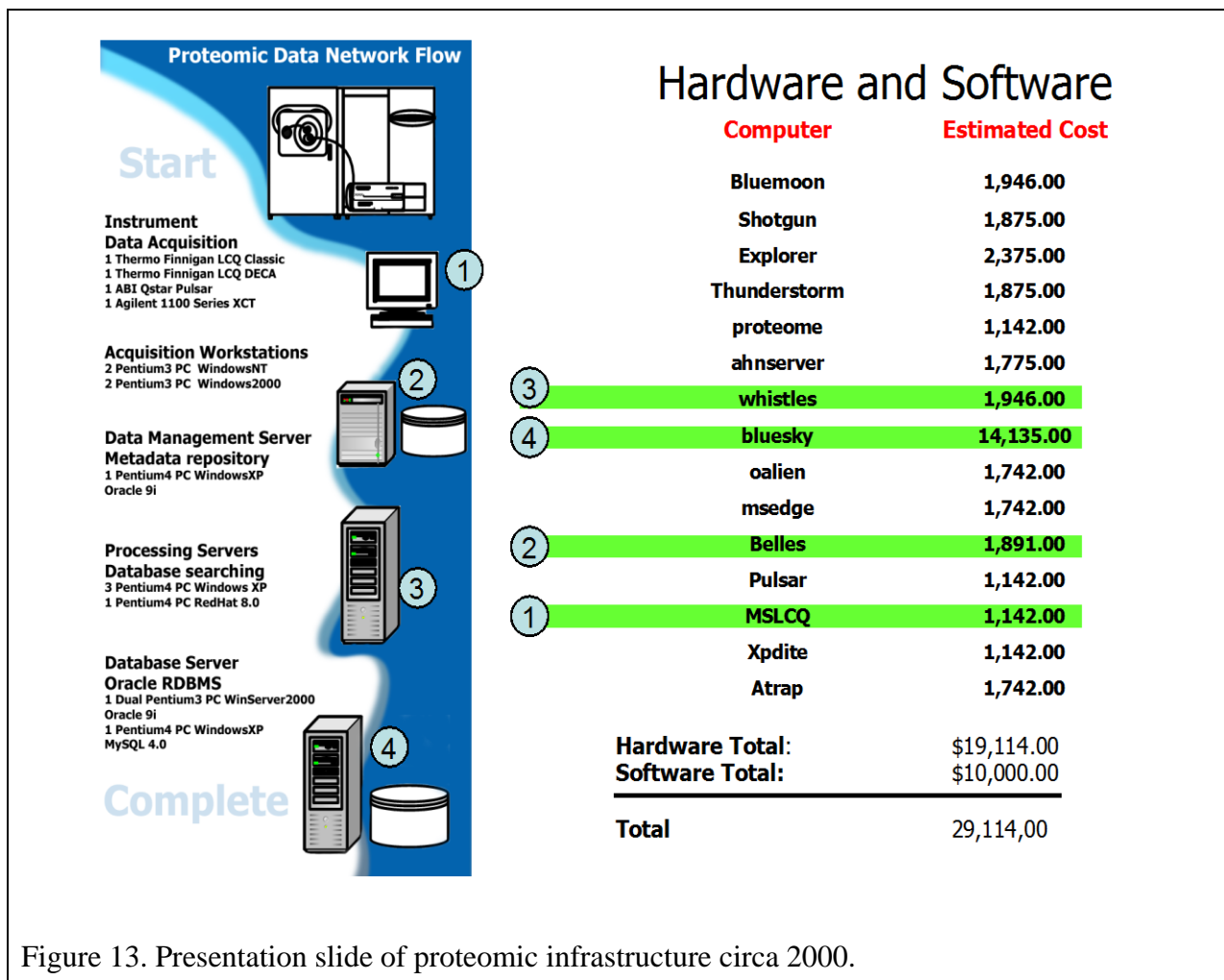
The acquisition workstation creates its instrument output file in a standard disk location. Automation scripts copy these files to a data management server where they are automatically archived. Metadata about these files is reported to the database server, which queues the RAW files for automated processing. The data management servers are high-end workstations able to process large numbers of files with web services, and database connectivity. This data repository is handy for catching instrument malfunctions. The metadata will report file size, or MS/MS scan variations outside the expected range. Instrument output file statistics, when viewed objectively over time, can be a reliable gauge of instrument health.

Protein search algorithm(s) act on the uninterrupted MS/MS datasets on the processing nodes. This step is processor intensive, so the machines designated are more powerful machines and integrated together to generate a high performance cluster. These computers cost \$2000 or more each and in some cases require additional software licenses per node for the protein search engines, Mascot (~\$4000 additional/node) or TurboSequest (~\$6000 additional/node).

Finally, the data from the protein searches, the metadata from the processing and any additional data, like generated statistics are housed in the central Oracle database. This machine has both processing power and large storage capabilities in the form of RAID storage or attached storage network appliances. There is RDBMS license expenses

## SHOTGUN PROTEOMICS

when using Oracle Standard Edition license for academic use of about \$10,000. Figure 13 illustrates the laboratory data flow from the mass spectrometer to the database and breakdowns the approximant costs of the computer hardware.



## Chapter Seven: Proteomics Software Development



In the last decade we have seen a steady increase in software specifically intended to store, manage and analyze proteomic data. Laboratories have a wide variety of both open source and vendor supported products to choose from if they do not actively create

## SHOTGUN PROTEOMICS



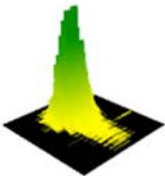
code. Developing code in-house is expensive and may not be feasible as it adds tremendous cost and overhead. Off the shelf implementations are a practical option for small or developing proteomic laboratories that may not have the resource to develop software.

### Proteomic Data Management and Analysis Software

Many of the third party proteomic applications available provide a management framework for proteomic datasets and utilities to analyze results. One area of focus is to improve the confidence of the protein identification by comparing results from the different search engines and effectively visualizing that data in a Graphic User Interface (GUI). Protein search engine data, the various parameters, sequence databases, and MS/MS instrument data are managed within the framework itself. Sometimes called proteomic *pipelines*, users can import, compare, quantify and validate their protein search results with these products. Figure 14 lists five different proteomic data solutions that have enabled labs not currently developing in-house solutions to process results.

	<p><b>Scaffold</b> Tools to help medical researchers confidently identify proteins in biological samples. Using output from SEQUEST®, Mascot®, or X! Tandem, Scaffold validates, organizes, and interprets mass spectrometry data, so you can more easily manage large amounts of data, compare samples, and search for protein modifications (Scaffold, 2011).</p>
	<p><b>Phenyx</b> Developed in collaboration with the Swiss Institute of Bioinformatics (SIB), Phenyx is GeneBio's renowned software platform for the identification, characterization and quantitation of proteins and peptides from mass spectrometry data. Specifically designed to meet the concurrent demands</p>

## SHOTGUN PROTEOMICS

	of high-throughput MS data analysis and dynamic results assessment, it offers a highly flexible user interface and an adaptable architecture that helps instill confidence in results assessment (GeneBio, 2011).
	<b>ProteinProphet</b> An Open Source product that automatically validates protein identifications made on the basis of peptides assigned to MS/MS spectra by database search programs such as SEQUEST. Allows filtering of large-scale proteomics data sets with predictable sensitivity and false positive identification error rates (Nesvizhskii, A. I., 2003).
	<b>PRISM</b> Proteomics Research Information Storage and Management system (PRISM) provides a platform that serves the needs of the accurate mass and time tag approach developed at Pacific Northwest National Laboratory. It incorporates a diverse set of analysis tools and allows a wide range of operations to be incorporated by using a state machine that is accessible to independent, distributed computational nodes (Kiebel G.R., 2006).
	<b>MaxQuant</b> An integrated suite of algorithms specifically developed for analyzing large mass-spectrometric data sets. Using correlation analysis and graph theory, MaxQuant detects peaks, isotope clusters and stable amino acid isotopes–labeled (SILAC) peptide pairs as three-dimensional objects in m/z, elution time and signal intensity space (Cox, J., & Mann, M., 2008).
Figure 14. Leading proteomic applications.	

**Proteomic Data Standards**

The Human Proteome Organization's (HUPO) Proteomics Standards Initiative has developed guidance modules for reporting the use of techniques such as gel electrophoresis and mass spectrometry (Taylor C. F., et al 2007). They have addressed the issue by creating a framework of modules that provide specific guidelines for reporting proteomic data. These guidelines are useful in mapping database attributes with

## SHOTGUN PROTEOMICS

the community standards. The minimum information about a proteomics experiment or (MIAPE) guidelines, intent to represent the data with two general criteria:

1. Sufficiency.

The MIAPE guidelines should require sufficient information about a dataset and its experimental context to allow readers to understand and critically evaluate the interpretation and conclusions, and to support their experimental corroboration.

2. Practicability.

Achieving compliance with MIAPE should not be so burdensome as to prohibit its widespread use.

Major contributors in the field are seeing the value in maximizing returns on datasets that are expensive to produce. Publicly available datasets have enabled collaboration and an open source environment for analyzing, annotating, and creating software together. This increases the value of the dataset and provides incentives to create policies to encourage standard compliance. For example, the UK Biotechnology and Biological Sciences Research Council (BBSRC) have finalized a policy statement that requires plans to be established for prevising the access of datasets that were generated in the course of BBSRC-funded work. Many other funders, such as the US National Institutes of Health and the National Science Foundation also require adherence to agreed community standards, where they exist. Figure 15 presents the MIAPE-compliant data management dataflow. Notice the heavy emphasis on metadata collection.



## SHOTGUN PROTEOMICS

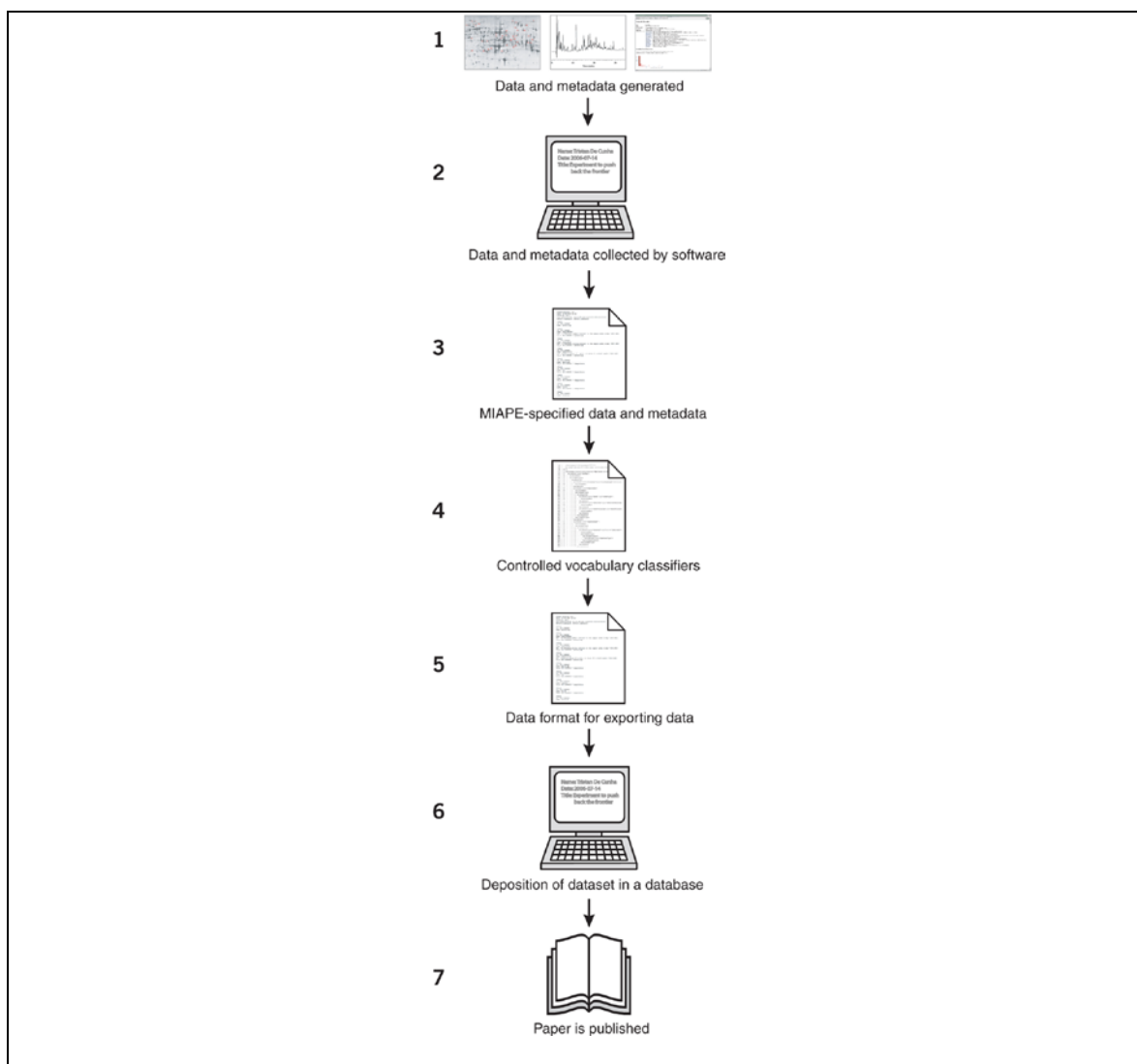


Figure 15. An example of MIAPE-compliant data management dataflow (1) Data and metadata are generated by an experiment; (2) some form of software collects the data and metadata, either by importing from computer-controlled instruments (better) or from manual data entry; (3) MIAPE specifies the data and metadata to be requested by the software tool; (4) a controlled vocabulary supplies classifiers via the software; (5) the software uses a data format specification when exporting a MIAPE-compliant dataset; (6) the dataset is stored in a MIAPE-compliant database and assigned an accession number; (7) data, including the appropriate accession number, is published in a journal (Taylor C. F., et al 2007).

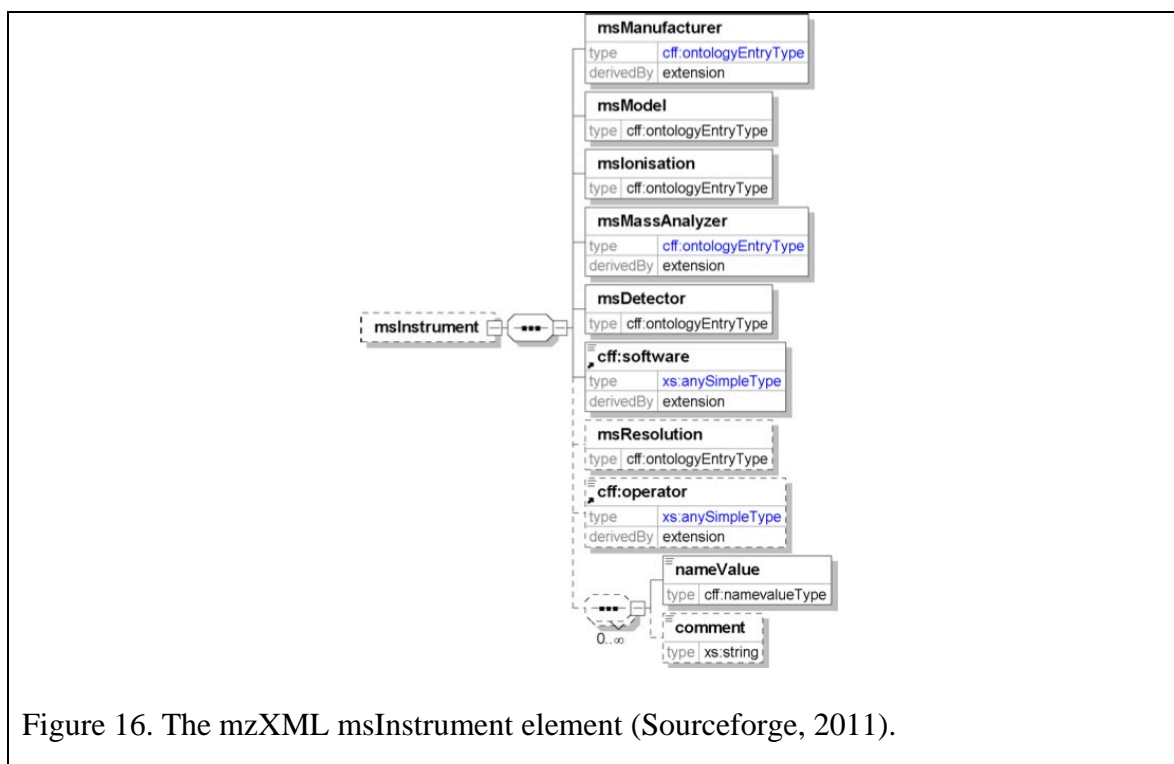
## SHOTGUN PROTEOMICS

The MIAPE-compliant data management highlights many significant areas in proteomic data processing. Shotgun proteomics had to prove its merit as a technique before the community could agree on a data management framework. Furthermore, a standard data format took considerable time to construct and even longer for the community to recognize.

**Open Data Standards for Proteomics**

The Institute for Systems Biology (ISB) leads the development of an open data standard for proteomics. They have created XML open formats in an effort to streamline data pipelines and software collaborations. The open mzXML and more recently, early 2009, mzML provide standard data containers for MS/MS data directly from the raw file using a *raw-to-mzXML* converter. These converters support most instrument output files. This open format is ideal for collaboration and web services. The data format can represent a number of different aspects and details about the data file including its metadata. Figure 16 is the XML container to store information about the mass spectrometer. This element captures the specifications of the MS instrument (e.g. resolution, manufacturer, model, ionization type, mass analyzer type, detector type) and the acquisition software used to generate the data. Having the metadata stored in the file format is a step in the right direction, but it becomes incredibly redundant with large datasets and can consume storage resources.

## SHOTGUN PROTEOMICS



### Software Explosion

Software products range from basic data mash-ups to comprehensive analysis platforms and pipelines. Efficiently analyzing large datasets, protein quantitation, sharing data for collaboration, visualization, and algorithm development are just a few areas of interest. Drug discovery and healthcare are currently two of the largest market sectors. Software focused on critically testing and validating high-coverage peptide profiles for drug development and data management systems have recently flooded the market.

### Proteomic Software Market Value

Functional genomics and proteomics have been quite successful in identifying cellular functions of potential therapeutic targets and their market values have reflected this. According to a recent report released by Global Industry Analysts, Inc., The global

## SHOTGUN PROTEOMICS

proteomics market is projected to reach \$6.1 billion by the year 2015, driven by increasing adoption in life sciences research in both developed and emerging markets (PRWeb, 2011).

Their report cites the sales of existing products such as microarrays, and biochips as well as new products including test kits for detecting airborne chemical warfare agents and protein structure modeling tools is steadily increasing. This technology is being embraced in a diverse range of industries. For proteomics to continue growing, it will need to develop better methods and technologies to speed validation of critical components.

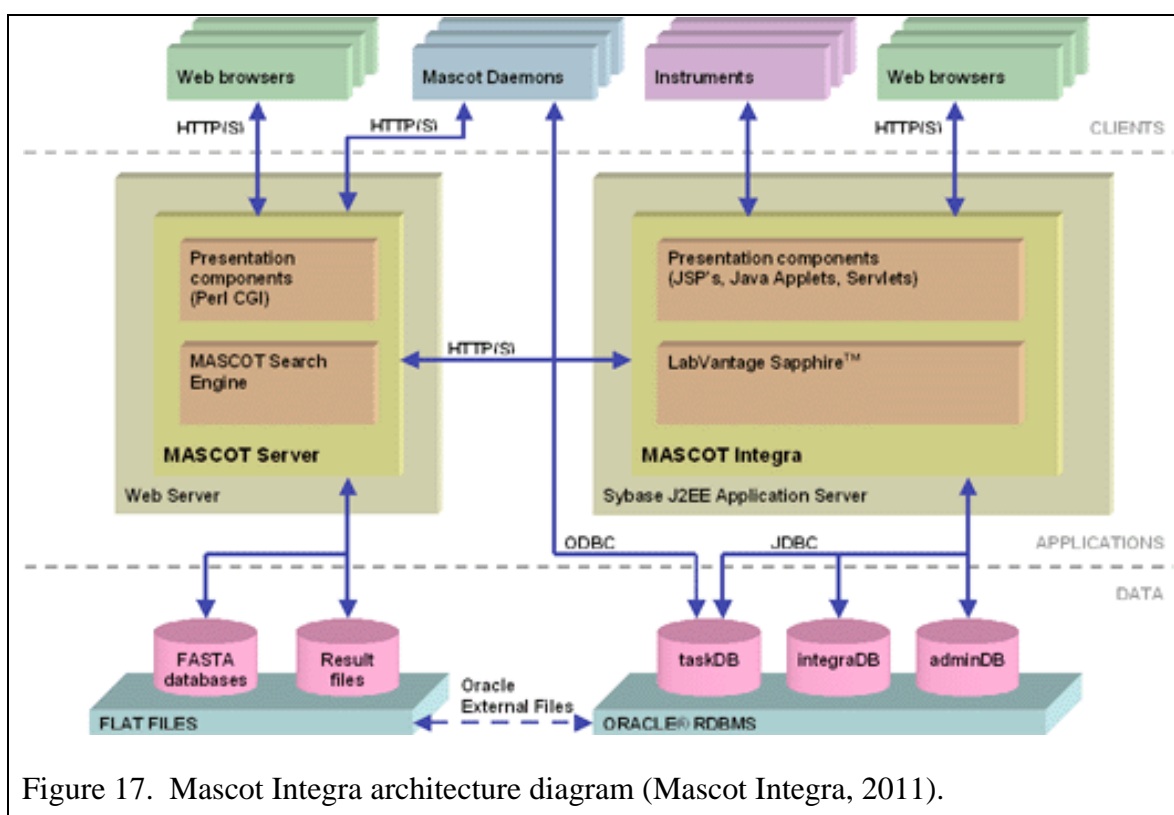
Validation of high-throughput proteomic technologies used to discover potential biomarkers and drug targets is critical for minimizing the associated costs and risk with producing drug candidates. Automated collection, validation and analysis software are opportunities in the current market. Software products being developed by research institutes can directly contribute to new products in software and lead to method development used in drug production. Proteomics technologies were worth \$1.3 billion in 2008 and an estimated \$1.5 billion in 2009. This segment is expected to increase at a rate of 13.9% to reach \$2.9 billion in 2014 (PRWeb, 2011).

### **Third Generation Products**

The proteomic software industry is now seeing second and third generation software to manage and analyze proteomic data in one application. One such system is called *Mascot Integra*. This Matrix Science product was developed in collaboration with

## SHOTGUN PROTEOMICS

Lab Vantage Solutions Inc., a well-known producer of laboratory information management systems also called LIMS. This turn-key system was created to cover all the steps involved with processing and analyzing data in a typical proteomic experiment. Shipping with hardware specifically designed for the Oracle database and software (web services) used in the system. Figure 17 is an overview of the Mascot Integra architecture.



### Mascot Integra

Mascot Integra is an application server that accesses a database of both relational tables and flat files in a multi-tier architecture. Leveraging the protein search engine Mascot at the web browser, this architecture utilizes well-developed and time-tested technologies to produce a robust system custom tailored for proteomic studies. The Java

## SHOTGUN PROTEOMICS

2 Enterprise Edition platform powers the presentation component of the system for reporting. A Sapphire thin client is utilized for both viewing and managing proteomic information over the Internet. The system also uses the legacy Mascot Daemon via the Perl CGI to automate Mascot processes. These two pieces exchange data via https and leverage Oracle's role base permissions for access control.

**Trans-Proteomic Pipeline (TPP)**

A different approach in creating a complete *pipeline* for proteomic data has been developed at the Institute for System Biology (ISB) and uses a wide range of existing open source software products to create what they call the *Trans-Proteomic Pipeline (TPP)*. This product can be installed on a variety of operating systems, Microsoft Windows, UNIX/Linux, and MacOS X. It aims to standardize the data format and provide a single platform for proteomic processes. The TPP claims to be the oldest and most comprehensive software suite available and has tools for MS data representation, MS data visualization, peptide identification and validation, quantification, and protein inference (Deutsch, E. W. 2010). The TPP like most software suites for processing proteomic data has a number of file converters, sequence searching tools, data visualization, and statistical modeling packages in the suite. This software, instead of housing the data in a database, converts the data into a vendor-neutral format, mzXML or mzML, and stores it in the file system.

## SHOTGUN PROTEOMICS

**TPP Dataflow Explained**

Figure 18 is a dataflow for the TPP product. Raw MS/MS data files are first converted to an open XML format such as mzXML and then analyzed with a search engine, either embedded in the TPP or used externally. Pep3D allows visualization of the data. First, the search results, in pepXML format, are processed with PeptideProphet for initial spectrum-level validation, iProphet for peptide-level validation, and finally ProteinProphet for protein-level validation and final protein inference. Quantification tools like XPRESS, ASAPRatio, or Libra can be used on labeled data. The final output is protXML, which can be imported into a variety of analysis tools (Deutsch, E. W. 2010).

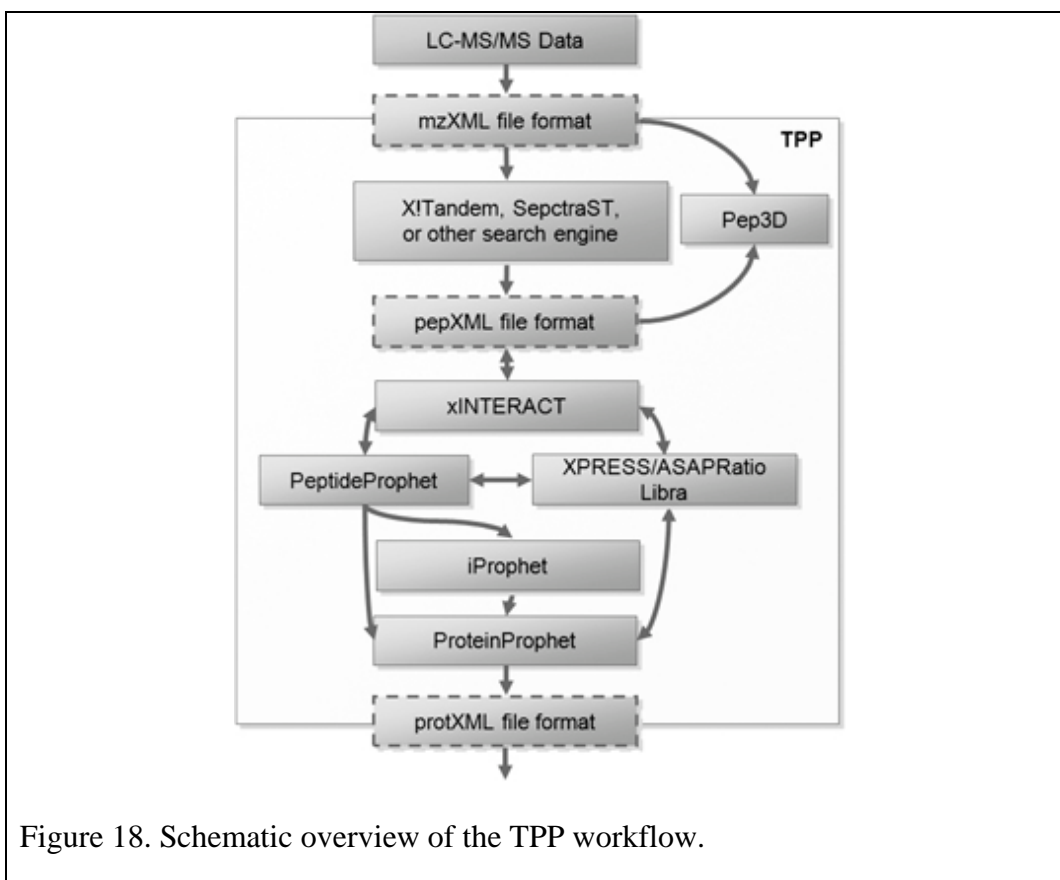


Figure 18. Schematic overview of the TPP workflow.

## SHOTGUN PROTEOMICS

Among TPP strengths is its comprehensive search algorithms support. The product covers most of the leading search engines including, X!Tandem, Mascot, TurboSequest, OMSSA 23, Phenyx 24, and ProbiD 25, although only ships with the open source X!Tandem, other products must be purchased and installed separately. Another key feature with this product is the ability to search previously identified spectral libraries. Basically FASTA files created with sequences already identified in the project. The spectral searches benefit from a smaller search space (fewer candidates to choose from) and the use of real reference spectra as opposed to theoretical ones predicted, often simplistically, by sequence search engines (Deutsch, E. W. 2010).

**Discussion of Mascot Integra vs. Tans-Proteomic Pipeline (TPP) Data Storage**

One difference with the two products is, Mascot Integra uses a database to house the experiment data and Tans-Proteomic Pipeline utilizes the file system and the mzXML format. TPP converts all experiment data to a standard XML format and interacts with it through various applications on the files system, whereas Mascot Integra uses more of a database centric architecture, relying on a multi-tier application layer to present the data. Both approaches have merit, advantages and disadvantages.

Instrument vendors have their own flavor of operations, their own file format and preferred software, converting everything to a standard XML format enables an open source community, like that developing TPP, to quickly work together and avoid complex format variations. The HUPO Proteomics Standards Initiative endorses the standard and specifies the XML schema definitions. The self-described schema contains



## SHOTGUN PROTEOMICS

metadata that can be read, understood and parsed easily by software. This format is ideal for sharing and collaboration and has been highly adopted in the software industry.

Because xml has been widely used for data exchange it is too easily accepted as a data storage model. Although a storage model can be built on XML's tree structure, XML was never designed to store data. These formats represent the mass spectrometry data ready for exchange or collaboration. Large XML file systems can be difficult to leverage and backup over time. Better solutions for storing the XML datasets exist in the database.

**Native XML Databases (NXD)**

Most database platforms support a XML data type. This data type while adding some overhead in query development and to the overall size of the dataset has some benefits. One of which is a native XML database type can perform faster than an application accessing a file system of XML files. Storing, maintaining and querying large datasets is much more concise in the database because the DBA can leverage built in database utilities and tools. The relational mapping and storage of the data in the database has greatly improved XML storage efficiency, flexibility and transparency. Whether or not this storage method is better than traditional relational database modeling, where the XML file is parsed into relational tables, is an argument of style and preference. Some argue that XML is ill suited for specifying complex metadata with dynamic dependences, as we see in shotgun proteomics. TPP and other services that use

## SHOTGUN PROTEOMICS

an XML based model would be wise to integrate a database option into their software suites.

### **Chapter Eight: Proteomic Information Future**

Software products have accomplished many of the significant challenges presented in early proteomic studies. Adapting known information technologies to solve processing and informatics issues has enabled the next stage of discovery. As instruments become even more sensitive, method development and bioinformatics exploration will challenge today's information systems and pushes the industry to produce new ideas, databases and software products.

#### **Microarray Technology**

Microarray technology is being applied in some proteomic methods to analyze proteins in specific tumor cells. Ciphergen Biosystems, Inc. has developed the surface-enhanced laser desorption-ionisation (SELDI) Protein ChipR, that involves the affinity capture of specific subgroups of proteins based on their biochemical and/or physical properties, coupled with automated MS analysis (Verrills, 2006). These techniques have proven useful in analyzing the protein patterns of serum from ovarian cancer patients and development of a commercial test, termed OvaCheck, for diagnosing ovarian cancer currently in clinical trials.

These chips for analyzing known biomarkers have potential for studying specific signaling pathways for both enzymatic activity of secreted and membrane proteomes as well as kinase activity via specific detection of phosphoproteins. This type of analysis

## SHOTGUN PROTEOMICS

could also be applied to monitor the response of patients to chemotherapy (Verrills, 2006).

### Mass Spectrometry Imaging

Another potentially exciting development, from Vanderbilt School of Medicine is the direct mass spectrometry imaging of proteomic data in frozen tissue samples. Figure 19 shows images created from proteomic datasets from mouse tumor cross sections.

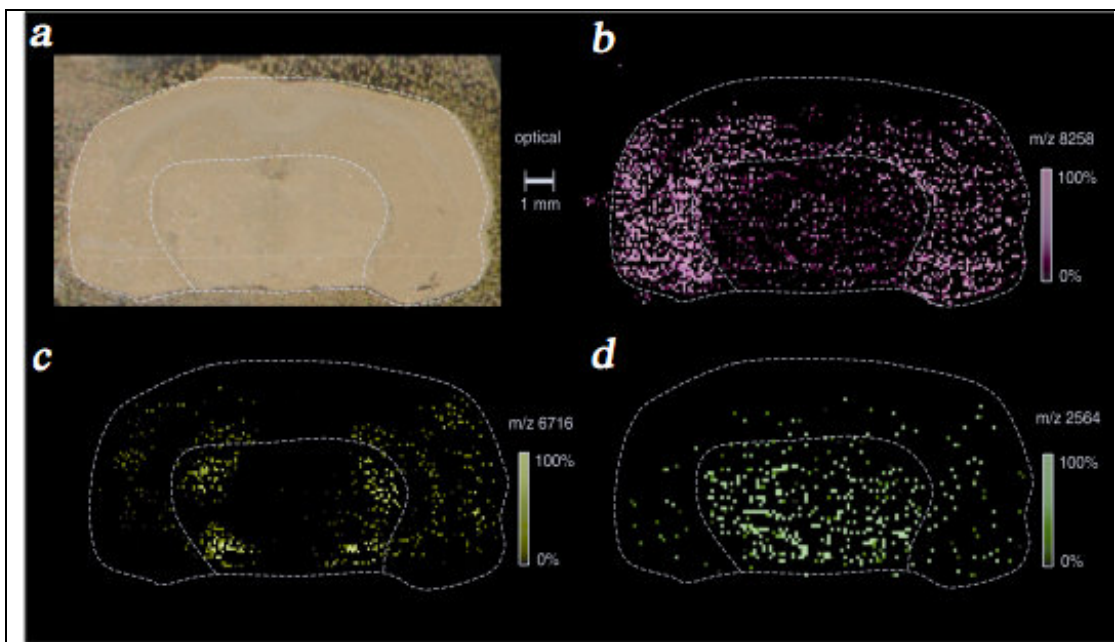


Figure 19. Mass spectrometric images of a mouse brain section.

a. Optical image of a frozen section mounted on a gold-coated plate. b.  $m/z$  8,258 in the regions of the cerebral cortex and the hippocampus. c.  $m/z$  6,716 in the regions of the substantia nigra and medial geniculate nucleus d.  $m/z$  2,564 in the midbrain (Stoeckli, 2001).

The molecular analysis and imaging of peptides and proteins in brain tumors is essential for locating specific proteins that are more highly expressed in tumors. It is thought that locating specific areas of the brain most affected by the tumor could be used

## SHOTGUN PROTEOMICS

in the intra-operative assessment of the surgical margins and/or clinical validation of diagnosis in patients.

### **Closing Thoughts**

The technology to support shotgun proteomics exists in a sliver of time. What was cutting edge a few years ago is now obsolete. Driven by vigorous research over decades, proteomic technology seems to be hitting its stride. Many second and third generation software products today represent years of hard work by pioneers in the field. These are people who take on the risk of developing new methods and technologies often under criticism from colleagues. Albert Einstein once said, “If we knew what it was we were doing, it would not be called research, would it?” Ultimately these pioneers steer science and every once in a while will discover breakthroughs that change our perception of reality.

## SHOTGUN PROTEOMICS

### References

- Ahn, N. G., Shabb, J. B., Old W. M., & Resing K. A. (2007). Achieving In-Depth Proteomics Profiling by Mass Spectrometry. *ACS Chemical Biology*, 2.1, 39-52.
- Ahram, M., Petricoin, E. F., (2008). Proteomics Discovery of Disease Biomarkers (2008). CiteSeerX Retrieved from:  
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.168.3744>
- Anderson, N. L., Anderson, N. G. (1998). Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, 19 (11), 1853-1861.
- Blackstock, W. P., Weir, M.P. (1999). Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* 17 (3): 121–7.
- Blast & FASTA, what's the difference? - Yahoo! Answers. (n.d.). Retrieved from  
<http://answers.yahoo.com/question/index?qid=20071003195335AAadk3z>
- BLAST: Basic Local Alignment Search Tool (2011). Retrieved from: Blast Program Selection Guide web site:  
[http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\\_TYPE=BlastDocs&DOC\\_TYPE=ProgSelectionGuide](http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=ProgSelectionGuide)
- Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase, 406, 89-112. Retrieved from  
[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=18287689](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18287689)
- Cox, J., Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26, 1367 – 1372.
- Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazan, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I. and Aebersold, R. (2010), A guided tour of the Trans-Proteomic Pipeline. *PROTEOMICS*, 10: 1150–1159. doi: 10.1002/pmic.200900375
- Donnelly CCBR. (n.d.). Retrieved from  
<http://www.torontodiscoverydistrict.ca/Page.asp?IdPage=5758&WebAddress=Discovery>
- Drews, J. (2000). Drug discovery: a historical perspective. *Science*,

## SHOTGUN PROTEOMICS

287:1960–4.

EMBL-EBI PRIDE. (2011). The PRIDE PRoteomics IDentifications database.  
Retrieved from: <http://www.ebi.ac.uk/pride>

Fenstermacher, D. A., Hemminger, B. M. (2002). Defining an Open Metadata Framework for Proteomics: The PROMIS Project. AMIA 2002 Annual Symposium Proceedings, 1093.

Flavin, M. (1981). Fundamental Concepts of Information Modeling. New York, NY: Yourdon

GO: Gene Ontology Project Website (2011). Retrieved from:  
<http://www.geneontology.org/>

GeneBio: Phenyx Product (2011). Retrieved from:  
<http://www.genebio.com/products/phenyx/>

Glossary of Bioinformatics Terms - Oak Ridge National Laboratory. (n.d.). Retrieved from <http://www.ornl.gov/hgmis/posters/chromosome/genejargon.shtml>

Kajdan, T., Cortes, T. H., Kuppanan, K., Young, S., (2008). Development of a Comprehensive Multidimensional Liquid Chromatography System with Tandem Mass Spectrometry Detection for Detailed Characterization of Recombinant Proteins. Journal of Chromatography, A 1189.1-2, 183-95.

KEGG: Kyoto Encyclopedia of Genes and Genomes. (nd). Retrieved from:  
<http://www.genome.jp/kegg/>

Lundgren, D. H., Han D.K., Eng J.K., (2005) Protein identification using TurboSEQUENT. Curr Protoc Bioinformatics. 13:Unit 13.3.

Mascot Integra (2011). Data Management for Proteomics. Retrieved from:  
<http://www.matrixscience.com/integra.html>

Matrix Science. (2006). [Thermo Fisher Scientific and Matrix Science]. Retrieved from [http://www.matrixscience.com/press\\_releases.html](http://www.matrixscience.com/press_releases.html).

Matrix Science. (2011). Mascot Search Overview. Retrieved from:  
<http://www.matrixscience.com/>

## SHOTGUN PROTEOMICS

- Mcgraw-hill: 2D Electrophoresis Animation. (2011). Retrieved from:  
<http://highered.mcgraw-hill.com/sites/dl/free/0072835125/126997/animation43.html>
- Minimum Information About a Proteomics Experiment (MIAPE):. (n.d.). Retrieved from  
[http://psidev.sourceforge.net/miape/MIAPE\\_Principles\\_5.1.doc](http://psidev.sourceforge.net/miape/MIAPE_Principles_5.1.doc)
- Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., (2003). A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Analytical Chemistry*, 75, 4646-4658
- Old, W. M., (2005). Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics. *Molecular & Cellular Proteomics*, 4.10, 1487-502.
- PRWeb. (2011). Global Proteomics Market to Reach \$6.1 Billion by 2015, According to New Report by Global Industry Analysts, Inc. [Press release]. Retrieved from:  
<http://blog.apastyle.org/apastyle/2010/09/how-to-cite-a-press-release-in-apastyle.html>
- Resing, K. A., Meyer-Arendt, K., Mendoza, A., Aveline-Wolf, L., Jonscher, K.G., Pierce, K. G., Old, W. M., Cheung, H., Russell, S., Wattawa, J. L., Goehle, G. R., Knight, R. D., & Ahn N. G., (2004). Improving Reproducibility and Sensitivity in Identifying Human Proteins by Shotgun Proteomics. *Analytical Chemistry*, 76.13, 3556-568.
- Ryan, T., (2002). Proteomics: Drug Target Discovery on an Industrial Scale. *Trends in Biotechnology* 20.12, S45-51.
- Scaffold. (2011). Visualize and Validate Complex MS/MS Proteomics Experiments, Retrieved from:  
[http://www.proteomesoftware.com/Proteome\\_software\\_prod\\_Scaffold.html](http://www.proteomesoftware.com/Proteome_software_prod_Scaffold.html)
- Sourceforge. (2011). mzXML Schema Documentation, Retrieved from:  
[http://sashimi.sourceforge.net/schema\\_revision/mzXML\\_2.1/Doc/mzXML\\_2.1\\_tutorial.pdf](http://sashimi.sourceforge.net/schema_revision/mzXML_2.1/Doc/mzXML_2.1_tutorial.pdf)
- Taylor C. F., Paton N. W., Lilley K. S., Binz, P., Julian, R. K., Jones A. R., Zhu, W., Apweiler, R., Aebersold R., Deutsch E. W., Dunn M. J., Heck, A. J. R., Leitner, A., ... Hermjakob H., (2007). The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology* 25, 887 - 893.

## SHOTGUN PROTEOMICS

The University of Aberdeen, (2006). 2D Gel Image of *Candida glabrata*., Reteived from:  
<http://www.abdn.ac.uk/cogeme/ap-Candida-glabrata-2d-gel.htm><http://www.abdn.ac.uk/cogeme/>

The University of New South Wales Staff Listing, (2010). [ School of Biotechnology and Biomolecular Science]. Retrieved from:  
<http://www.babs.unsw.edu.au/directory.php?personnelID=12>.

UniProt: The Universal Protein Resource (2011). Retrieved from:  
<http://www.uniprot.org/>

X! Tandem (2011). X! Search Engine Development . Retrieved from:  
<http://www.thegpm.org/tandem/>

Yen, C.Y., Meyer-Arendt, K., Eichelberger, B., Sun, S., Houel S., Old W. M., Knight R., Ahn, N. Gg., Hunter, L. E., & Resing K. A. (2009). A Simulated MS/MS Library for Spectrum-to-spectrum Searching in Large Scale Identification of Proteins. *Molecular & Cellular Proteomics*, 8.4, 857-69.