FACTORS THAT AFFECT PROTEIN IDENTIFICATION BY MASS SPECTROMETRY

by

HAOFEI TIFFANY WANG

(Under the Direction of Ron Orlando)

ABSTRACT

Mass spectrometry combined with database search utilities is a valuable protein identification tool. The success of database mining is dependent upon mass accuracy, protein purity, peptide yield, and the genomic complexity of the target organism. Three proteins were selected to investigate the dynamic interaction of these variables and their effect on database mining. With variables of interest controlled, simulated spectra were searched using two searching programs. Results suggest that high mass accuracy improves database searching confidence in the protein identification. With the addition of random noise peaks, some searching programs require a significant increase in the number of peptide ions and the mass accuracy required. Placing limits on database searches usually improves searching efficiency by allowing fewer peptide ions for a successful identification.

INDEX WORDS: Mass spectrometry, MALDI-TOF, Tandem mass spectrometry, Proteomics, Peptide mapping, Database search, and Mass accuracy

FACTORS THAT AFFECT PROTEIN IDENTIFICATION BY MASS SPECTROMETRY

by

HAOFEI TIFFANY WANG

B.S., The Capital University of Medical Sciences, China, 1999

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment

of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2003

FACTORS THAT AFFECT PROTEIN IDENTIFICATION BY MASS SPECTROMETRY

by

HAOFEI TIFFANY WANG

Major Professor: Ron Orlando

Committee: Jaxk Reeves
Claiborne Glover

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2003

DEDICATION

This thesis is dedicated to my family in China, to my parents, my grandparents, and my supportive friends. I never feel alone with all the love from my excellent family and friends. I love you all.

ACKNOWLEDGEMENTS

This work cannot be done without the direction from my director, Dr. Ron Orlando. Thank you for showing me the direction in my research, for giving me room to make mistakes, for allowing me to make decisions or even change mind sometimes.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Introduction to Mass Spectrometry-Assisted Peptide Mapping

Mass spectrometry (MS) is now a widely applied technique in a number of fields, including biochemistry [1], pharmacology [2], and the proteomics/genomics field [3]. In proteomics research, MS is used to determine the identity of proteins, to analyze the primary structure at the protein level and determine post-translational modifications [4,5].

In protein identification practice, MS technique combining with protein database search is considered as a rapid, sensitive and automation-possible tool [4-6]. The idea is that an unknown protein is digested using a specific proteolytic enzyme, such as trypsin, to generate a set of peptides. Peptide masses are accurately measured and recorded during MS process, such as matrix-assisted laser desorption/ionization (MALDI)-time of flight (TOF) MS [7,8]. The pattern of masses generated from MS process is reported in the form of a mass spectrum. The peaks in a spectrum provides a specific "fingerprint" of the precursor protein, which are called peptide mass fingerprint (PMF) for the precursor unknown protein [9,10]. The PMF then undergoes computer-assisted proteome database searching, i.e., the experimentally obtained PMF is compared with every theoretical PMFs from individual proteins in a database [11]. The protein or proteins matching best are identified.

In tandem mass spectrometry, a sample is weighed in the first mass spectrometer; precursor ions are broken into pieces in the second mass spectrometer. A piece or pieces (fragments) are weighed in the second mass spectrometer. Tandem mass spectrometry (MS/MS) [12] assisted protein identification is getting more and more attention nowadays in that it allows

identifying an unknown protein from a single precursor fragment (peptide) ion. It is a powerful and convincing way for protein identification because the sequence information can be achieved as well.

Currently, there are several on-line searching software packages available to which users may submit their peptide/fragment queries [13-15]. Appendix A gives a list of the commonly used ones with their on-line web site. Following certain algorithms, these searching utilities will evaluate the matches between the submitted PMF with every protein query in an assigned proteome database. The evaluations will be quantified and ranked based on the confidence of non-randomness in matching [16,17]; usually an associated probability-based score will give users an idea of how confidently the protein is identified.

1.2 Factors that Affect Peptide Mapping

In the MS process to generate a spectrum for unknown proteins, there are a series of factors that could affect protein identification outcome. The most widely noticed of these include:

(1) The number of peptide ions / fragment ions constituting the PMF [18,19],

(2) The mass accuracy of the peptide ions / fragment ions [20-23],

(3) The relative number of irrelevant and potentially interfering mass submitted along with peptide ions / fragment ions [20, 23].

(4) Database Restriction [13,14].

Once a peptide ions pool is generated, i.e. the peptide mass fingerprint (PMF) is constructed from a real spectrum; it will be searched against a specific database. We noticed that protein identification results vary with the assigning of different databases. Limiting searches to a certain target database affects outcome confidence, searching efficiency, or both. We consider

database restrictions as the fourth factor of our interest. Most on-line searching programs allow users to limit the scope of their searches in certain ways. Usually, those user-controllable specifications include limiting the database to a specific range, of molecular weight (MW), a specific range of isoelectric point (pI), and a specific species. In all, the potentially affecting factors involved in database searching include:

(1) Limiting the target database by the protein's organism resources (Species)

(2) Limiting the target database by the protein's molecular weight (MW)

(3) Limiting the target database by the protein's isoelectric point (pI)

1.3 Goals

The potential influence of the factors discussed has been noted in the field of proteomics research. Nevertheless, the understanding of the way these factors affect peptide mapping is still insufficient. In this thesis, we report a systematic and quantitative evaluation on the effect of these variables. A further goal is to discuss how to optimize protein identification on the basis of the knowledge of what role they play in identifying a protein and how significant the role is.

CHAPTER 2

METHODS

2.0 Experimental Design

Factors that influence the protein identification results include mass accuracy, number of peptide ions submitted, the impurity in a peptide mass fingerprint (PMF) and the database restrictions as explained in Section 1.2. To analyze the effects from these factors and their dynamic interaction, a cross experimental design was applied.

There are four main factors in our design, they are:

(1) Mass Accuracy (7 levels) [1, 10, 100, 200, 400, 800, 1000] ppm

(2) Number of Real Peaks (9 levels) [1, 2, 3, 4, 5, 6, 7, 8, 9]

(3) Contamination Proportion (4 levels) [0%, 50%, 67%, 75%]

(4) Database Search Restriction ($2^3 = 8$ levels for all possible combinations of presence/absence of restrictions on {MW, pI, species})

This is a completely crossed design, in that all 7*9*4*8 = 2016 combinations of factors are simulated for each of 3 sampled proteins, with each method's results being evaluated by both Mascot™ and ProFound™ (as will be discussed in Section 2.8). Two responses will be observed for each experiment: (1) whether or not the "correct" protein is identified as the best choice, and (2) the probability score reported for the true protein.

2.1 Materials

Chemicals and solvents used in this study were of analytical grade [24]. Trypsin for in-gel tryptic digestions was sequencing grade-modified trypsin from Promega®. One protein-

extract from *Sus scrofa domestica* (Domestic Pig*)* and one protein-extract from *Arabidopsis thaliana* are selected for parallel experiments [24,25].

2.2 Sample Preparations and MS Processes

Each protein extract was subjected to 2-D gel electrophoresis for separation and purification. After visualization, the bands containing the protein were excised from SDS-polyacrylamide gels and subjected to digestion with trypsin [26]. The information of MW and pI for each spot was recorded before cutting to be used as a future reference when setting database search parameters. Digested peptides were extracted and submitted to MALDI-TOF MS [24, 27, 28] and LC-MS/MS processes [12]. Real spectra were obtained for each protein spot. Figure 2.1 shows a MALDI-TOF spectrum for Protein A.

2.3 Selecting Appropriate Spectra to Analyze the Factors of Interest

To choose spectra suitable for analyzing the effect of factors of interest, each spectrum was processed and searched using Masslynx 3.4 featuring BioLynx™. BioLynx was used to determine the ID of the precursor protein from a real spectrum and to differentiate tryptic peptide ions from noise peaks. Three steps were followed to choose suitable spectra:

(1) Each spectrum was analyzed by BioLynx functioning MaxEnt3™ process with default parameters set-up.

(2) The spectrum processed with MaxEnt3 was searched using database search function of BioLynx to determine the protein's identity (ID);

Three proteins successfully identified using BioLynx were selected for parallel experiments. The three selected proteins and their IDs in NCBI nr database are listed in Table 2.1.

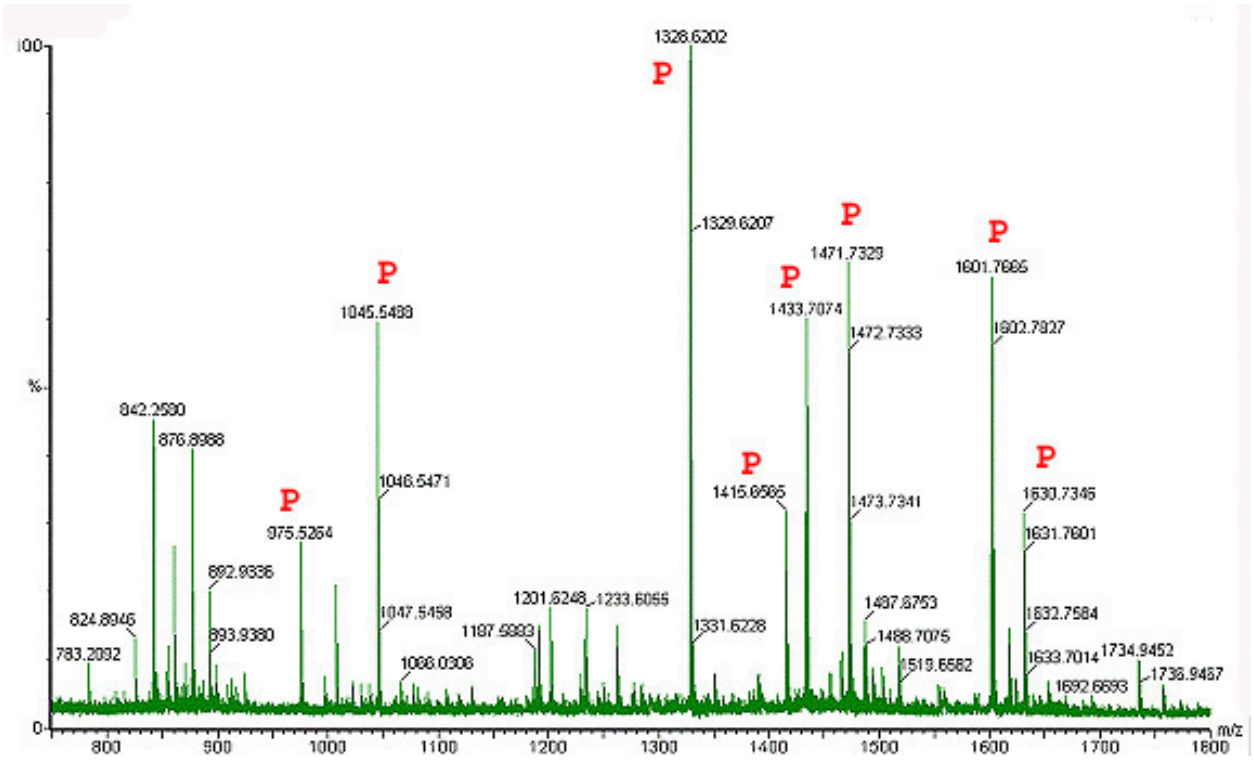

Figure 2.1 MALDI-TOF Spectrum for Protein A

Figure 2.1 shows a MALDI-TOF spectrum for Apolipoprotein AI, the red “P” denotes peptide ions, peaks without “P” are considered as noise peaks.

Table 2.1 Proteins Sampled to Investigate the Factors that Affect Peptide Mapping Result

PROTEIN	ID	ORGANISM	SEQUENCE LENGTH (AA)	MW (KDA)	PI
A	Apolipoprotein AI	*Sus scrofa domestica* (Domestic Pig)	265	30.31	5.4
B	ATPase beta subunit	*Arabidopsis thaliana*	498	53.92	5.4
C	Fructose bisphosphate-aldolase like protein	*Arabidopsis thaliana*	358	38.52	6.1

2.4 Generating Peptide/Noise Pool

A complete sequence of selected proteins was retrieved from NCBI nr database and was then subjected to simulated trypsin digestion using the protein/peptide editor function of BioLynx. Tryptic fragments with their sequence and mass information were saved for each sampled protein as an ideal peptide file. Figure 2.2 shows the tryptic files for protein A. The masses of peaks shown in the real spectrum (Figure 2.1) are compared with this theoretical tryptic mass to determine the real peptide ion peaks from noise peaks. In Figure 2.1, the peaks denoted by a red “P” represent those which matched a theoretical peptide in Figure 2.2, thus is considered as an peptide ion peak. Those that failed to match any tryptic fragment in figure 2.2 are considered as noise peaks. Both peptide peaks and noise peaks were ranked by intensity to construct peptide ion pool and noise ion pool.

Trypsin:K-\P /R-\P

Frag#	Res#	Sequence	Theor(Bo)	[M+H]
T33	313-317	(R)ITSTK(K)	548.32	549.32
T42	419-423	(R)LTVAR(A)	558.35	559.36
T4-5	18-22	(K)KNLGR(I)	586.36	587.36
T39	393-397	(K)QTLQR(Y)	644.36	645.37
T33-34	313-318	(R)ITSTKK(G)	676.41	677.42
T43-45	424-429	(R)ARKIER(F)	771.47	772.48
T15	128-134	(R)TTSPIHK(S)	782.43	783.44
T42-43	419-425	(R)LTVARAR(K)	785.49	786.49
T3-5	16-22	(R)EKKNLGR(I)	843.49	844.50
T42-44	419-426	(R)LTVARARK(I)	913.58	914.59
T39-40	393-399	(K)QTLQRYK(E)	935.52	936.53
T21	168-178	(K)IGLFGGAGVGK(T)	974.55	975.56
T17	146-154	(K)LSIFETGIK(V)	1006.57	1007.58
T47	448-456	(K)YVGLAETIR(G)	1020.56	1021.57
T49	487-495	(K)ATNLEMESK(L)	1021.48	1022.48
T18	155-163	(K)VVDLLAPYR(R)	1044.60	1045.60
T11	76-86	(R)AVAMSATEGLK(R)	1076.55	1077.56

Figure 2.2 Theoretical Tryptic Files for Protein A

Protein A was theoretically digested by trypsin using BioLynx. The file shows, from left to right, the fragment series number, amino acid location, sequence, theoretical mass and [M+H] mass.

2.5 Constructing MALDI-TOF Mass Data Files

Peptide and noise pools for each selected protein were optimized to construct searchable mass data files. Data files were designed to meet the needs of testing for different potentially affecting factors.

(1) Constructing basic mass data file

For each selected protein, the 9 peptide ions with the highest intensities were chosen to construct a basic mass data file. The basic mass data file contains the real masses of these peptide ions as determined by MALDI-TOF and the simulated masses under each mass accuracy level

(Section 2.0). For each mass accuracy level, the exact peptide ion mass was randomized to simulate an appropriate level of mass error. Simulated peptide pools with mass accuracy from 1 ppm to 1000 ppm were created for each protein. Table 2.2 shows a basic mass data file for protein A. Basic mass data files for proteins B and C were constructed in the same way for parallel experiments.

(2) Data files for the analysis of the effect of mass accuracy on searching confidence

Searching confidence is defined as the confidence level that a match (between a queried mass data file and an entry protein in a database) is not random. The basic mass data file was used to analyze the effect of mass accuracy on protein identification confidence. For each mass accuracy category, we submitted a fixed number of 6 peptide ions with the highest intensity. Six peptide ions are reasonable and representative for most peptide ion pools. Thus, with the number of submitted peptide ions controlled, the searching confidence's variation, if there is any, is expected to come from the changing of mass accuracy assignment solely.

Table 2.2 Basic Mass Data File Constructed for Protein A

	REAL MASS	RANDOMIZED MASSES TO SIMULATE SPECIFIC MASS ACCURACY						
		1ppm	10ppm	50ppm	100ppm	200ppm	400ppm	800ppm
1	991.50	991.50	991.51	991.49	991.47	991.44	991.53	991.68
2	1404.71	1404.71	1404.70	1404.73	1404.74	1404.74	1404.48	1404.78
3	1797.90	1797.90	1797.90	1797.88	1797.84	1797.95	1797.94	1797.69
4	1510.86	1510.86	1510.85	1510.89	1510.85	1510.78	1510.77	1510.90
5	1012.58	1012.58	1012.58	1012.55	1012.56	1012.67	1012.59	1012.68
6	1589.80	1589.80	1589.80	1589.80	1589.82	1589.81	1589.95	1589.68
7	1305.64	1305.64	1305.65	1305.65	1305.65	1305.74	1305.76	1305.66
8	1271.54	1271.54	1271.53	1271.56	1271.56	1271.42	1271.54	1271.54
9	1851.96	1851.91	1851.91	1851.96	1851.89	1852.03	1851.89	1851.84

(3) Data files for the analysis of effect of mass accuracy on searching efficiency

Searching efficiency is defined here as the minimum number of peptide ions required to ensure a successful protein identification with 95% confidence. To determine this minimum number, for a mass accuracy level, a full-size basic mass data file containing 9 peptide ions was submitted, with the initial confidence score returned. Deleting the peptide ion with the lowest intensity from the basic mass data file and resubmitted the remaining peaks, a new confidence score was returned and recorded. Repeat this delete-and-resubmit loop until the searching utility failed to return a confidence score satisfying a 95% confidence for successful protein identification. If N is defined as the number of peptide ions left in the submitted mass data file

when the search fails, then N+1 is the minimum number of peptide ions required for successful protein identification at the 95% confidence level. Under each mass accuracy category, the result of N+1 was recorded and analyzed.

(3) Constructing data files for analysis of effects of impurity on protein identification

To analyze the effect of potential impurity (noise peaks submitted along with peptide peaks) in the peptide pool, for each protein, peaks in the noise pool (Section 2.4) were randomized to accommodate mass accuracy levels from 1 ppm to 1000 ppm. For each mass accuracy level, a certain number of noise peaks were introduced into the basic mass data file. These noise peaks, combined with peptide peaks, generated an "impure" mass data file containing 50%, 67%, and 75%, respectively, of impurity. Here the percentage refers to the proportion of noise peaks out of all peaks submitted. For example, "50%" indicates that for all the masses submitted, half of them are peptide ions and the other halves are impurities. Under a various mass accuracy levels, peptide peaks combined with various amount of noise peaks were submitted. The searching efficiencies under each of these conditions were recorded and compared with those obtained under the basic data conditions (i.e. no impurity introduced).

(5) Data file used to analyze the effect of factors involved in database searching

Factors involved in database searching include database restriction on the protein's molecular weight, pI, and /or the species of the organism that produces the protein. We use the basic mass data file to examine the effects of these factors on the protein identification results at each mass accuracy level.

2.6 Constructing MS/MS mass data files

LC-MS/MS spectrum for protein C in Table 2.1was chose to analyze the factors in tandem mass spectrometry. Figure 2.2 shows its real MS/MS spectrum.

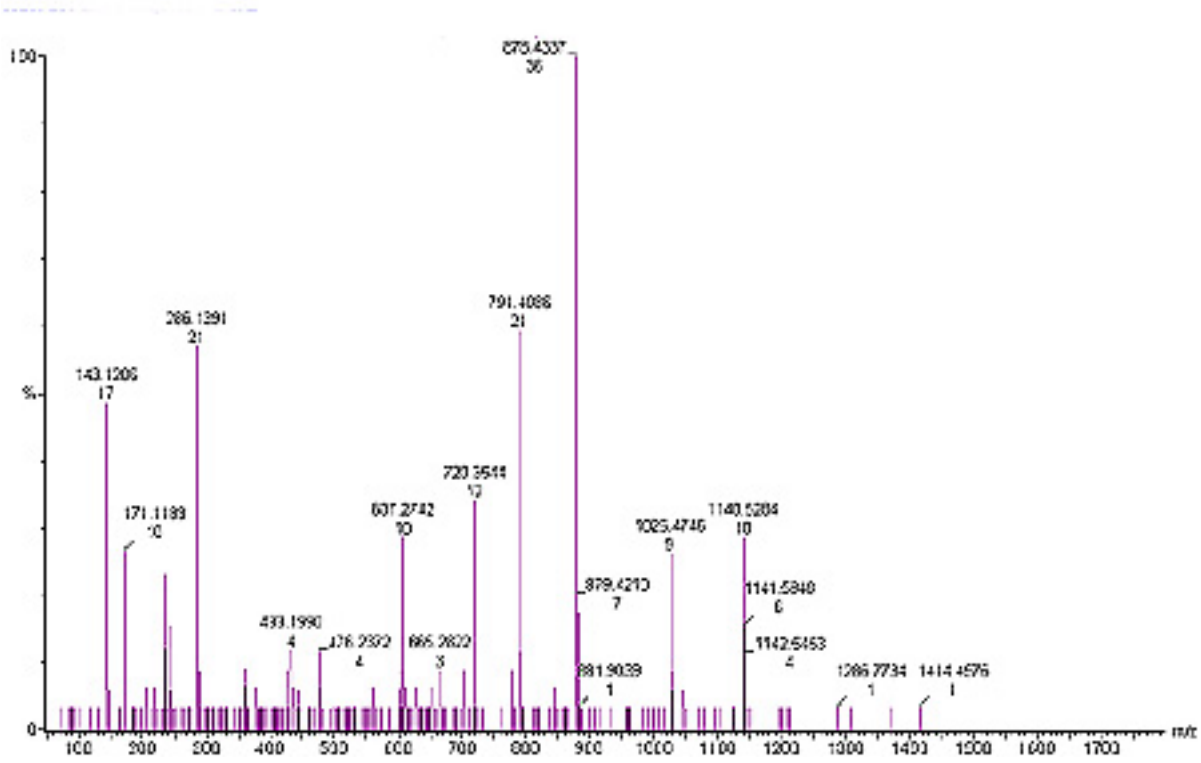

Figure 2.3 MS/MS Spectrum for Protein C

The variables of our interest include precursor peptide ion's mass accuracy (in ppm), MS/MS fragment mass tolerant (in Dalton), and the minimum number of fragment ions submitted to a search utility. To analyze these variables, searchable mass data files were constructed following the steps below:

(1) Simulate precursor mass accuracy level and fragment mass accuracy level

The masses of precursor ions were randomized to accommodate the mass accuracy levels of 10 ppm and 100 ppm. Under each precursor mass accuracy level, i.e. [10 ppm and 100ppm], the fragment ion's mass was randomized to simulate eight fragment mass tolerant levels: [0.01 Da, 0.02 Da, 0.05Da, 0.1Da, 0.5Da, 1Da, 1.5Da, 2Da]. Same as the way we constructed the basic mass data files for the analysis on factors in peptide fingerprint mapping (Section 2.5), here we

included nine fragment ions with highest intensity in the basic mass data files for MS/MS analysis. The ion peaks in a data file were ranked based on the intensity, with top one representing the most abundant ion peak. Considering the interaction between peptide mass accuracy (2 levels: [10 ppm, 100 ppm]) and fragment ion mass accuracy (8 levels: [0.01 Da, 0.02 Da, 0.05Da, 0.1Da, 0.5Da, 1Da, 1.5Da, 2Da]), a complete cross experimental design was applied with a total of 2*8=16 crossed mass accuracy levels analyzed and compared.

(2) Mass Data files used to analyze the effect from the number of fragment

A full-size basic mass data file for MS/MS containing 9 fragment ion peaks was submitted under 16 crossed mass accuracy levels. If with the initial confidence score is higher than the homology confidence threshold given by Mascot (http://www.matrixscience.com/home.html), we deleted the fragment ion peak with the lowest intensity from the basic mass data file and resubmitted the remaining peaks. We repeated this delete-and-resubmit loop until the searching utility failed to identify a protein homology with 95% confidence. If we define N as the number of peptide ion s left in the submitted mass, then N+1 is the minimum number of fragment ions required for a successful MS/MS fragment mapping, which will be used as an index to the searching efficiency under certain conditions.

2.7 Searching Scheme

Two on-line searching utilities, Mascot [31] (http://www.matrixscience.com/search_form_select.html) (Matrix Science Ltd., London) and ProFound [29, 30] (http:// http://prowl.rockefeller.edu) (Rockefeller University, Genomic Solutions) were chosen to search MALDI-TOF data to analyze the factors in peptide mapping. For parallel experiment, simulated mass data files for each sampled protein (Section 2.5) were submitted on-line to both Mascot and ProFound. Results from the two search utilities were

analyzed independently and compared in Chapter 3 to generalize the influences from the factors of interest.

To analyze the factors involved in tandem mass spectrometry-assisted fragment mapping, Mascot is further used to search MS/MS mass data files constructed for Protein C (Section 2.6). The search results will be discussed in Chapter 4.

CHAPTER 3

FACTORS IN PEPTIDE MASS FINGERPRINTING USING MALDI-TOF MS

Mass data files were constructed as described in Chapter 2. They were then searched to analyze the factors of interest in protein identification using peptide mass fingerprint technique. With the factors in question controlled, searching confidence (the confidence gained that a match is not random) and/or searching efficiency (the minimum number of peptide ions required for a successful search with 95% confidence) are observed and discussed as the two response variables affected by influencing factors.

3.1 Effect of Mass Accuracy and the Number of Peptide Ions Submitted

(1) The effect of mass accuracy on searching confidence

A basic mass data files containing 6 peptide ions with mass accuracies of 1 ppm, 10 ppm, 50 ppm, 100 ppm, 200 ppm, 400ppm and 800 ppm were searched to investigate the effect of mass accuracy on searching confidence. The confidence scores by Mascot and ProFound are plotted in Figure 3.1 (A-C) and Figure 3.3 (A-C) respectively.

A Mascot score is defined as $M= -10*Log_{10} (P)$ (http://www.matrixscience.com/help/scoring_help.html), where P evaluates the probability that the observed match is a random event. Higher confidence score suggests a less random match. A red threshold line indicates 95% confidence that the observed match is not a random event. Mascot searching results are shown in Figure 3.1(A-C) for the three sampled proteins, they are Apolipoprotein AI from domestic pig, ATPase beta subunit from *Arabidopsis thaliana*, and Fructose bisphosphate-aldolase like protein from *Arabidopsis thaliana* respectively.

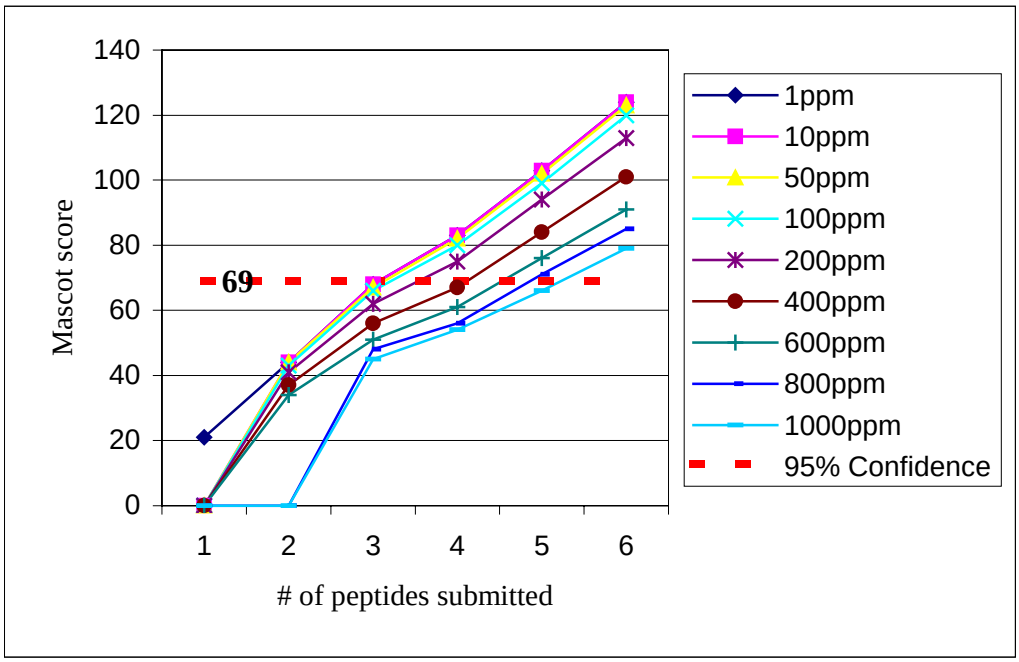

Figure 3.1 (A). Mascot Scores: Mass Accuracy vs. Search Confidence (Protein A)

The basic mass data file for Apolipoprotein AI from domestic pig was searched using Mascot under different mass accuracy levels. The red line indicates a 95% confidence threshold considering the size of current database. (Data was recorded in 2000). The search scores higher than 69 suggest a successful search.

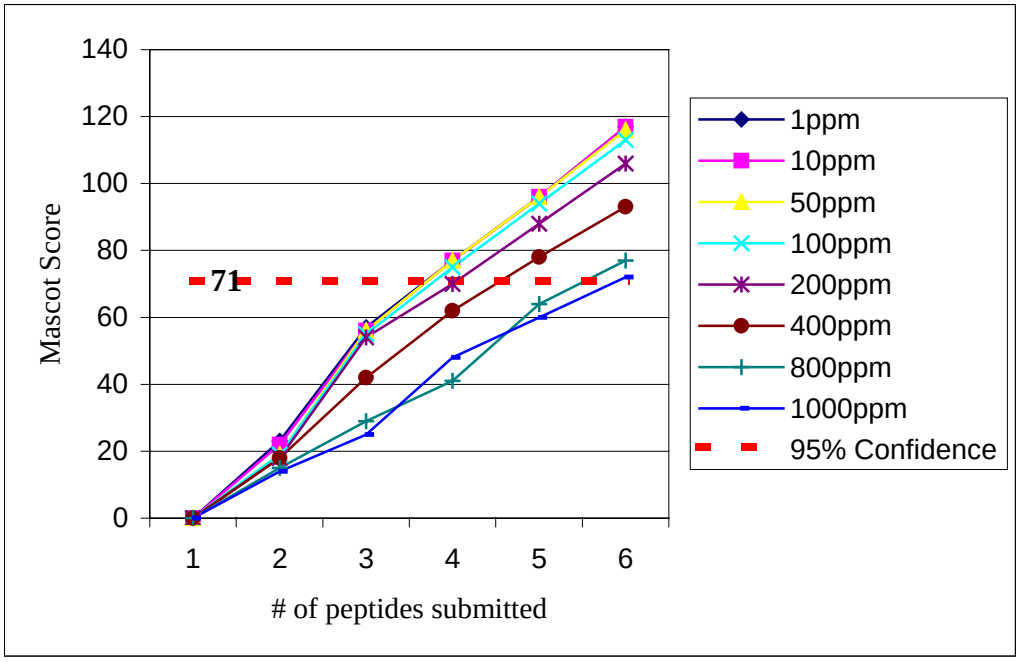

Figure 3.1 (B). Mascot Scores: Mass Accuracy vs. Search Confidence (Protein B)

The basic mass data file for ATPase beta subunit from *Arabidopsis thaliana* was searched using Mascot under different mass accuracy levels. The red line indicates a 95% confidence threshold considering the size of current database. (Data was recorded in 2001). The search scores higher than 71 suggest a successful search.

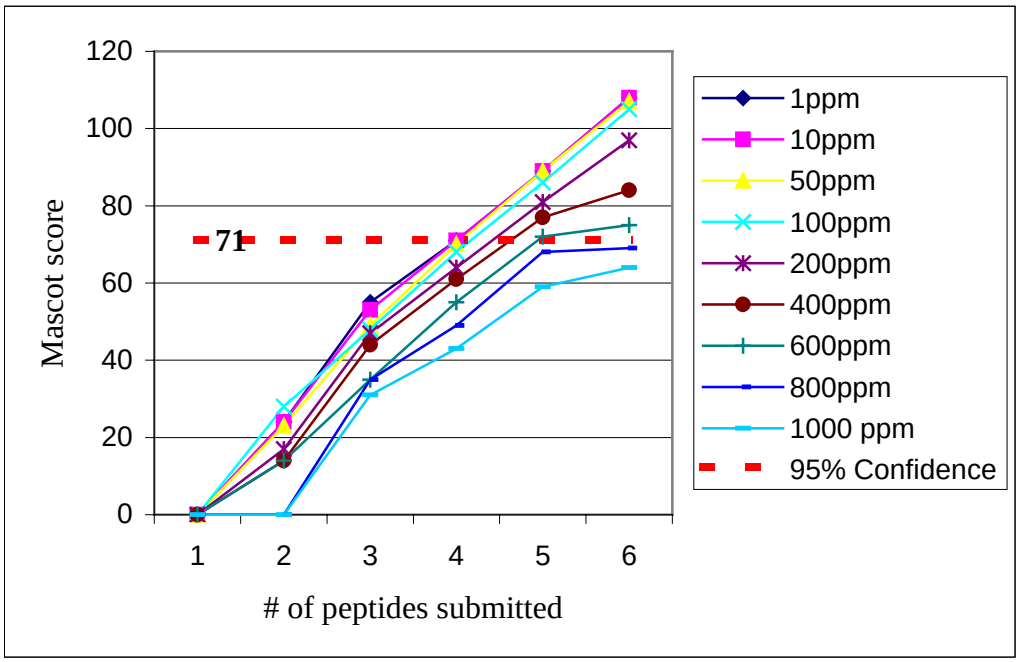

Figure 3.1 (C). Mascot Scores: Mass Accuracy vs. Search Confidence (Protein C)

The basic mass data file for Protein C (Fructose bisphosphate-aldolase like protein) was searched using Mascot under different mass accuracy levels. The red line indicates a 95% confidence threshold considering the size of current database. (Data was recorded in 2001). The search scores higher than 71 suggest a successful search.

Mascot search results show that for each protein, the confidence scores decreased gradually when mass accuracy was lowered (i.e. when ppm values is increased). The decreasing trends for the 3 proteins were very similar. With 6 peptide ions submitted, highest confidence scores of 124, 117, 108 were returned for protein A, B, and C, respectively, when the best mass accuracy data (1 ppm) were submitted. Figure 3.2 shows the decreasing trend of confidence score when relaxing mass accuracy from 1 ppm to 1000 ppm when six peptide ions submitted. Since six peptide ions represent a reasonable and also, favorable peptide coverage, the results

will show the effect from mass accuracy without intervention from other factors such as peptide coverage. It suggests that compromising mass accuracy from 1 ppm to 100 ppm has negligible effect on searching confidence. An apparent decrease of the confidence score is shown when an average mass accuracy (>200 ppm) is lowered to a poor mass accuracy (800 ppm)

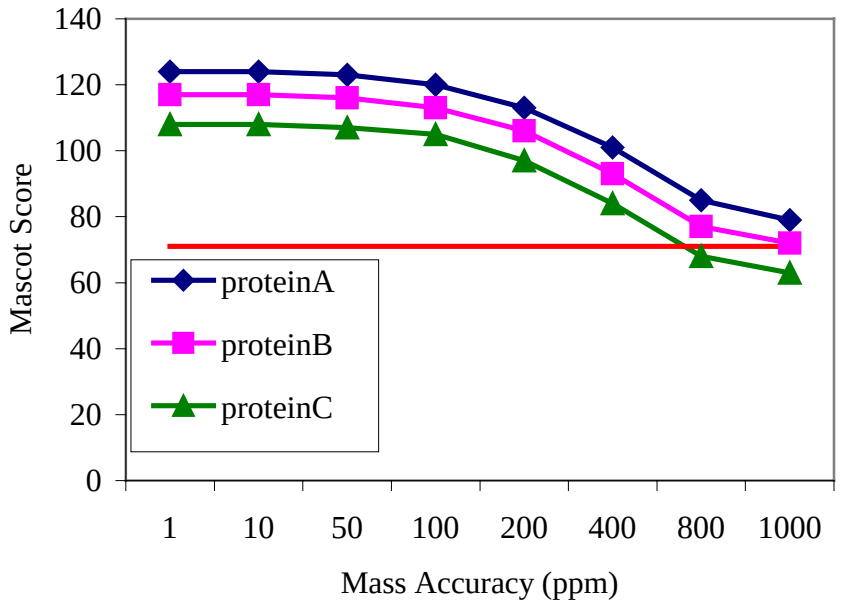

Figure 3.2. Decreasing Trend of Confidence with the Relaxation of Mass Accuracy (Mascot)

The searching results when using ProFound (http://65.219.84.5/service/prowl/profound/profound_E_adv.html) are shown in Figure 3.3(A-C). To verify the relationship between mass accuracy and protein identification confidence observed from Mascot search results, the same peptide mass data files were submitted to ProFound under the same mass accuracy categories for Apolipoprotein AI (Figure 3.3(A)), ATPase beta subunit (Figure 3.3(B)), and Fructose bisphosphate-aldolase like protein (Figure 3.3(C)).

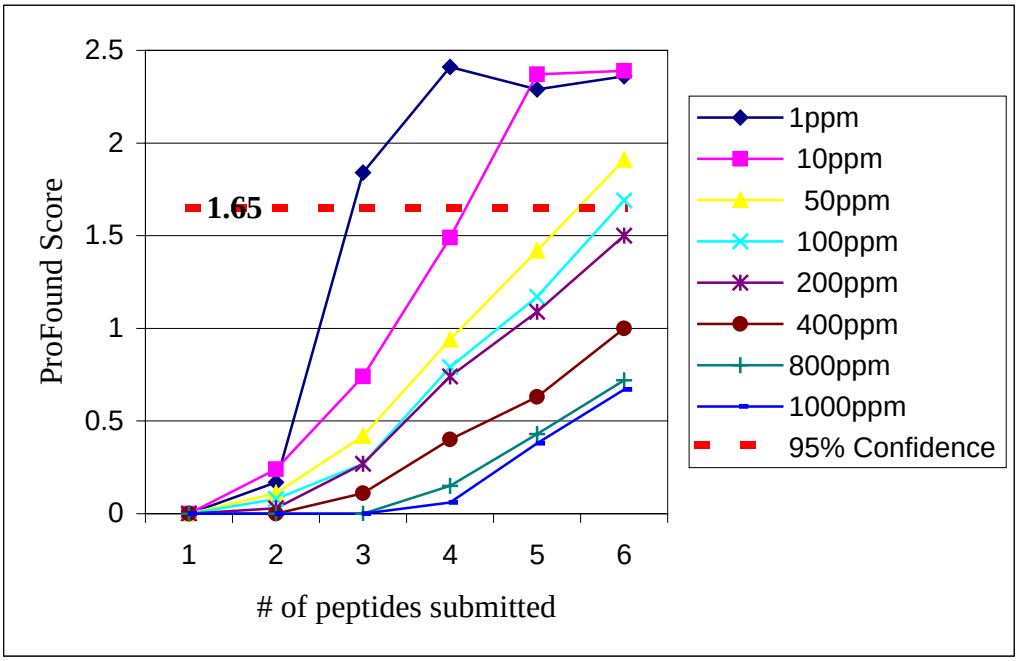

Figure 3.3 (A). ProFound Z-Scores: Mass Accuracy vs. Search Confidence (Protein A)

The basic mass data file for Protein A (Apolipoprotein AI) was searched using ProFound under different mass accuracy levels. The red line indicates a 95% confidence threshold. (Data was recorded in 2001). The search scores higher than 1.65 suggest a successful search.

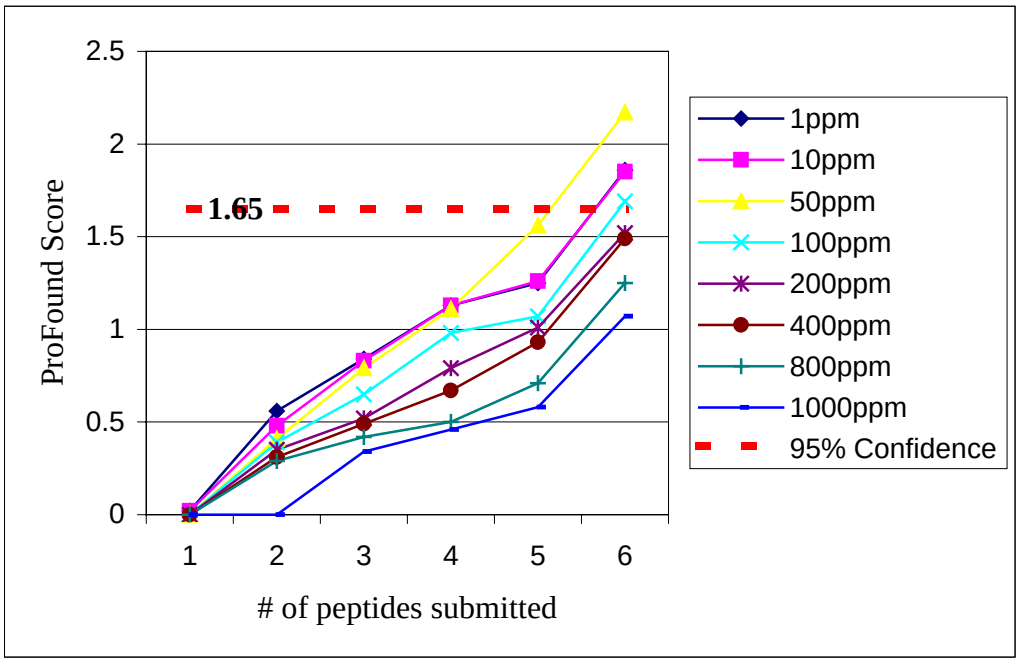

Figure 3.3 (B). ProFound Z-Scores: Mass Accuracy vs. Search Confidence (Protein B)

The basic mass data file for Protein B (Apolipoprotein AI) was searched using ProFound under different mass accuracy levels. The red line indicates a 95% confidence threshold. (Data was recorded in 2001). The search scores higher than 1.65 suggest a successful search.

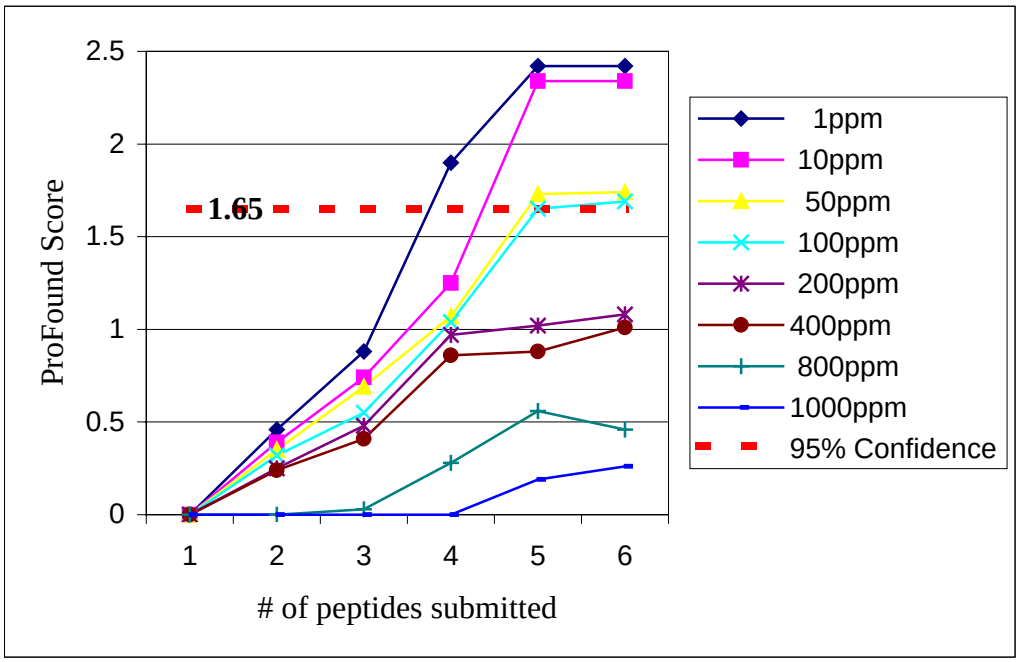

Figure 3.3 (C). ProFound Z-Scores: Mass Accuracy vs. Search Confidence (Protein C)

The basic mass data file for Protein C (Fructose bisphosphate-aldolase like protein) was searched using ProFound under different mass accuracy levels. The red line indicates a 95% confidence threshold. (Data was recorded in 2001). The search scores higher than 1.65 suggest a successful search.

Similar to the Mascot search results, a decreasing trend of ProFound confidence score, Z score [29, 30], was observed when mass accuracy was reduced in Figure 3.3 (A-C). Figure 3.4 compares the decreasing trends for the three sampled proteins. Unlike the Mascot results, which showed similar decreasing trends for the three sampled proteins, the three decreasing curves generated with ProFound suggested different decreasing trends. Rapid decreases in searching confidence in response to the relaxation of mass accuracy is observed for protein C, which had the highest MW among the three tested proteins. Results for protein C also suggested an apparent decrease of confidence even under fairly good mass accuracy of 100 ppm. In contrast, the

searching result for protein A, which had the lowest MW among the tested proteins, was not affected by allowing more mass error in the same range.

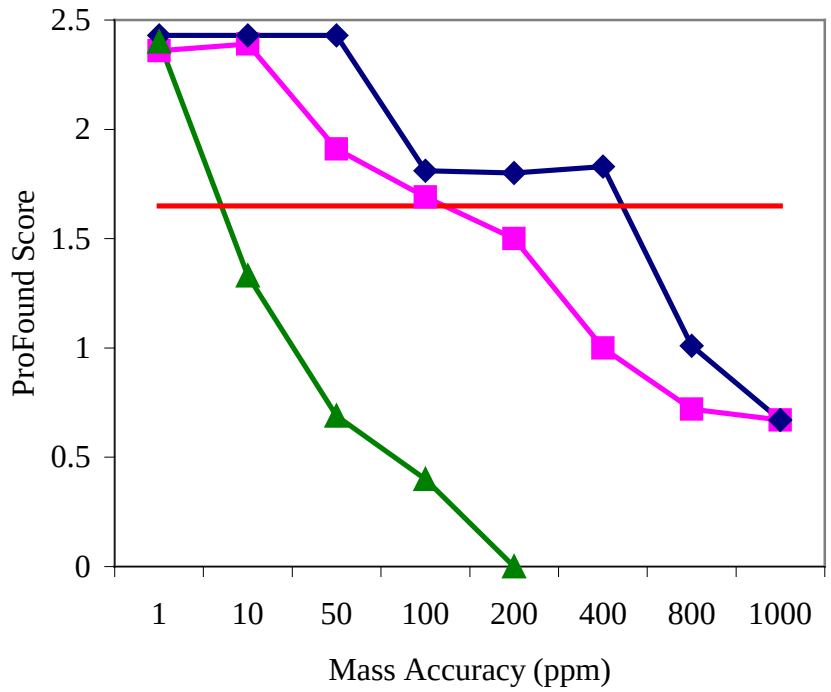

Figure 3.4. Decreasing Trend of Confidence with the Relaxation of Mass Accuracy (ProFound)

Searching results for protein A, protein B, and protein C with ProFound are plotted using the colored-lines consistent with Figure 3.3 to compare the decreasing trends shown by different searching programs.

(2) Effect of mass accuracy on searching efficiency

To analyze the effect of mass accuracy on searching efficiency, the minimum number of peptide ions that ensure successful protein identification with confidence level not less than 95 %

is used as an index of searching efficiency. To compare the searching efficiency under each mass accuracy level, the minimum number of peptide ions are listed and compared in Table 3.1.

Table 3.1. Minimum Numbers of Peptide Ions for a Successful Search

Mass Accuracy	Mascot Result			ProFound Result		
	Protein A	Protein B	Protein C	Protein A	Protein B	Protein C
1 ppm	4	4	4	3	4	6
10 ppm	4	4	4	5	5	6
50 ppm	4	5	4	6	5	6
100 ppm	4	5	4	6	5	6
200 ppm	4	5	5	7	10	7
400 ppm	5	5	5	8	10	8
800 ppm	5	7	6	9	>10	8
1000 ppm	6	8	6	>10	>10	9

Results show the minimum number of peptide ions required to identify the target protein with a 95% confidence level under different mass accuracy level.

Although the minimum number of peptide ions required for protein identification varies for the three sampled proteins and when different searching programs were used, in general this number increased when mass accuracy was compromised. Comparing searching results within one searching program (using the same algorithm) from Table 3.1, Figure 3.1(A-C), and Figure 3.3 (A-C), we observed that for each protein: (1) the least number of peptide ions for a successful peptide mapping was always associated with the best mass accuracy (1 ppm in our experiment), (2) lowering mass accuracy increased the minimum peptide requirement, thus lowering the

searching efficiency, (3) it was not uncommon that low mass accuracy data resulted in failure of protein identification. For instance, in Table 3.1, when the mass accuracy was lowered to 1000 ppm, ProFound failed to identify protein C, even with as many as 10 peptide ions submitted.

3.2 Effect of Impurity in PMF

To estimate the effects of impurity on protein identification, 1/2, 2/3, and 3/4 noise peaks were introduced into the peptide ion pool to generate “impure” data files (Chapter 2). Peptide ions along with certain level of extraneous noise were searched at a mass accuracy of 1 ppm, 10 ppm, 100 ppm, and 400 ppm.

(1) The result of Mascot searching:

Figure 3.5 (A-C) show the minimum number of peptide ions required to identify the target protein using data with no noise, 1/2 noise, 2/3 noise, and 3/4 noise under different mass accuracies. Searches were done using Mascot.

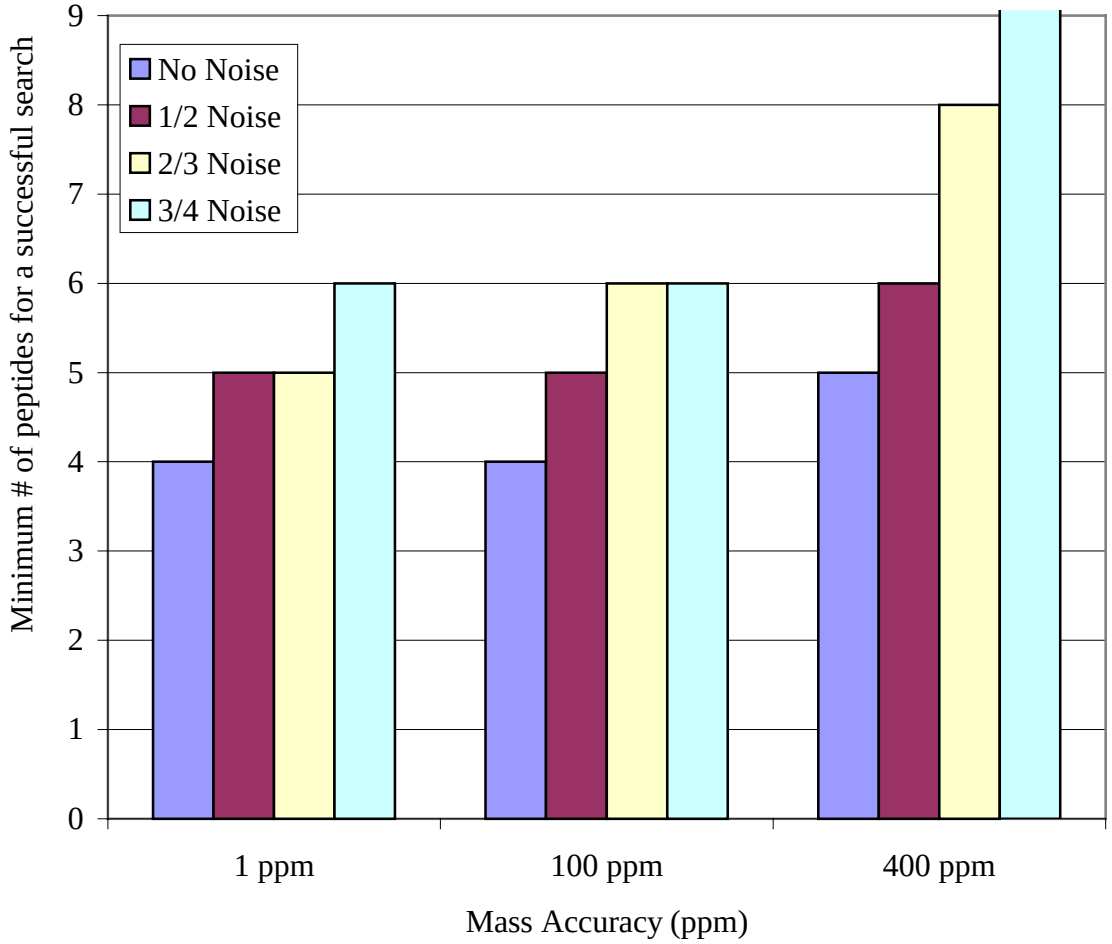

Figure 3.5 (A). Mascot result: The Effect of Impurity on Searching Efficiency (Protein A)

Results show the searching efficiency, i.e. the minimum number of peptide ions ensuring a successful search for apolipoprotein AI. Searches were under a mass accuracy level of 1 ppm, 100 ppm, and 400 ppm. Under each mass accuracy level, from left to right, the data files contain 0%, 50%, 67%, and 75% of noise peaks respectively.

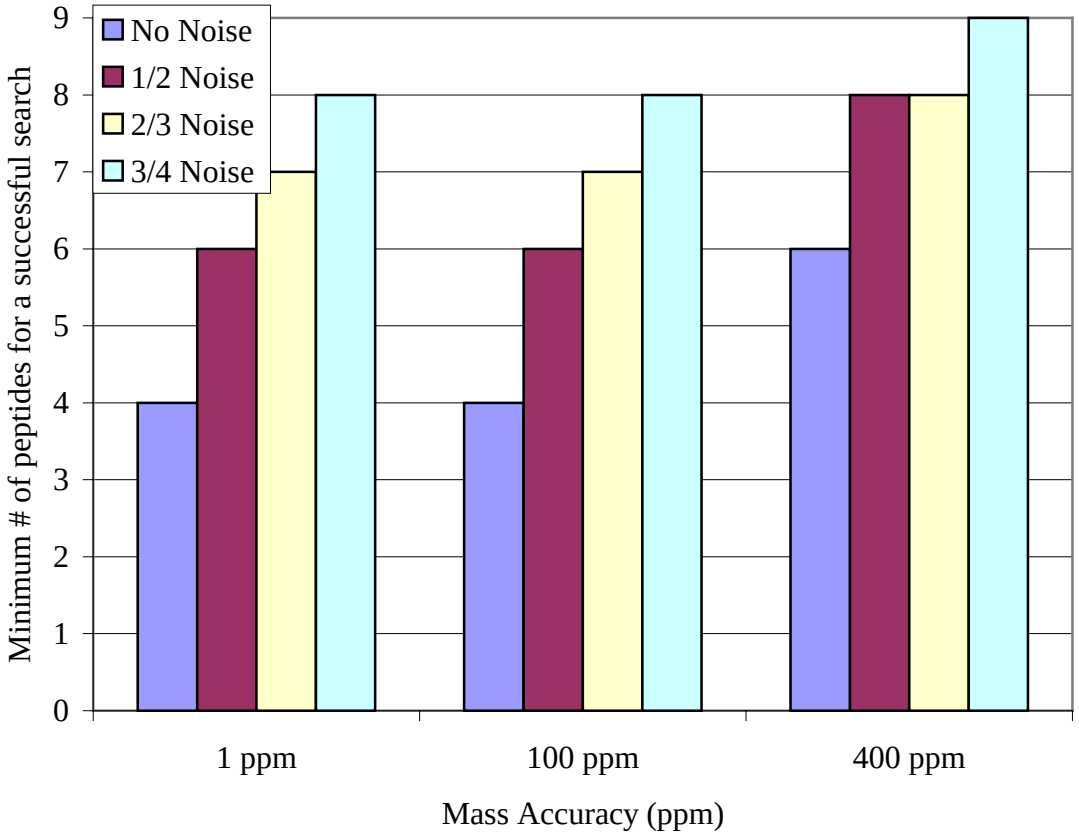

Figure 3.5 (B). Mascot result: The Effect of Impurity on Searching Efficiency (Protein B)

Results show the searching efficiency for ATPase beta subunit. Searches were under a mass accuracy level of 1 ppm, 100 ppm, and 400 ppm. Under each mass accuracy level, from left to right, the data files contain 0%, 50%, 67%, and 75% of noise peaks respectively.

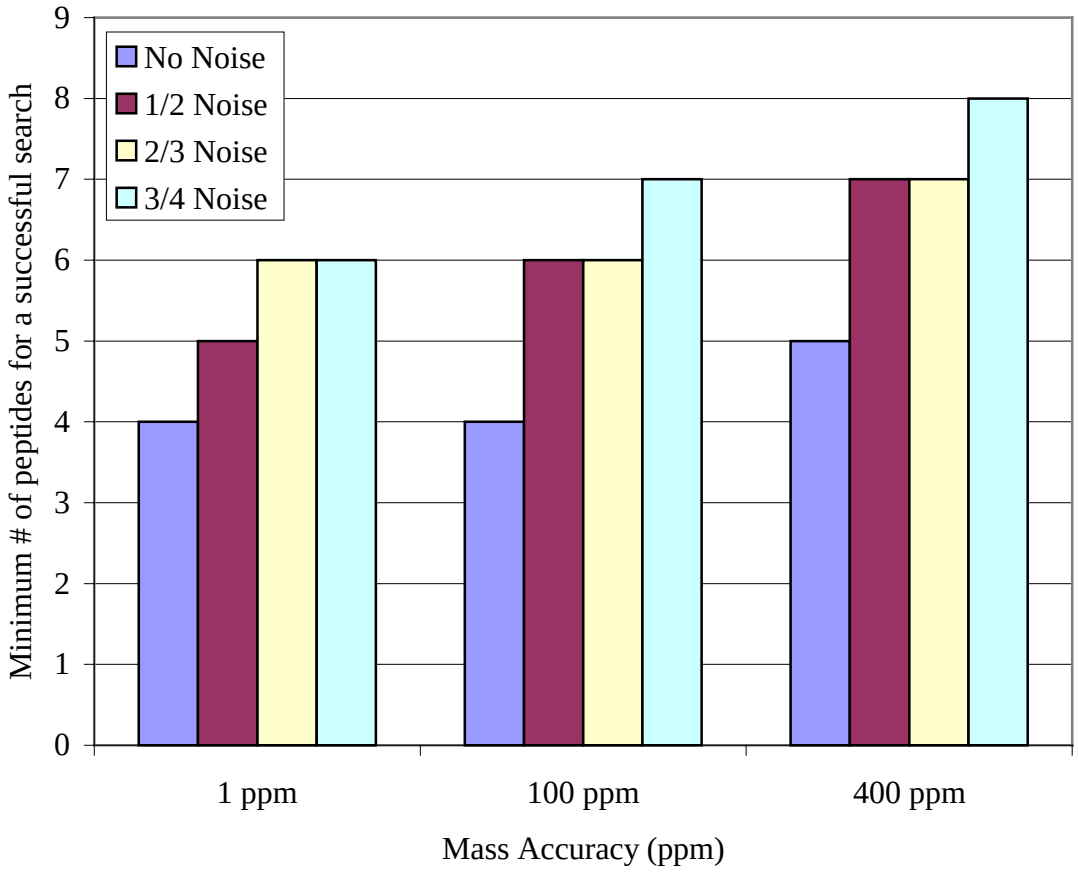

Figure 3.5 (C). Mascot result: The Effect of Impurity on Searching Efficiency (Protein C)

Results show the searching efficiency for Fructose bisphosphate-aldolase like protein. Searches were under a mass accuracy level of 1 ppm, 100 ppm, and 400 ppm. Under each mass accuracy level, from left to right, the data files contain 0%, 50%, 67%, and 75% of noise peaks respectively.

The Mascot result suggests that:

1) More peptide ions are required to identify a protein with increasing amount of impurity in the searched data.

Searching results for all 3 tested proteins showed a common effect of noise on protein identification. By allowing extraneous noise in the searched data file, more peptide peaks were required to identify a protein at all mass accuracy levels. Table 3 compares the minimum number of peptide ions required to identify protein A, B, and C using data with/without noise. Searches were performed under an ideal mass accuracy of 1 ppm to avoid potential effect from mass error. In Table 3, all three tested proteins showed that more peptide ions were required for a successful search when noise peaks were introduced. Submitting more peptide ions compensated for the effect from noise. Two, four and two more peptide ions were required to identify protein A, B, and C, respectively, after 3-fold noise peaks were introduced. Results also indicate that for different proteins, the ability to tolerate impurity might vary. Protein B responded to impurity more sensitively than did Proteins A or C.

Table 3.2 Minimum Number of Peptide Ions Required at Different Noise Levels (Mascot)

Noise Level	Protein A	Protein B	Protein C
No Noise	4	4	4
1/2 Noise	5	6	5
2/3 Noise	5	7	6
3/4 Noise	6	8	6

Searches performed using peptide ions with 1 ppm mass accuracy to exclude the potential influence from mass accuracy.

2) High mass accuracy is able to compensate for the effect of impurity in some cases

We noticed that although all tested proteins show a common trend when responding to increasing noise level, high mass accuracy data compensates for the influence from noise. For example, searching the data file under the1 ppm mass accuracy category, two more peptide ions were required to identify protein A after 3/4 noise peaks were introduced. With the same level of impurities submitted at 400 ppm, four additional peptide ions were needed to ensure a good search (Figure 3.5(A)). A similar difference due to mass accuracy levels was observed in the protein C results. In both cases, data with good mass accuracy showed a higher potential to tolerate impurities in the searched data files.

(2) The result of ProFound searching:

To determine the effect of noise on protein identification, parallel experiments were conducted using ProFound. Figure 3.6 (A-C) shows the search result with ProFound for proteins A, B, and C, respectively. Searches were performed using data with 1 ppm, 100 ppm and 400 ppm to compare with the Mascot results.

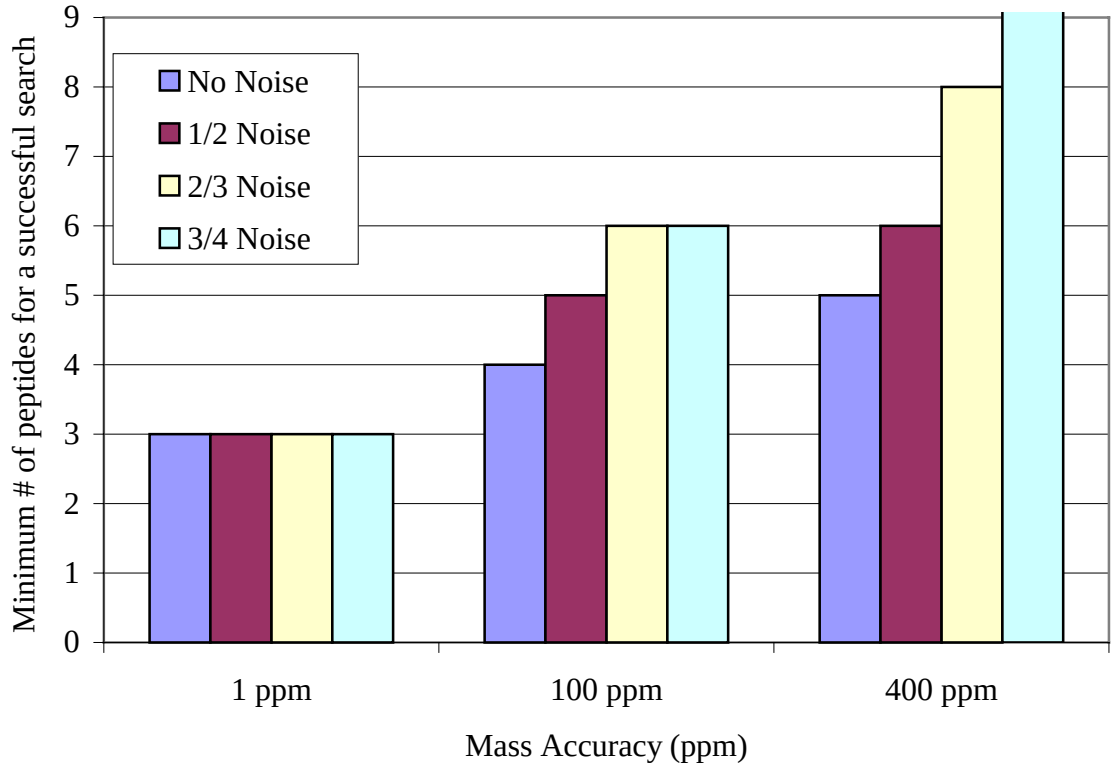

Figure 3.6 (A). ProFound result: The Effect of Impurity on Searching Efficiency (Protein A)

Results show the searching efficiency for Apolipoprotein AI protein. Searches were under a mass accuracy level of 1 ppm, 100 ppm, and 400 ppm. Under each mass accuracy level, from left to right, the data files contain 0%, 50%, 67%, and 75% of noise peaks respectively.

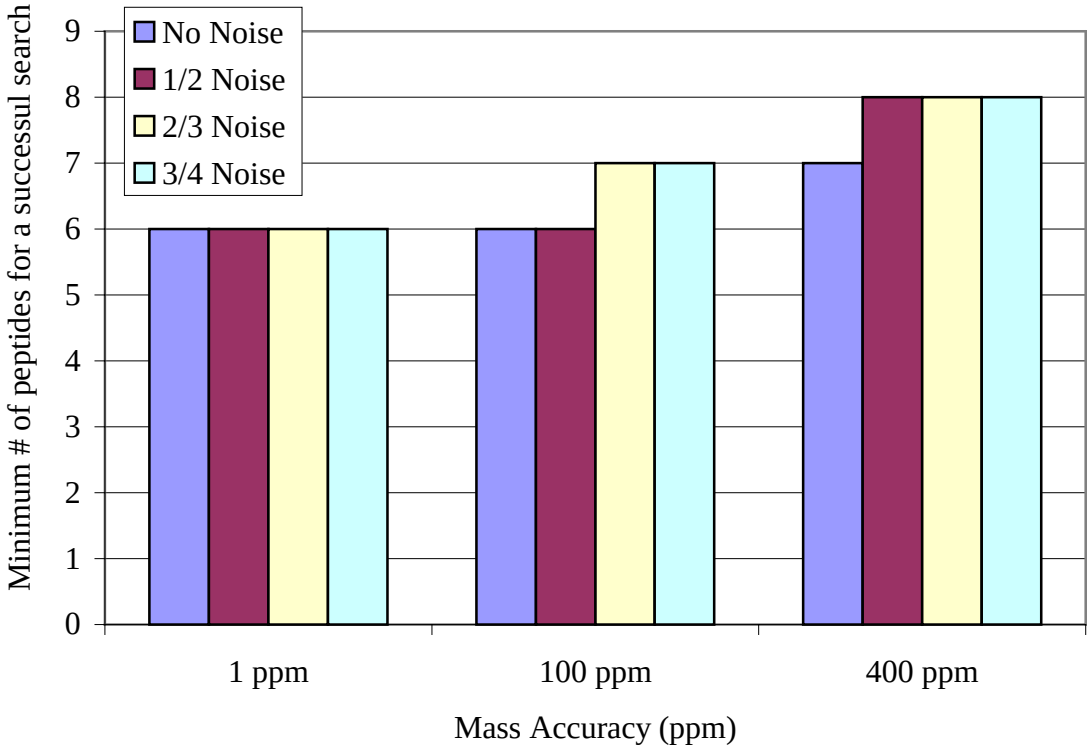

Figure 3.6 (B). ProFound result: The Effect of Impurity on Searching Efficiency (Protein B)

Results show the searching efficiency for ATPase beta subunit protein. Searches were under a mass accuracy level of 1 ppm, 100 ppm, and 400 ppm. Under each mass accuracy level, from left to right, the data files contain 0%, 50%, 67%, and 75% of noise peaks respectively.

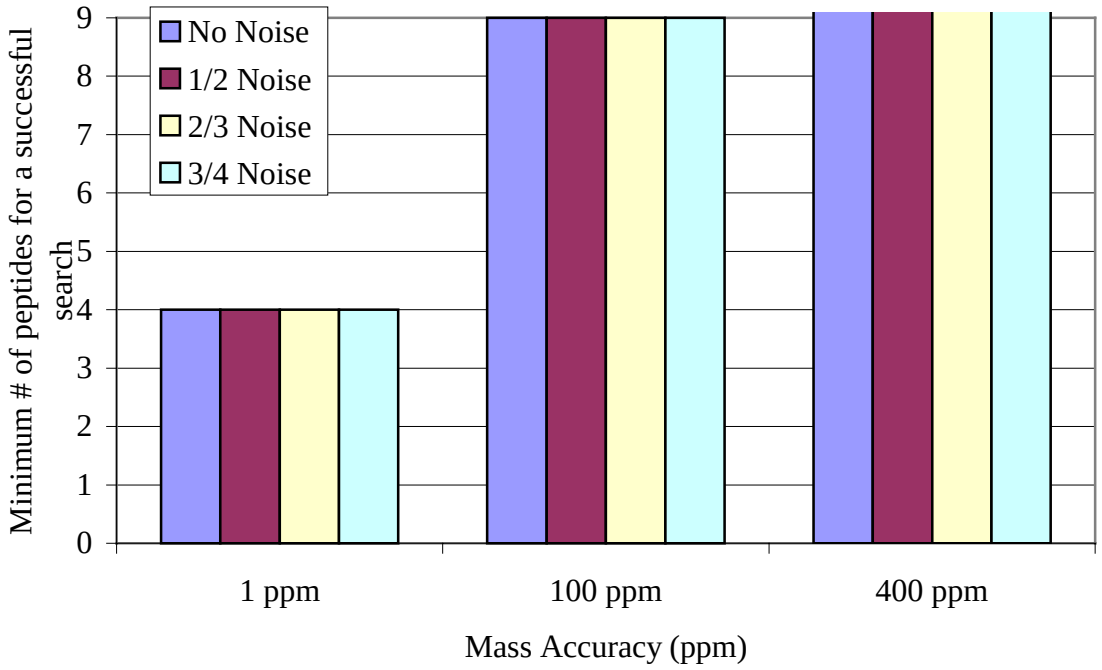

Figure 3.6 (C). ProFound result: The Effect of Impurity on Searching Efficiency (Protein C)

Results show the searching efficiency for Fructose bisphosphate-aldolase like protein. Searches were under a mass accuracy level of 1 ppm, 100 ppm, and 400 ppm. Under each mass accuracy level, from left to right, the data files contain 0%, 50%, 67%, and 75% of noise peaks respectively.

In general, the ProFound results agree with Mascot in that impurity has a negative effect on protein identification: more peptide ions are required to identify a protein when extraneous noise peaks are introduced. Both protein A and B showed an increase in the

minimum peptide requirement with mass accuracy of 100 ppm or worse. Comparing with the Mascot result, as Table 3.3 suggested, ProFound showed a better tolerance of impurity under very good mass accuracy such as 1 ppm. With high quality data submitted, the minimum peptide requirement for proteins A, B, and C hold the same after introducing noise peaks to as much as a 3-fold level.

Table 3.3. Minimum Peptide Ions Required at Different Noise Levels (ProFound)

Noise Level	Protein A	Protein B	Protein C
No Noise	3	6	4
1/2 Noise	3	6	4
2/3 Noise	3	6	4
3/4 Noise	3	6	4

Searches performed using peptide ions with 1 ppm mass accuracy to exclude the potential influence from mass accuracy.

3.3 Effect of Database Restrictions

To investigate the effect of limiting the database searches on peptide mapping results, proteins A, B, and C were searched using Mascot and ProFound against the NCBI nr database. Since ProFound provides more options for database limitation, we analyze this issue using ProFound, with Mascot results as an addition.

ProFound allows users to limit molecular weight, isoelectric points, and sample organism. Based on the knowledge of sample resources and 2-D results (discussed in Chapter 2), the limiting parameters were set up in six different ways to compare the dynamic effect of limits on searching results.

The on-line searching by ProFound was performed with the following constraints:

1) Limiting molecular weight (MW);

2) Limiting isoelectric point (pI);

3) Limiting species;

4) Limiting MW, PI, and species simultaneously.

The on-line searching by Mascot was performed in a similar design except for that Mascot does not provide the option to limit pI in searches.

The limit parameters were set up as shown in Table 3.4.

Table 3.4. Parameters Set-up when Limiting Searches

Protein	Molecular weight	Point of isoelectric	Database
Protein A	27-33 KD	4.9-5.9	Mammal
Protein B	50-56 KD	4.9-5.9	*Arabidopsis*
Protein C	36-42KD	5.5-6.5	*Arabidopsis*

We evaluated the effects of the four (three for Mascot) different limiting methods by comparing the minimum peptide requirement to identify proteins A-C with 95% confidence (Figure 3.7, Figure 3.8, and Figure 3.9).

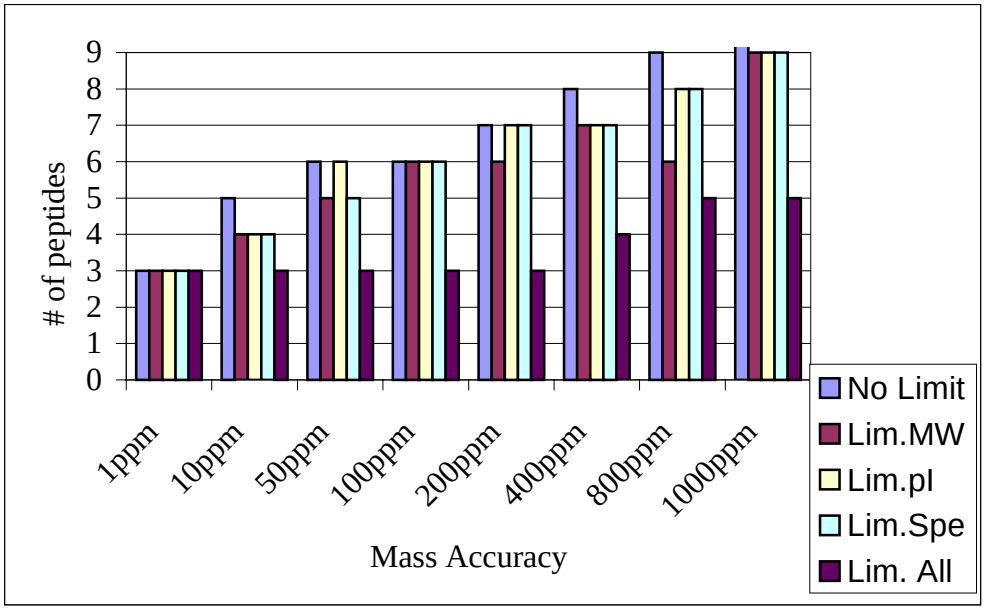

Figure 3.7 (A). Effect of Placing Limits on Searching Efficiency for Protein A (ProFound)

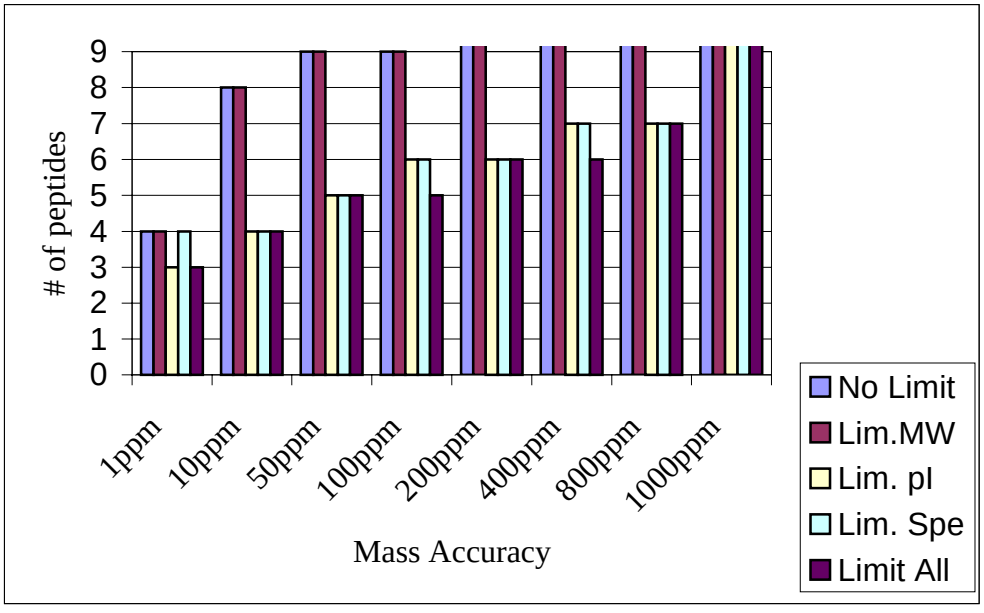

Figure 3.7 (B). Effect of Placing Limits on Searching Efficiency for Protein B (ProFound)

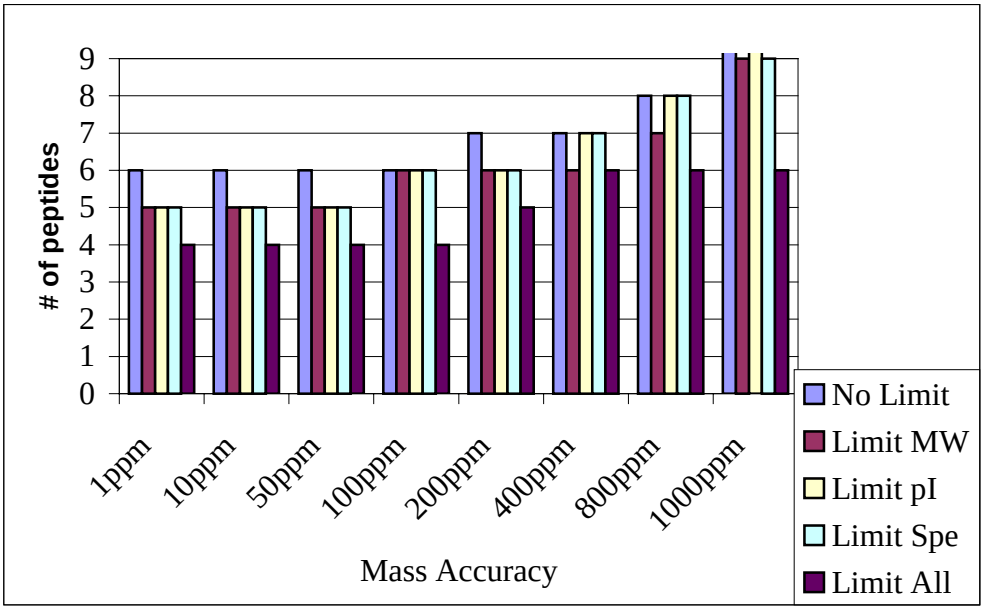

Figure 3.7 (C). Effect of Placing Limits on Searching Efficiency for Protein C (ProFound)

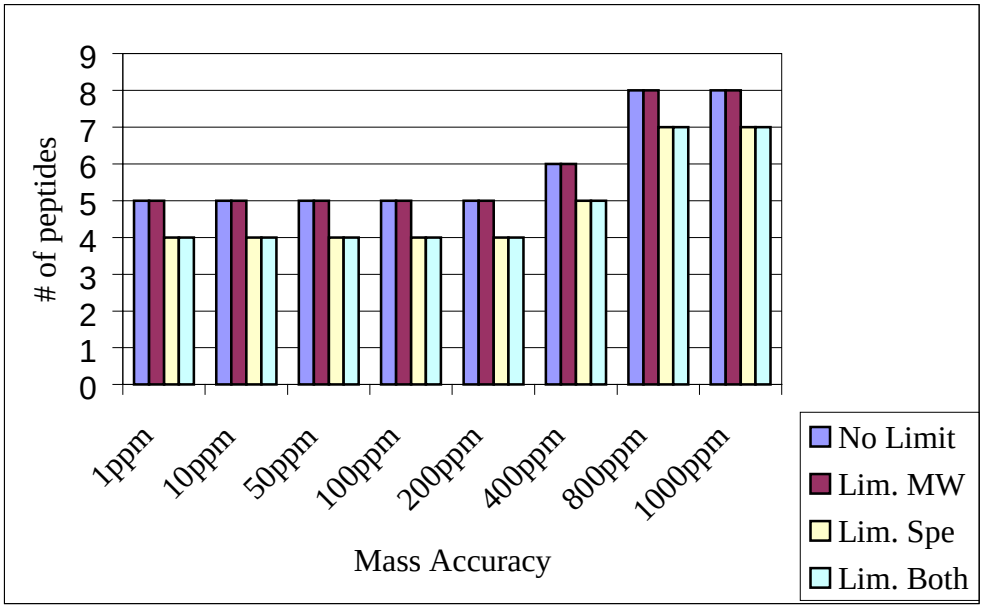

Figure 3.8 (A). Effect of Placing Limits on Searching Efficiency for Protein A (Mascot)

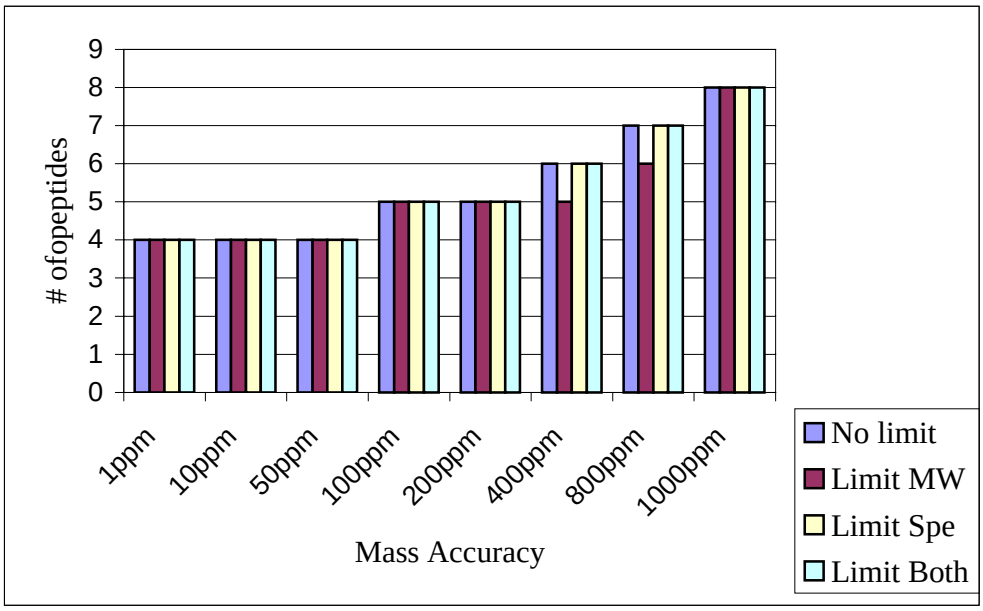

Figure 3.8 (B). Effect of Placing Limits on Searching Efficiency for Protein B (Mascot)

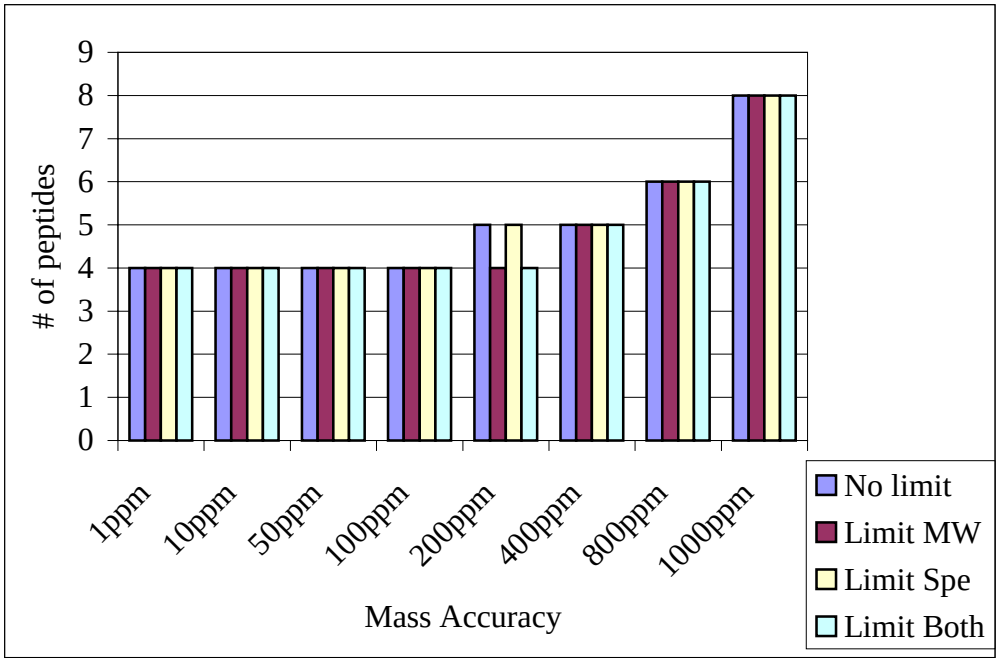

Figure 3.8 (C). Effect of Placing Limits on Searching Efficiency for protein C (Mascot)

(1) The effect of limiting the database on searching efficiency

ProFound searching results shown in Figure 3.7 (A-C) suggest that properly setting up limit(s) increases the searching efficiency by allowing fewer peptide ions for a successful identification. For each protein, compared with the no limit, limit MW, pI, and/or organism lessened the number of peptide ions required in general. The search results suggest that the effectiveness of limiting the database is more apparent when the mass accuracy is comparatively poor. For instance, in Figure 3.7 (A), at1000 ppm mass accuracy, enabling all possible limits allows as many as five fewer peptide ions for a successful search. While for the same protein, limiting the database did not show a significant effectiveness when searching data at 1 ppm mass accuracy. Similar results were observed for proteins B and C.

Unlike the ProFound results, Mascot in general did not show a significant effect of limiting the database search at all mass accuracies tested (Figure 3.8 (A-C))

(2) The effect of limiting the database on searching confidence

To test the effect of limits on searching confidence, we searched the peptide pool from protein A with a fixed size of six peptide ions at different mass accuracies. We chose to use ProFound, which, in the above analysis, was significantly more effective in allowing fewer peptide ions for a successful search. Table 3.5 compares the original probability-based profound Z-score [28-30] with the ones returned when limit(s) is enabled.

Table 3.5 Comparison Searching Confidence for Protein A when Limiting Database (ProFound)

Mass Accuracy	No Limit	Lim. MW	Lim. pI	Lim. Spe.	Lim. All
1 ppm	2.36	2.36	2.36	2.43	2.43
10 ppm	2.39	2.38	2.41	2.38	2.43
100 ppm	1.69	2.08	1.97	1.98	2.43
1000 ppm	0.67	1.05	1.00	1.01	2.24

The results shown in Table 3.5 suggest that with a fixed number of peptide ions submitted, at each mass accuracy category, the probability-based ProFound Z-score is generally increased by database limitation. It also shows that the improvement of searching confidence by limits is more apparent when searching data with poor mass accuracy. Comparing different limit methods, we found that different methods affect peptide mapping to a variable degree. The effectiveness of limiting MW, pI, or species varied with mass accuracy, different protein, and/or number of peptide ions submitted. But in general, best searching confidence scores were always associated with the combination of limits.

CHAPTER 4

FACTORS IN FRAGMENTMASS FINGERPRINTING USING TANDEM MASS SPECTROMETRY

4.1 Effect of Peptide and Fragment Mass Accuracies

Mascot MS/MS search result for fructose bisphosphate-aldolase like protein are shown in Figure 4.1 and Figure 4.2. To analyze the dynamic interaction between the m/z accuracy from precursor peptide ions and fragment mass tolerance, two precursor mass accuracy levels [10 ppm, 100 ppm] and 8 fragment tolerance levels [0.01Da, 0.02Da, 0.05Da, 0.1Da, 0.5Da, 1Da, 1.5D, 2Da] are completely crossed, i.e. under each precursor mass accuracy level, spectra with eight levels of fragment tolerance are searched. Analysis of search scores within and between precursor levels show that:

(1) Unlike the results obtained with peptide mass fingerprint (PMF), which suggests that high mass accuracy gained high confidence search score, the MS/MS protein identification score is independent of mass accuracy. Comparing Figure 4.1 with Figure 4.2, although the precursor mass accuracy is changed from 10 ppm to 100 ppm, a same search score of 29 were observed from both search. Within a fixed precursor mass accuracy level, changing fragment mass tolerance did not alter the searching score either.

(2) Altering precursor mass accuracy from 10 ppm to 100 ppm raised the confidence thresholds for identify and homology identifications. Using the Mascot scoring system [31], the threshold for identify was raised by ten; the threshold for homology was raised by about five when fragment mass tolerance was lower than 0.05 Da. When fragment mass tolerance is better than 0.05 Da, the confidence threshold did not affected by precursor mass accuracy.

(3) With precursor mass accuracy fixed, trends of confidence score with changing fragment mass tolerance are similar for Figure 4.1 and Figure 4.2. Both figures suggest that relaxing fragment mass tolerance affects search results in that it requires a higher confidence score for successfully identifying protein homology. Searches at a precursor mass accuracy of 10 ppm revealed that four more Mascot score units were required to identify protein homology when fragment mass accuracy was compromised from 0.01 Da to 0.5 Da. When precursor mass accuracy was lowered to 100 ppm, as many as nine more Mascot score units were required to identify a protein homology. Meanwhile, both search score and score threshold remained unchanged with relaxing fragment mass accuracy for protein identification.

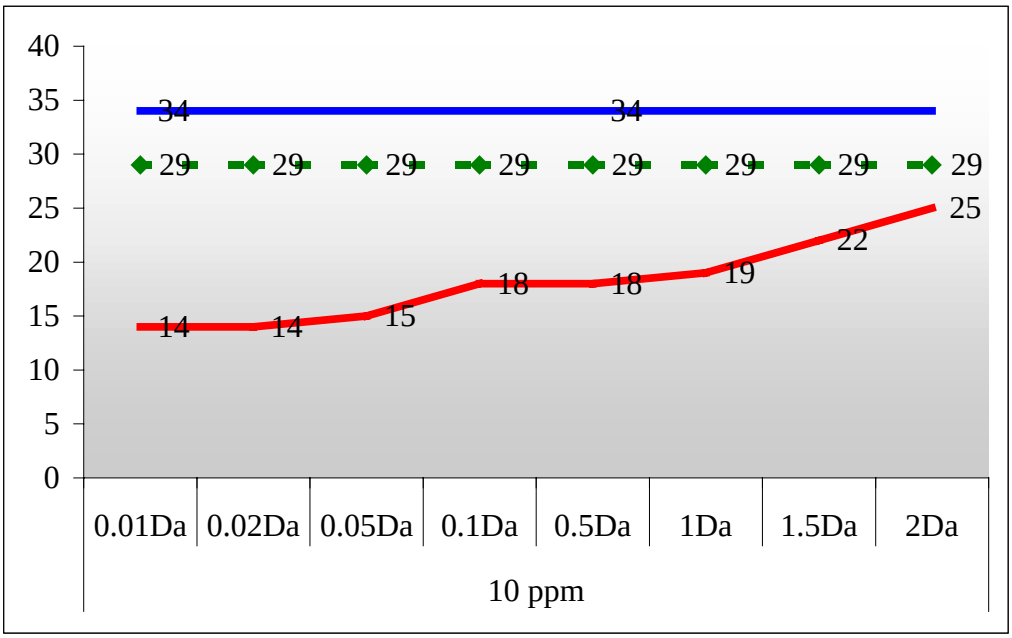

Figure 4.1. Effect of Peptide Mass Accuracy and Fragment Mass Accuracy (Precursor mass accuracy 10 ppm)

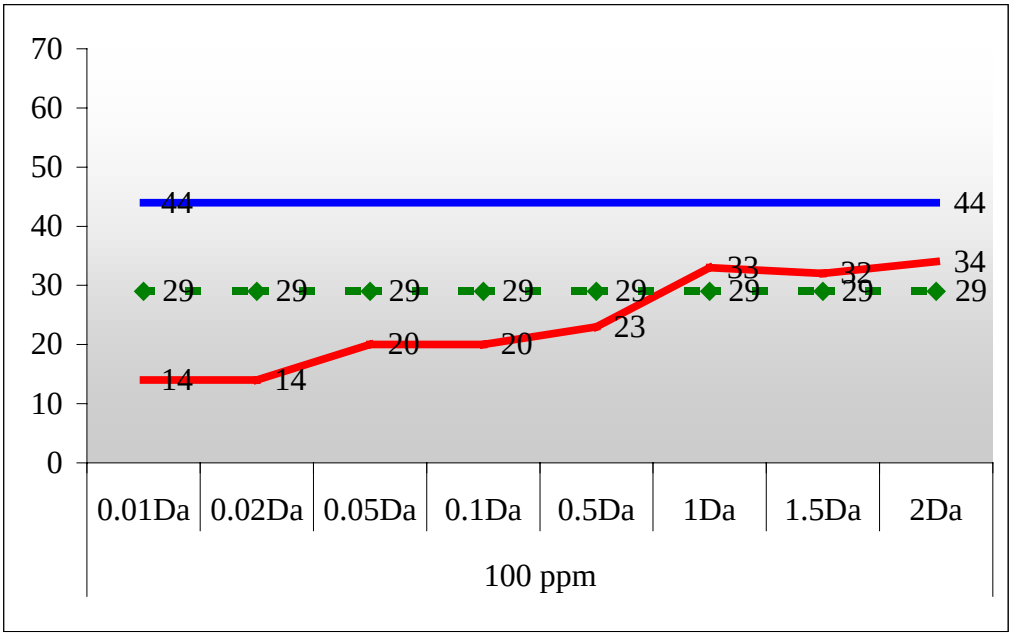

Figure 4.2. Effect of Peptide Mass Accuracy and Fragment Mass Tolerance (Precursor mass accuracy 100 ppm)

4.2 Effect of the Number of Fragments Submitted

Figure 4.3 and Figure 4.4 show that for each fragment mass accuracy level tested; decreasing the number of fragments submitted lowers the Mascot search score and results in failure of protein identification. Table 4.1 lists the minimum number of fragments required for identifying protein C. We observed that two more fragments were required for identifying protein homology when fragment mass accuracy was relaxed from 0.02 Da to 2 Da; while at least one more fragment was required for protein identity.

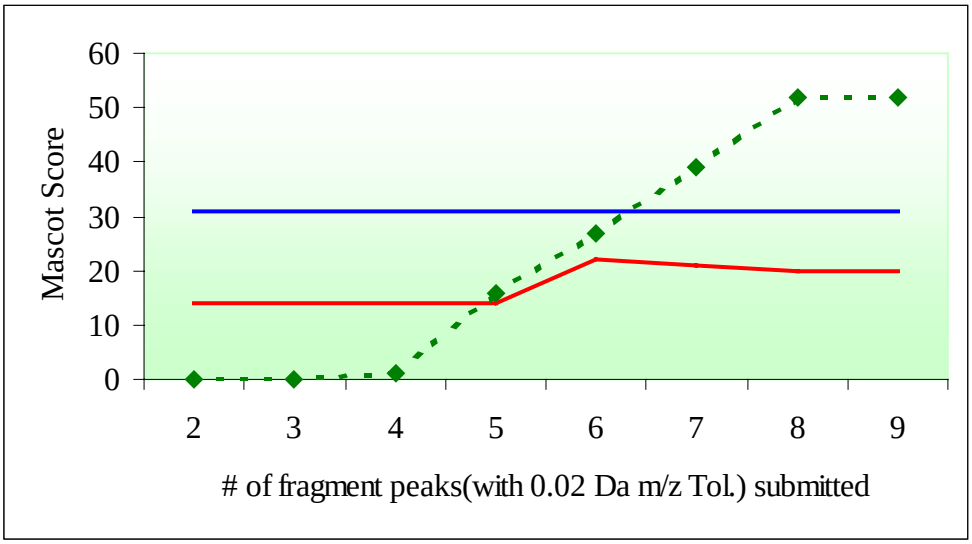

Figure 4.3 Mascot Searching Score vs. Number of Fragment Submitted (0.02Da)

(Fragment mass accuracy: 0.02 Da)

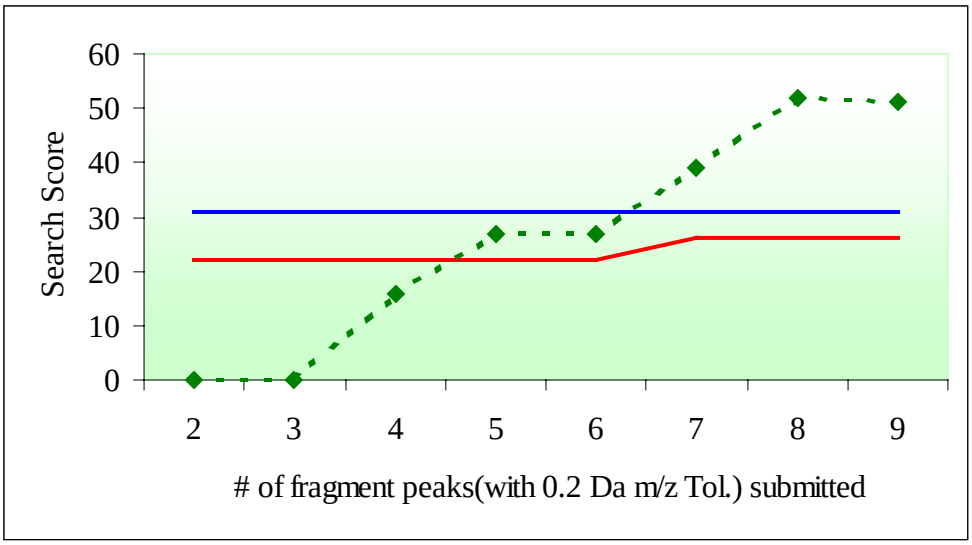

Figure 4.4. Mascot Searching Score vs. Number of Fragment Submitted (0.2 Da)

(Fragment mass accuracy: 0.2 Da)

Table 4.1 Minimum Number of Fragments Required for Identifying Protein Homology and Identity

Fragment mass accuracy (Peptide accuracy = 10 ppm)	Minimum number of fragments to identify	
	Homology	Identity
0.02 Da	5	7
0.2 Da	5	7
2 Da	7	8

CHAPTER 5

CONCLUSIONS

5.1 Mass Accuracy and the Number of Peptide Ions Submitted

The results for searching three proteins using two different programs indicate that high mass accuracy data require fewer peptide ions for a successful search. Mascot shows that using low mass accuracy data (allowing wider mass error window) increases the chance of a random match in peptide mapping. ProFound verified the observations by showing that high mass accuracy peptide ions minimized the possibility of random match.

Searching results also show that high mass accuracy usually helps to improve searching confidence. Since most of searching programs will give a relatively fixed confidence threshold for user's reference, an improved confidence score sometimes allows identifying a protein that which cannot be identified with 95% confidence level using data of low quality.

In practice, because of the instrument used, mass error is sometimes inevitable. If average or low mass accuracy data are searched, enlarging the peptide ion pool submitted is an effective way to increase searching confidence. Also, because the mathematical and statistical methods utilized by different program vary in how they consider and deal with mass error, multiple programs are recommended for poor mass accuracy data to enhance searching accuracy and efficiency.

In tandem mass spectrometry, the precursor peptide mass accuracy has negligible effect on determination of protein identity. Peptide mass accuracy can be relaxed from 10 ppm to 200 ppm without changing the 95% identity confidence threshold and the search score, though relaxing peptide mass accuracy from 10 ppm to 100 ppm does increase the protein identity and

homology confidence threshold apparently (i.e., peptide mass accuracy of 100 ppm or lower decreases the chance to identify proteins with 95% confidence). MS/MS fragment accuracy affects searching results by altering the confidence scores required for detecting protein homology and identity. Allowing an MS/MS mass error of greater than 1 Da increases the homology confidence threshold. Changing MS/MS fragment m/z accuracy from 0.02 Da to 1 Da has negligible influence on search result.

5.2 Impurity in PMF

Searching efficiency is weakened by impurity of the data; and severe impurity of data affects search results to great extent. Using Mascot, the addition of random noise peaks has a significant effect in increasing the number of peptide ions and the mass accuracy required. The addition of three random noise peaks per sample peak eliminates the ability to confidently identify the protein even with 1 ppm mass accuracy searching with ten peptide ions. On the other hand, searching results show that ProFound has a higher ability to tolerate extraneous noise than Mascot especially with good mass accuracy.

5.3 Database Restrictions

On-line peptide mapping results indicate that properly limiting the searches based on the available information of the submitted peptide pools improved peptide mapping accuracy and efficiency. In our searches, the help proved by limiting the database varied from case to case; thus, we cannot conclude arbitrarily that a particular limit is more effective in peptide mapping than others. A significant improvement is observed when searches are limited by MW, pI, and database when using ProFound. The major result of placing all of the limits on a search is that it reduces the mass accuracy needed for significant protein identification.

REFERENCES

[1] Siuzdak G. (1994) The emergence of mass spectrometry in biochemical research. Proc. Natl. Acad. Sci. USA 91: 11290–11297.

[2] Fenselau C. (1992) Tandem mass spectrometry: the competitive edge for pharmacology. Annu. Rev. Pharmacol. Toxicol. 32: 555–578.

[3] Roepstorff P. (1997) Mass spectrometry in protein studies from genome to function. Curr. Opin. Biotech. 8: 6–13.

[4] Mann M., Højrup P. and Roepstorff P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22: 338–345.

[5] Pandey A. and Mann M. (2000) Proteomics to study genes and genomes. Nature 405: 837–846.

[6] Jonsson A. P., Griffiths W. J., Bratt P., Johansson I., Strömberg, N., Jörnvall H. et al. (2000) A novel Ser O-glucuronidation in acidic proline-rich proteins identified by tandem mass spectrometry. Febs Lett. 475: 131–134.

[7] Annan R. S., Huddleston M. J., Verma R., Deshaies R. J. and Carr S. A. (2001) A multi-dimensional electrospray MS-based approach to phosphopeptide mapping. Anal. Chem. 73: 393–404.

[8] Karas M., Bachmann D., Bahr U. and Hillenkamp F. (1987) Matrix-assisted ultraviolet laser desorption of non-volatile compounds. Int. J. Mass Spectrom. Ion Proces. 78: 53–68.

[9] Wiley W. C. and McLaren I. H. (1955) Time-of-flight mass spectrometer with improved resolution. Rev. Sci. Instrum. 26:1150–1157.

[10] Guilhaus M., Mlynski V. and Selby D. (1997) Perfect timing: time of flight mass spectrometry. Rapid Commun. Mass Spectrom.11: 951–962.

[11] Roepstorff P. (2000) MALDI-TOF mass spectrometry in protein chemistry. In: Proteomics in Functional Genomics: Protein Structure Analysis, pp. 81–97.

[12] McLafferty F. W. (1983) Tandem Mass Spectrometry. Wiley, New York.

[13] Hellman U., Wernstedt C., Gonez J. and Heldin C. H. (1995) Improvement of an “in-gel” digestion procedure for the micropreparation of internal protein fragments for amino acid sequencing. Anal. Biochem. 224: 451–455.

[14] Mann M., Højrup P. and Roepstorff P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22: 338–345.

[15] Pappin D. J. C., Højrup P. and Bleasby A. J. (1993) Protein identification by peptide mass fingerprinting. Curr. Biol. 3: 327–332.

[16] Yates J. R., Speicher S., Griffin P. R. and Hunkapiller T. (1993) Peptide mass maps: a highly informative approach to protein identification. Anal. Biochem. 214: 397–408.

[17] Henzel W. J., Billeci T. M., Stultz J. T., Wong S. C., Grimley C. and Wantanabe C. (1993) Identifying proteins from twodimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc. Natl. Acad. Sci. USA 90: 5011–5015.

[18] Scheler, C., Lamer, S., Pan, Z., Li, X. P., Salnikow, J., and Jungblut, P. (1998) Peptide mass fingerprint sequence cover-age from differently stained proteins on two-dimensional electrophoresis patterns by matrix assisted laser desorption/ionization-mass spectrometry (MALDI-MS). Electro-phoresis19, 918-927.

[19] James P., Quadroni M., Carafoli E. and Gonnet G. (1993) Protein identification by mass profile finger printing. Biochem. Biophys. Res. Commun. 195: 58–64.

[20] Green, M. K., Johnston, M. V., and Larsen, B. S. (1999) Mass accuracy and sequence requirements for protein database searching. Anal. Biochem.275, 39-46.

[21] Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (+/-10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal. Chem.71, 2871-2882.

[22] Fukai, T., Kuroda, J., and Nomura, T. (2000) Accurate mass measurement of low molecular weight compounds by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. J. Am. Soc. Mass Spectrom 11, 458-463.

[23] Gharahdaghi, F., Weinberg, C. R., Meagher, D. A, Imai, B. S., and Mische, S. M. (1999) Mass spectrometric identification of proteins from silver-stained polyacrylamide gel: a method for the removal of silver ions to enhance sensitivity. Electro-phoresis 20, 601-605.

[24] Kussmann, M. and Roepstorff, P. (2000) Sample preparation techniques for peptides and proteins analyzed by MALDI-MS. Methods Mol. Biol.146, 405-424.

[25] Jensen, O. N., Larsen, M. R., and Roepstorff, P. (1998) Mass spectrometric identification and microcharacterization of proteins from electrophoretic gels: strategies and applications. Proteins Suppl. 2,74-89.

[26] Hellman, U., Wernstedt, C., Gonez, J., and Heldin, C. H. (1995) Improvement of an "In-Gel" digestion procedure for the micropreparation of internal protein fragments for amino acid sequencing. Anal. Biochem.224, 451-455.

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

[28] Karas, M. and Hillenkamp, F. (1988) Laser desorption ioniza-tion of proteins with molecular masses exceeding 10,000 daltons. Anal. Chem. 60, 2299-2301.

[29] Zhang, W and Chait, BT. (2000) ProFound: An Expert System for Protein Identification Using Mass Spectrometric Peptide Mapping Information Anal. Chem.2000, 72,2482-2489.

[30] Zhang, W and Chait, BT. (1995) Protein Identification by Database Searching: A Bayesian Algorithm. Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics, Atlanta, Georgia, 1995

[31] Pappin, DJC, Hojrup, P and Bleasby, AJ. (993) Rapid identification of proteins by peptide-mass fingerprinting. Curr. Biol. 3(6): 327-332

APPENDIX A

ON-LINE DATABASE SEARCH PROGRAMS

Mascot (http://www.matrixscience.com/search_form_select.html)

David M Creasy, John S Cotrell, matrix sciences

ProFound (http://prowl.rockefeller.edu/profound_bin/WebProFound.exe)

Wenzhu Zhang and Brian T. Chait.

MS-FIT (http://prospector.ucsf.edu/ucsfhtml3.4/msfit.htm)

Karl Clauser and Peter Baker, Alma Burlingame, UCSF

PeptIdent (http://www.expasy.org/tools/peptident.html)

Under ExPASy server

SEQUEST (not available on the web)

PeptideSearch

(http://www.mann.emblheidelberg.de/GroupPages/PageLink/peptidesearchpage.html)

Mathias Mann

PeptideMapper (http://wolf.bms.umist.ac.uk/mapper/)

PepSea (http://195.41.108.38/PepSeaIntro.html)

Matthias Mann, PROTANA / MDS Proteomics