UNIVERSIDADE DO ALGARVE

Faculdade de Ciências e Tecnologia

A Computational Tool for Peptide Mass Fingerprinting

Eman Saad Ali AL-Hawri

Mestrado em Engenharia Informática

2013

UNIVERSIDADE DO ALGARVE

Faculdade de Ciências e Tecnologia

A Computational Tool for Peptide Mass Fingerprinting

Eman Saad Ali AL-Hawri

Tese orientada por

António dos Anjos

Tese coorientada por

Gareth Pearson

Mestrado em Engenharia Informática

2013

A Computational Tool for Peptide Mass Fingerprinting

Declaração de autoria de trabalho Declaro ser a autora deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam na listagem de referéncias incluída.

Abstract

Protein identification using Mass Spectrometry (MS) is essential in the study of proteomics. Two popular techniques are used in the identification: Tandem Mass Spectrometry (MS/MS) and Peptide Mass Fingerprinting (PMF), which is considered in this work. PMF is widely used in the proteomics field. It is faster and more economic when compared to MS/MS.

This work focuses on the development of a computational tool for protein identification using PMF data. The main objective for any PMF tool is to identify the correct protein (if it exists) by searching a peak list, produced by MS, against a protein database. However, one of the great challenges to these tools is related to the size of the databases that result in many random matches. In fact, the main difference between these tools is the scoring method which is responsible of minimizing these random matches. Therefore, a review of PMF tools and their scoring methods is presented and discussed.

There are many tools on the Internet (both commercial or academic) for PMF protein identification using public databases. These tools do not offer a locally installable version, and do not allow the use of in-house databases, a feature that is of great importance to biologists who work on non-model systems. In contrast, the tool developed in this work is free, can be installed locally, and can be used with both public and local databases. Additionally, it supports different sorts of protein modifications and contaminants suppression, features that are not available by some of the existing tools.

A new scoring method is proposed and incorporated in the proposed tool. The proposed tool is compared with two of the most popular software packages (commercial and academic), showing a good accuracy and being very competitive with the most popular and robust commercial software (Mascot). The developed prototype is platform-independent and is very easy to install. To allow users to work and interact with the system in an easy-to-use environment, a friendly graphical user interface is developed to allow them to manage their files very efficiently. In addition, it can work with single or multiple query files to support different work scales. The features this new tool offers make it an important assist to the biological laboratories concerning the PMF task.

Resumo

A identificação de proteínas utilizando Espectrometria de Massa (MS) é essencial no campo da proteómica. Há duas técnicas muito populares utilizadas para a identificação: Tandem Mass Spectrometry (MS/MS) and Peptide Mass Fingerprinting (PMF), sendo esta última a abordada nesta tese. A PMF é vastamente utilizada no campo da proteómica. Quando comparada com a MS/MS, esta é mais rápida e mais económica.

O foco deste trabalho é o desenvolvimento de uma ferramenta computacional para a identificação de proteínas utilizando dados resultantes da PMF. O objetivo principal de qualquer ferramenta de PMF é o de identificar a proteína correta (se esta existir) por procurar uma lista de picos, produzidos através de MS, numa base de dados de proteínas. No entanto, um dos grandes desafios destas ferramentas prende-se com o grande tamanho das bases de dados, que levam a que haja muitos matches aleatórios. De facto, a principal diferença entre as ferramentas existentes é o método de scoring, o qual é responsável por minimizar os matches aleatórios. Desta forma, apresenta-se uma revisão e faz-se uma discussão das ferramentas de PMF e respetivos métodos de scoring.

Existem várias ferramentas na Internet (tanto comerciais como académicas) para identificação de proteínas através de PMF utilizando bases de dados públicas. Estas ferramentas não oferecem uma versão que permita a instalação local, e não permitem a utilização de bases de dados caseiras, uma funcionalidade que é de grande importância para biólogos que trabalham em sistemas não-modelo. Em contraste, a ferramenta desenvolvida neste trabalho, além de livre, pode ser instalada localmente, e pode ser utilizada tanto com bases de dados públicas como caseiras. Além disso, também suporta diferentes tipos de modificações de proteínas e supressão de contaminantes, funcionalidades não disponíveis em algumas das ferramentas existentes.

Propõe-se um um novo método de scoring e incorpora-se o mesmo na ferramenta proposta. Esta é comparada com dois dos mais poderosos pacotes de software disponíveis, sendo que a ferramenta proposta apresenta uma boa prestação e é bastante competitiva com o mais popular e robusto software comercial (i.e. Mascot). O protótipo desenvolvido é independente da plataforma onde corre e de muito fácil instalação. Para permitir que os utilizadores pos-

sam trabalhar e interagir com o sistema de uma forma simples, foi desenvolvida uma interface gráfica bastante amiga do utilizador. Esta permite a gestão dos ficheiros de projeto de forma muito eficiente. Adicionalmente, a ferramenta proposta pode trabalhar com ficheiros de uma ou múltiplas queries. Esta ferramenta e as funcionalidades oferecidas pela mesma, contribuem de forma relevante para assistir os laboratórios da área da biologia no que diz respeito à PMF.

Keywords: Protein, peptide, mass, molecular weight, fingerprint, Mass Spectrometry, FASTA, amino acid, enzyme, digestion, trypsin.

Acknowledgments

I would like to thank my advisors Prof. António dos Anjos and Prof. Gareth Pearson for their support, patience, guidance, and encouragement throughout this work. Many thanks for Miss Catarina Mota for her help, and great comments to improve this work.

I would like to dedicate this work to:

My parents for their endless love and support, and for giving me the strength throughout my life.

My husband for his support, love, and for teaching me the real life.

My brothers and sisters for their love and emotional encouragement.

My nephews Mony, Shado, and Fahood.

Contents

List of Figures

List of Tables

Acronyms

PMF Peptide Mass Fingerprinting

MS Mass Spectrometry

Da Dalton

PTM Post-Translation Modification

MW Molecular Weight

MALDI Matrix Assisted Laser Desorption/Ionization

TOF Time-Of-Flight

NCBI National Center for Biotechnology Information

MS/MS Tandem mass spectrometry

GUI Graphical user interface

2-DE Two-dimensional gel electrophoresis

Chapter 1

Introduction

Bioinformatics is an interdisciplinary field that relies on mathematics, computer science and biochemistry to analyze biological data. It provides a set of practical tools and methods to biologists for studying and analyzing these data very effectively. For example, DNA, RNA, and protein sequences are, usually, massive, and manual processing is very time-consuming and error-prone, if not impossible. Therefore, computer-based solutions are extremely necessary to perform such tasks.

A Protein can be defined as a large molecule made up of amino acids. Linear strings of amino acids in each protein are arranged in a specific way that allows it to fold into a certain shape which in turn determines its function [22]. The primary elements of amino acids are carbon, hydrogen, oxygen, and nitrogen. Some other secondary elements may be found in the side chains of the amino acids string. Figure 1.1 illustrates the amino acids' structure while Table 1.1 shows a list of standard amino acid codes.

In practice, proteins consist of very long amino acid chains. They can be digested (cut) into smaller fragments (peptides) by proteolytic enzymes, some of which have well defined digestion patterns. For example, the enzyme trypsin is commonly used in MS experiments because of its features: specificity, availability, and low cost [46]. Usually, peptides contain a sequence ranging from 2 to 50 amino acids. This range allows the peptide to be analyzed and identified easily and effectively by MS, using one of several available methods.

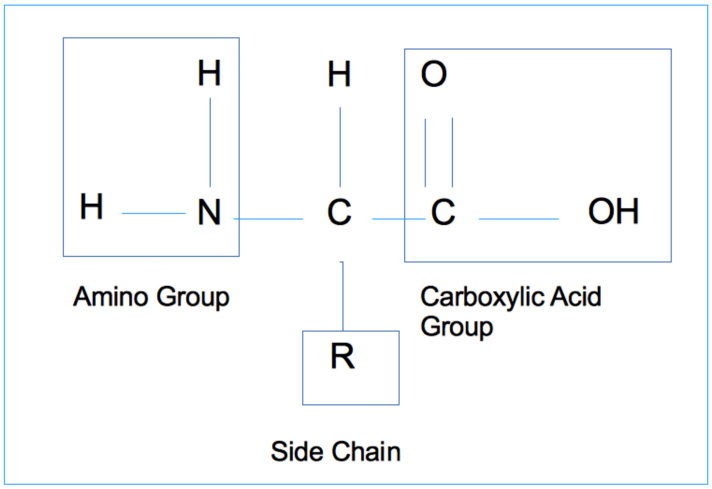


Figure 1.1: The chemical structure of amino acids. The primary structure is read from the N-terminal to the C-terminal. Each amino acid has a different structure in its side chain (R group). Amino acids all have a carboxylic acid on one end of the main carbon chain and an amine group on the very next carbon atom in the chain.

There are two major methods used in the protein identification process: peptide mass fingerprinting (PMF) [22], which is the focus of this thesis, and tandem mass spectrometry (MS/MS) [26].

Table 1.1: Amino acid codes. The first column contains the name of the amino acid, the second and third columns contain the corresponding amino acid code with three and one letters respectively.

Amino-Acid	**3-Letter**	**1-Letter**	**Amino-Acid**	**3-Letter**	**1-Letter**
Alanine	Ala	A	Arginine	Arg	R
Asparagine	Asn	N	Aspartic acid	Asp	D
Cysteine	Cys	C	Glutamic	Glu	E
Glutamine	Gln	Q	Glycine	Gly	G
Histidine	His	H	Isoleucine	Ile	I
Leucine	Leu	L	Lysine	Lys	K
Methionine	Met	M	Phenylalanine	Phe	F
Proline	Pro	P	Serine	Ser	S
Threonine	Thr	T	Tryptophan	Trp	W
Tyrosine	Tyr	Y	Valine	Val	V
Selenocysteine	Sec	U	Pyrrolysine	Pyl	O

1.1 Overview

A biological background is presented in the rest of this chapter. Chapter 2 presents a literature review of methodologies and tools used in PMF. The matching and scoring approaches are described in this chapter. In Chapter 3, the preprocessing of the input data is described. This includes the methodologies used in handling ambiguous amino acids in protein databases. The proposed method is presented in Chapter 4. Chapter 5 presents and discusses the experimental results. The conclusion and the future work are covered in Chapter 6.

1.2 Motivation

This work focuses on creating a computational tool for protein identification by mass spectrometry (MS) using peptide mass fingerprinting (PMF) [22]. There are many existing tools in the web (Section 1.13) to perform PMF but, most of them have the following limitations:

- In order to have a locally installed copy of the tool, the client must pay an expensive license. Many laboratories cannot afford it.
- The query space is restricted to public protein databases, only.

The aim of this work is to provide an efficient, user-friendly, and reliable software tool (Appendix A, B) for the identification of protein samples generated with PMF. That allows biologist to have a **free tool** for local search of **any** protein database.

1.3 Protein Identification

In proteomics, protein identification is the process of defining the probable primary sequence of an experimental sample protein by relating it to a specific database protein sequence. A protein can be identified from its peptide composition after digestion into fragments, *i.e.* search in databases for proteins whose peptide compositions (masses) are closest to the peptide compositions of the given experimental protein. In general, this process is based on several phases: extraction, separation, digestion, mass spectrometry, matching, and score calculation. In the

extraction phase, samples are extracted from a certain organism. These samples are separated by using specific techniques, *e.g.* Two-dimensional gel electrophoresis (2-DE) [34]. Separated proteins are digested by specific proteases into peptides and mass spectrometry is then performed for each unknown (separated/ isolated) protein to produce a mass spectrum. Finally, these peptide masses are searched against a database of theoretical proteins by matching and using a scoring algorithm to find the best identification of the input masses [47]. In this phase, the design of the scoring method determines the quality of the identification. It is worth pointing out that, the steps up to Mass Spectrometry are experimentally performed.

1.4 Peptide Mass Fingerprinting

PMF is one of the most important and widely used methods for protein identification using mass spectrometry [45, 48, 21]. The following is a brief explanation of how PMF works to identify the protein from a database. It involves the following steps:

- Separate the proteins, *e.g.* using 2-D gel electrophoresis.
- Digest the separated proteins into peptides with an enzyme that cleaves specific amino acid bonds.
- Perform Mass Spectrometry (MS) analysis to determine peptide masses, usually Matrix assisted laser desorption/ionization Time-of-flight (MALDI-TOF). The resulting MS data makes up the experimental data (peak list).
- The experimental data is searched against *in silico* digested protein database entries:
  - Compare the peptide masses of each protein in the database with the peptide masses of the experimental protein. This involves the digital (*in silico*) digestion and peptide mass calculation of each protein in the database .
  - Calculate the scores and measure how well the experimental proteins match the theoretical proteins.

– Present the results in which the best hits (proteins with highest scores), along with their scores and significance, are shown.

Figure 1.2 illustrates the steps needed for protein identification using PMF.

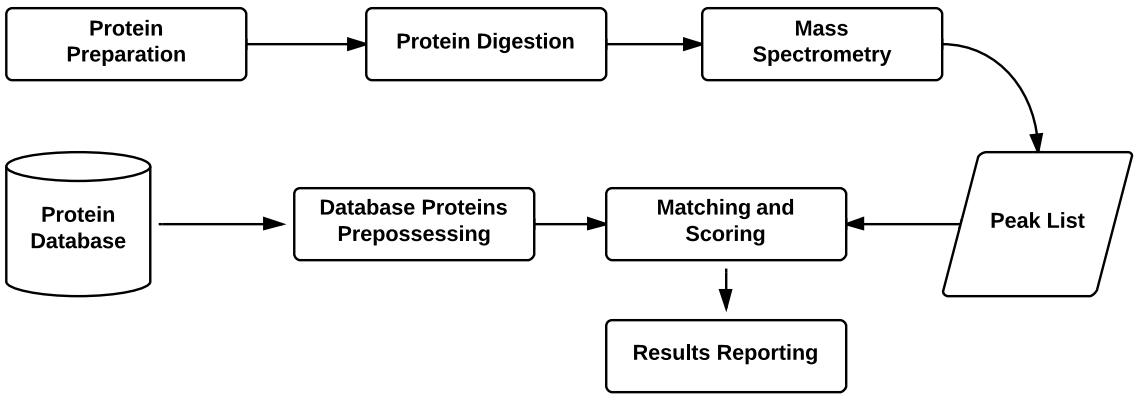


Figure 1.2: PMF steps flowchart.

The most significant parameters in PMF that can affect the results of identification are: the number of matched masses, mass error threshold, percentage of matched masses that covers the full sequence, post-translation modification, and the number of missed cleavages (Section 1.6) [41].

1.4.1 Problems in PMF

Protein identification using PMF has some constraints [1]:

- The experimental protein should exist in the search database. A new or modified protein may not be identified (correctly).
- Large **proteins** in the database have more peptides than smaller ones. Therefore, the probability of large proteins to match the experimental peptides will increase.
- Smaller **peptides** in the database have higher chance to match experimental peptides when compared to larger peptides.

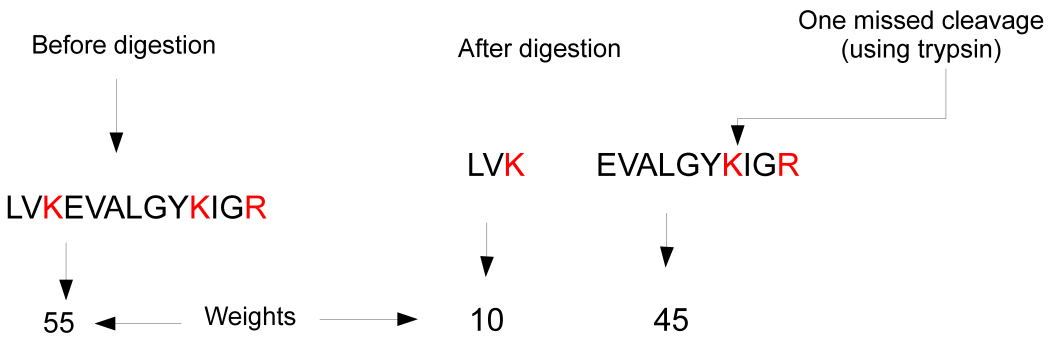
However, many new software packages have developed an advanced statistics and probability based scoring to overcome these problems.

1.5 Protein Preparation

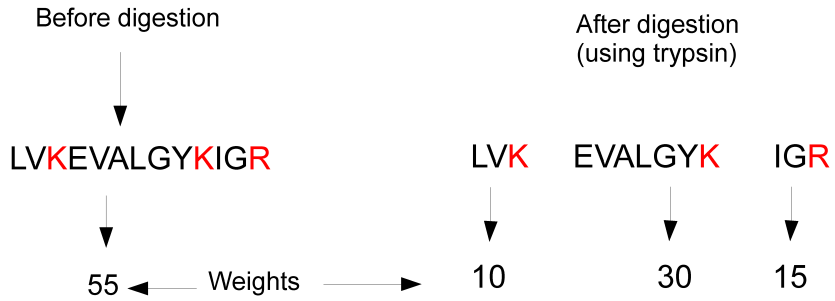
Sample preparation involves removing the contaminants and reducing the complexity of a protein sample [29]. Careful preparation is very important for performing mass spectrometry successfully. To achieve this task, tools like sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) or Reversed-Phase liquid chromatography (LC) can be used. These tools may be considered as an interface between biology and MS. They are used to purify proteins before they undergo MS, by cleaning, separating, quantifying, and assessing the post-translational modification (PTM) [25].

1.6 Protein Digestion

Digestion is the task of cutting the protein into peptides by using a specific enzyme. The most common enzyme is Trypsin, because it produces a cleavage with high specificity, availability, and low cost [46]. It converts the proteins into peptides by cleaving them at the carboxylic side of Arginine (K) and Lysine (R) residues [47, 33]. It is important to carefully perform this task, because missed cleavages make it difficult to successfully identify the protein. Experimental proteins are digested in a natural biological process. That process may fail sometimes *i.e.* one or more cleavages may be missed. In this case, the protein will contain fewer peptides and consequently the weights (resulting from the MS spectra) of these peptides will be affected. On the other hand, digital digestion for theoretical proteins (database proteins), can be done without any flaw. Therefore, computational tools should include a parameter to simulate the missed cleavages allowing the enhancement of the matching process. Figures 1.3a and 1.3b demonstrate a perfect digestion and a digestion with some missing cleavages.



(a) Incomplete digestion. Trypsin is used, it cleaves the amino acids chain after Arginine (K) and Lysine (R) in red color. In the figure one cleavage site remains after digestion.



(b) Complete digestion. Trypsin is used, it cleaves amino acids chain after Arginine (K) and Lysine (R) in red color. No missed cleavage is shown in the figure.

Figure 1.3: Digestion types.

1.7 Mass Spectrometry (MS)

Mass Spectrometry of a gel-separated protein is a protein identification technology that plays a major role in analyzing biological samples [39]. MS produces a spectrum of peptide sample masses by ionizing the digested protein sample and separating the resulting ions according to their mass-to-charge ratio, known as *m/z*, where *m* is the mass of the ion in Daltons, and *z* is the charge of the ion [32]. These ions, produced by mechanisms like Matrix assisted laser desorption/ionization (MALDI) [42] and electrospray ionization (ESI) are separated (*e.g.* by Time-of-flight MS analysers) to produce a mass spectrum. [47]. These two techniques are the core of MS and are usually implemented as high-throughput techniques.

1.7.1 MALDI-TOF

MALDI is a soft ionization technique that uses a short laser pulse of nitrogen gas instead of continuous laser to ionize molecules in a matrix [20, 24]. It is useful in protein identification because it is suitable for determining the mass of the intact peptide. These molecules (protein and peptide) are easily broken and tend to fragment when ionized by other ionization techniques. Furthermore, MALDI has other features that make it the first choice when it comes to protein study [23]:

- It requires relatively less intense sample preparation.
- Its matrix has resistance to the interferences caused by salts and detergent.
- It facilitates the data interpretation by producing peptides containing only one charge and shows only one peak in spectrum.

MALDI is attached to a time of flight (TOF) analyzer which calculates the time that the molecules take to move a fixed distance.

How MALDI works: It screens the peptide masses that are tryptic digested. The protein or peptide is placed on a target plate and merged with an appropriate matrix on this plate. The mixture of protein or peptide sample and matrix are crystallized and then irradiated in vacuum environment with a short laser pulse, which leads to release of matrix, and sample ions from the plate. The ions are then accelerated in TOF analyzer [37].

1.8 Problems Associated with Biological Processing

Protein identification is susceptible to contaminants and modifications during biological processing. Consequently, identifying the correct protein becomes more difficult. The following subsections explain some contaminants and modifications that may occur in protein samples.

1.8.1 Protein Contaminants

The MS masses of the protein sequences may include masses of contaminants. When contaminant masses are used in database search, it increases both false positive and false negative results. Therefore, when the contaminants are identified and removed from the protein sample, the probability of getting an accurate match will be increased. Possible contaminants can originate from:

- Keratin, which is a common hair and skin protein contaminant.
- Protease used in digestion *e.g.* trypsin.
- Sample chemicals.
- MALDI matrix and the electrophoresis components.

Despite the unknown identity of some contaminants, they can be observed in a large number of samples. Furthermore, when the same masses exist in many samples, it is a good indicator that those masses come from contaminants [9, 52]. It is important to remove any mass related to Keratin if the sample is not human, because Keratin is abundant in human hair and skin. MS masses should be as clean as possible before starting the database search.

1.8.2 Post-Translational Modification (PTM)

Post-translational modifications are steps in the protein generation process in which the protein may undergo cleavage, extension, and other processes, including chemical modifications on some amino acids. It essentially affects the protein function due to the changes made to its chemical structure. Identification, characterization and mapping of these modifications is critical for understanding the function of proteins in a biological context [27, 10].

1.9 Protein Sequence File Format

Because of the fast growth of biological data (and databases), manual processing of sequences becomes very laborious and error-prone. Accordingly, more flexible, efficient, faster, and easier

processing tools become necessary. Automated tools have been developed to confront this massive growth of data. These tools process protein sequences automatically using algorithms, methods, and computer programs to organize and display them in a clear and understandable format [38, 30]. Protein sequences can be found in different formats, such as FASTA, GCG, and plain text, depending on the database they belong to. Nevertheless, the most popular format of these is still FASTA. Figure 1.4 shows an example of an entry in a FASTA file, which is considered in this thesis. This kind of file starts with one description line followed by lines of sequence data.

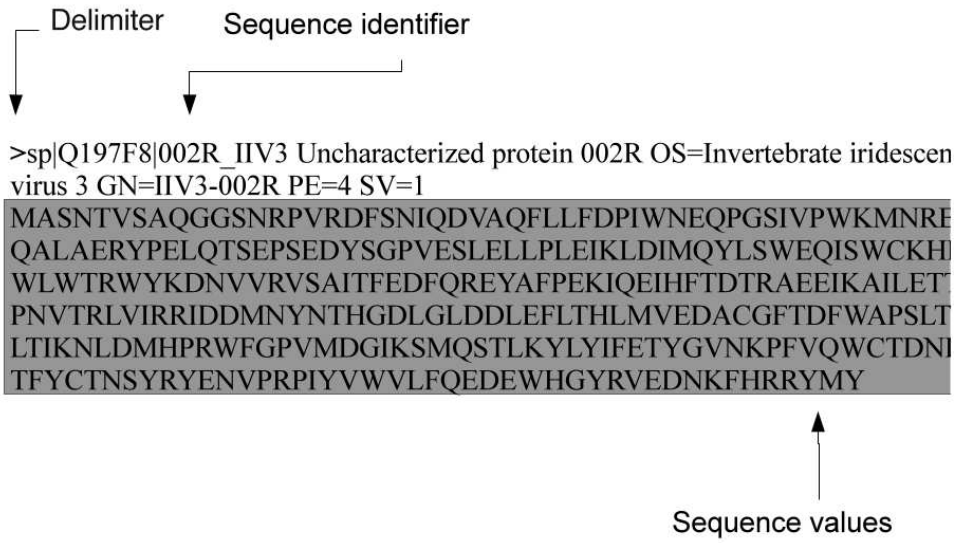


Figure 1.4: FASTA format for one protein sequence. The first character of the description line is the greater-than (>) symbol. The number of sequences in the input data is determined by the number of lines beginning with a '>'.

1.10 Protein Databases

Biological databases, either locally stored or published online, store a cumulative genetic knowledge about sequences, structures, and functions of biological data. The more reliable, complete, and well organized the database is, the better the results obtained through searching and matching processes [18]. Nowadays, there are many on-line databases that supply an essential support for protein identification. These databases not only store a series of protein sequences, but also contain annotation information for these sequences. However, databases may have limitations

that result in bad matching between experimental data and theoretical data, such as incomplete data, particularly where non-model organisms are being studied, for which public data are unavailable [33]. some popular protein databases are listed in Table 1.2.

Table 1.2: Protein databases.

Database	Features	Website
GenBank (NCBI)	Has an annotation system Has Blast algorithm It is a part of the International Nucleotide Sequence Database Collaboration	http://www.ncbi.nlm.nih.gov/genbank/
Protein Identification Resource (PIR)	Has an annotation system	http://pir.georgetown.edu/
UniProtKB/Swiss-Prot	Has annotation system	http://web.expasy.org/docs/swiss-prot

1.11 General Search Parameters

Almost all tools that perform PMF share the same general parameters. These are listed as follows:

- Enzyme: It is used to cut the protein into fragments. Many enzymes are available to perform the digestion. A good digestion does not cut the protein into very small peptides because it would result in a lot of random matches. It is usually better to have long peptides to get more specific results. Trypsin is the most popular enzyme used in the digestion process due to this nice property.

- Missed Cleavages: Sometimes, in the experimental digestion, some sites (usually one or two) are missed. This parameter is a positive integer value that represents the maximum allowed number of missed cleavage sites during the digital digestion.

- Mass Tolerance: It is a user defined threshold that represents the acceptable difference between the experimental masses and theoretical masses through the comparison step [6]. It may be entered as:

1. % : Fraction represented as percentage.
2. *mmu* : Milli mass unit.
3. *ppm* : Fraction represented as parts per million.
4. *Da* : Abbreviation for Dalton. It is the absolute unit of mass.

- Fixed Modifications: A list of fixed modifications that may be selected by user. This parameter represents the chemical modifications known to occur in the sample. A common fixed modification is carboxymethyl (C).

- Variable Modifications: A list of variable modifications that may or may not occur in the sample. Therefore, the masses for each modified symbol are calculated twice, with modification and without modification. A common variable modification is oxidation (M).

- Query (peak list): Consists of either a data file, or a list of peptide mass values and respective intensities (optional) typed in a query window. If the intensities of the masses are supplied, some tools can use them to get a better score by selecting the mass values with higher intensities.

1.12 Scoring Methods

The core of any protein identification process is its scoring method. Its quality determines the efficiency of the identification method. Old scoring methods were based on the number of matched masses (sample masses and database masses) [36]. This kind of scoring method is used in PeptIdent [3]. It was sufficient for PMF protein identification several years ago, but no longer due to the increase of database sizes. Most of new scoring algorithms are based on statistics and probabilities because of the advantages these systems provide. For example MOWSE, ProFound, Ms-Fit, and Mascot tools are based on statistical frameworks.

1.13 Popular Tools

In the last few years, proteomics has rapidly grown along with the development of protein identification and quantification techniques. Some of these techniques are new, while others are based upon older ones. Protein identification software effectively contribute to the study and exploration of information from 2-D gels using mass spectrometry. Recently, many existing tools such as ProFound [54], PeptideMass [16] , Mascot [13], and Ms-Fit [7] were developed. Some of these tools do not perform the complete protein identification pipeline, while others may perform the whole steps needed for protein identification. For example, PeptideMass just cleaves the a protein sequence into peptides with the specified enzyme and reports the masses of these peptides. On the other extreme, Mascot can perform the whole process needed for identification. It sequentially performs the following steps:

1. Cleave database protein into peptides.
2. Calculate the masses for the resulting peptides.
3. Apply the comparison between experimental masses and the masses resulting from the previous step.
4. Calculate scores to determine the accuracy of the matching.
5. Show the best protein hits and respective scores, based in the scoring.

Chapter 2

Related Work

PMF is commonly used in many biological laboratories around the world. Therefore, many existing PMF software packages became commercially successful. The key to the success of each software is its scoring method. A good scoring function takes into account several factors: mass tolerance, database size, number of missed cleavages, coverage of matched peptides, and number of variable modifications. These factors must be applied in such a way that maximizes the probability that the top-ranked database matches are true candidate proteins. At the same time, scores for non-matching proteins should be minimized. [47]. This chapter introduces a review of some popular tools and their scoring methods, if available.

2.1 MOWSE

MOWSE (MOlecular Weight SEarch) is an important scoring method in protein identification, used in PMF [51]. Several software packages now are built on this method, such as Ms-Fit and Mascot [36]. This method is based on the achievable matches between the theoretical proteins and the MS sample, and the occurrence of molecular weight for each theoretical peptide [47]. MOWSE takes into account some aspects like protein size and the frequency of each peptide in the database through scores calculation [3]. On the other hand, it does not provide a confidence measure for these scores. Because MOWSE will serve as the starting point of a method proposed in this work, it will be described in Section 4.2.1 in detail.

2.1.1 Limitations

MOWSE is a scoring method that has the following limitations:

- Can not filter random matches sufficiently.
- There is no contaminant removal mechanism.

2.2 Mascot

It is one of the most popular application for protein identification using mass spectrometry data. It performs both protein identification techniques: MS/MS and PMF. Mascot can be freely accessed at the Mascot server (http://www.matrixscience.com/server.html) but with limitations. The complete application's features, such as data size, confidential issues, and dealing with enzymes and modifications can only be accessed with the commercial version. Furthermore, the probability model details are not published and publicly unknown. It is fast due to its multi processor ability [8]. Mascot uses the probability-based MOWSE scoring algorithm. However, Mascot and MOWSE differ in a couple of things. First, Mascot directly deals with the FASTA format instead of prebuilt indexes (used by MOWSE). Second, Mascot uses both MOWSE and probability in scoring [8], where the matching between experimental data and each theoretical sequence can be considered as a random event. Theoretical proteins that have random match to the experimental data, are then ranked with decreasing order of probability [13]. Figure 2.1 shows the Mascot form for peptide mass fingerprinting identification.

2.2.1 Mascot Scoring

The main technique of probability based scoring is to calculate the probability that the observed match between the MS masses and each database entry is a random event. The match with the lower probability is the match with the higher score. Using probabilities as final score may be confusing. Therefore, Mascot reports the scores as [53, 8]:

$$-10\log_{10}(\mathrm{Pr}) \tag{2.1}$$

where the Pr is the absolute probability.

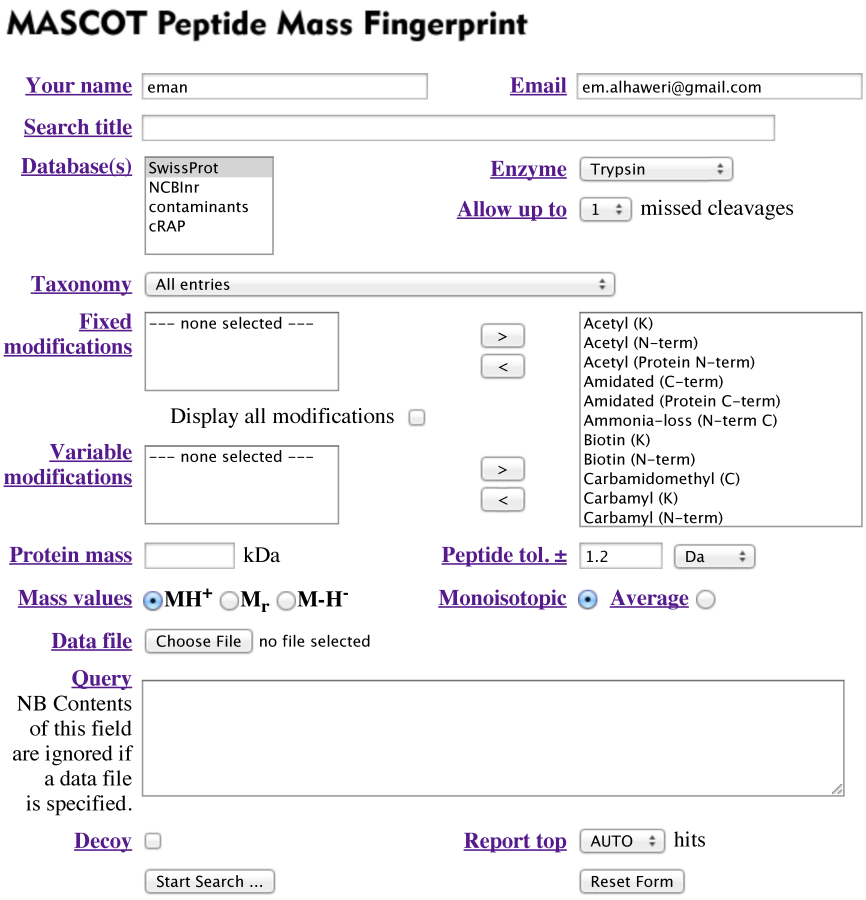


Figure 2.1: Mascot program form for peptide mass fingerprinting.

2.2.2 Mascot Score Significance

The difference between random and significance scores should be as high as possible to ensure confidence in the results. Identification reliability can be affected by the sequence coverage (SC), which can be defined as the proportion of the database protein covered by the query peptides, and the number of matched mass values (MM). To obtain more accurate results, the multiplication of SC by MM is used to reduce the random matches as much as possible [51, 53].

2.2.3 Limitations

Mascot has the following limitations:

- The licensed version restricts the search to a limited number of databases.

- It does not provide a free standalone software to handle large-scale PMF spectra.

2.3 ProFound

ProFound is another tool to identify proteins by PMF. It uses a Bayesian framework as the core of the scoring method to rank the theoretical sequences depending on their probability of occurrence. This tool takes into account the protein's properties to increase the accuracy of the scoring [13]. It identifies proteins even if the quality of database's proteins is fairly low. ProFound computes the probability of an experimental sample to be a specific protein in the database, based on the MS protein (protein sample resulting from MS) and any information related to it, *i.e* mass accuracy, previous experiments on this protein, and the enzyme used for the digestion. Likewise, it uses the available properties of the database proteins to restrict the search space. This restriction decreases the amount of random matches which consequently increases the confidence of the identification [54]. Figure 2.2 presents the ProFound form which displays the parameters needed by the program to perform the identification.

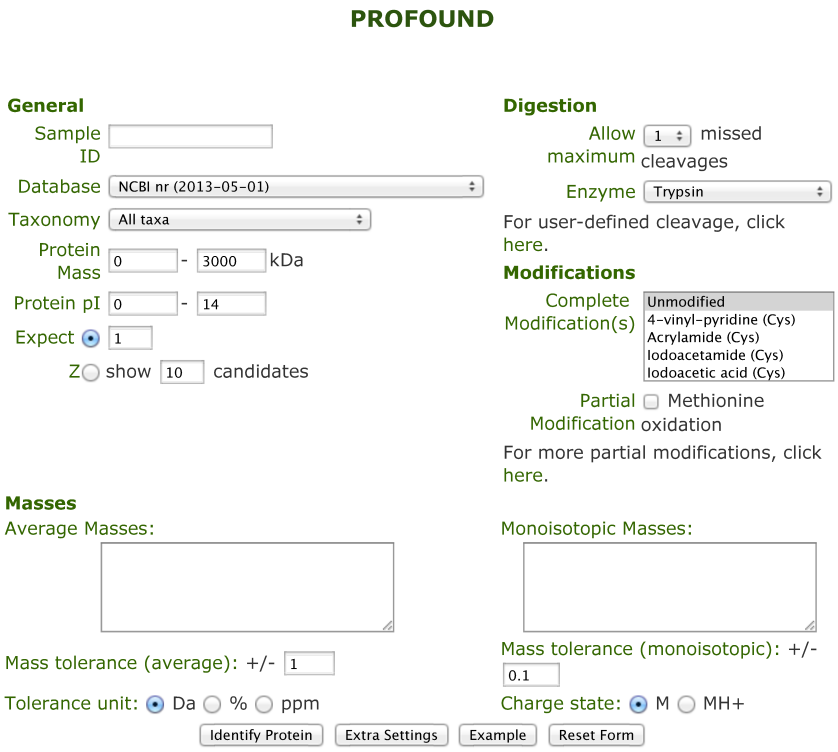


Figure 2.2: ProFound program form.

2.3.1 ProFound Scoring

ProFound considers theoretical protein properties as well as other properties relevant to protein sample in its scoring. It ranks the theoretical proteins using a Bayesian algorithm to obtain a reasonable inference when identifying a protein sample. ProFound introduces the hypothesis: let P_{th} be the theoretical protein in the database, P_{ex} be the experimental protein generated by MS, and K be the available information. Thus, the probability that the theoretical protein P_{th} is the intended protein given the experimental protein P_{ex} and the background information K, is calculated by the approximated equation:

$$\Pr(P_{\text{th}}|P_{\text{ex}}K) \sim \Pr(P_{\text{th}}|K)\left(\sqrt{\frac{2}{\pi}}\frac{\text{m}_{\text{max}}-\text{m}_{\text{min}}}{|\Psi(P_{\text{th}})|}\right)^{|\hat{\Psi}|} \times \prod_{i=1}^{|\hat{\Psi}|}\frac{1}{\sigma_i}\left\{\sum_{j=1}^{c_i} \text{e}^{\left[-\frac{(\text{m}_i-\text{m}_{ij})^2}{2\sigma_i^2}\right]}\right\}\text{F}_{\text{pattern}} \tag{2.2}$$

with the normalized condition: $\sum_{P_{\text{th}}\in\text{database}} \Pr(P_{\text{th}}|P_{\text{ex}}K) = 1$

where:

- $\Pr(P_{\text{th}}|\text{K})$ is the probability that the P_{th} given the background information K is the protein sample.
- $|\Psi(P_{\text{th}})|$ is the number of peptides resulting from the theoretical protein P_{th}.
- $|\hat{\Psi}|$ is the number of matches between P_{th} and P_{ex}.
- $\text{m}_{\text{max}} - \text{m}_{\text{min}}$ is the range of measured peptide masses.
- m_i is the measured mass of the ith match.
- c_i is the number of theoretical peptides that match m_i.
- m_{ij} is the calculated mass of the jth peptide in the ith match.
- σ_i is the standard deviation of the mass measurement at the mass m_i.
- $\text{F}_{\text{pattern}}$ is an empirical term. This term increases the probability when adjacent peptides are found or/and when overlapping occurs.

In summary, the theoretical protein P_{th} probability increases when the number of matches $|\hat{\Psi}|$, and the mass accuracy (small σ_i, and $m_i - m_{ij}$ values) increase, and when the number of fragments $|\Psi|$ decrease [54].

2.3.2 ProFound Confidence

The information related to the protein helps ProFound to get more accurate results. The Bayesian algorithm combines the information to restrict the database search process and to reduce the occurrence of peptides that give random matches. This increases confidence in ProFound scores [54]. In fact, probability is the main factor to get good results in ProFound. When the probability increases, the confidence level will increase and vice-versa. This tool can be accessed by Expasy server: http://prowl.rockefeller.edu.

2.3.3 Limitations

The restrictions of ProFound can be listed as:

- It only works with the NCBI database.
- There is no contaminant removal mechanism.
- It does not provide a free standalone software to handle large-scale PMF spectra.

2.4 MS-Fit

Ms-Fit is considered one of the most popular PMF softwares along with Mascot. Ms-Fit as well as Mascot, uses MOWSE to calculate scores, besides offering additional options to restrict and enhance the search [36, 7]. These options allow the user to control some parameters like the range of protein molecular weights and the minimum number of matched peptides. These parameters help to speed up searches and improve accuracy. Figure 2.3 shows the Ms-Fit form, which can be accessed on the Ms-Fit server (http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msfitstandard).

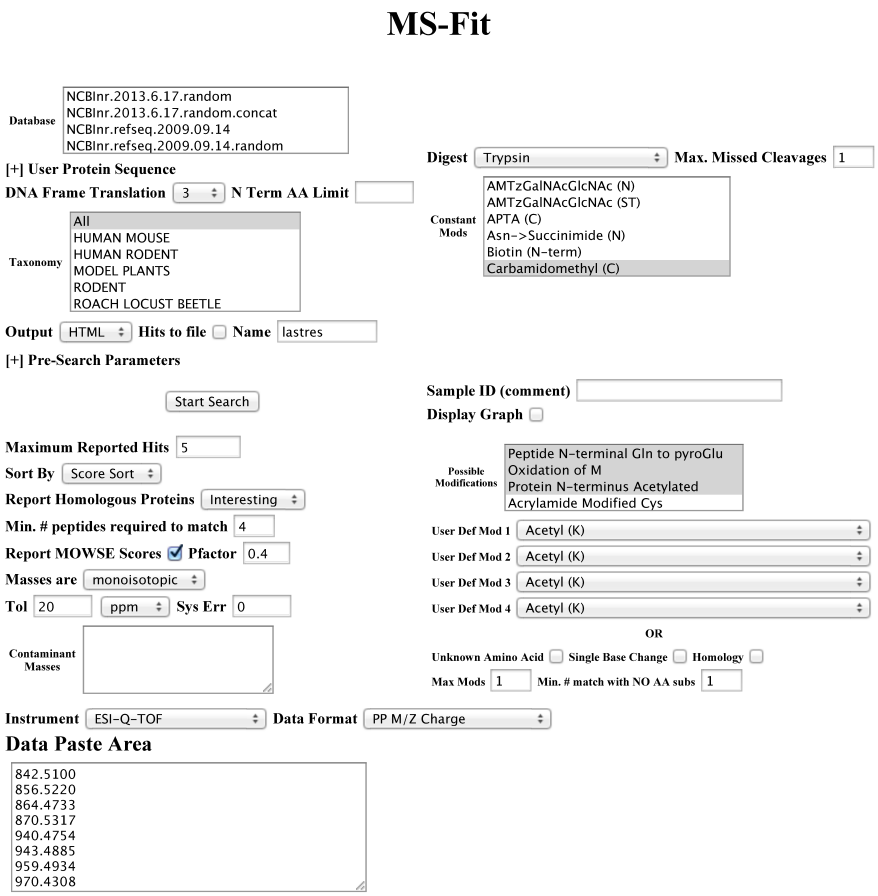


Figure 2.3: MS-Fit program form for peptide mass fingerprinting.

2.4.1 Limitations

Ms-Fit limitations are the following:

- Restricts searches to only some databases.
- It does not provide a free standalone software to handle large-scale PMF spectra.

2.5 ProteinDecision

To get the advantages of statistical properties and to handle these properties in PMF, probability based scoring is widely used in most protein identification methods.

ProteinDecision [49] is based on MOWSE and uses probability-based scoring. This approach starts by constructing a frequency table, which MOWSE and Mascot are also built on, for all

theoretical proteins. When a theoretical protein is selected to be preprocessed, this protein is digested and the peptides are mapped to the frequency table, based on the molecular weight of the peptides and the molecular weight of the respective protein. The masses of these peptides are compared to the experimental data, given a mass tolerance, to find matches. Formally, let Ψ_i be the set of peptides in the row i. Additionally, let Ω_j be the set of proteins in the column j. Given an experimental protein P_{ex}, a theoretical protein $P_{\text{th}} \in \Omega_j$, and a peptide $\psi \in \Psi_i$, the probability of ψ matching an MS mass is:

$$\Pr_1(\psi|P_{\text{ex}}) = \frac{m_{ij}}{M_j} \tag{2.3}$$

Where the m_{ij} is the number of peptides in cell ij (in frequency table) for Ω_j, and $M_j = \sum_{i=1}^{N} m_{ij}$ where N is the number of rows in the frequency table. For specificity, the probability of ψ not being from the cell ij is:

$$\Pr_2(\psi|P_{\text{ex}}) = 1 - \frac{m_{ij}}{M_j} \tag{2.4}$$

And the probability that at least one peptide from the protein P_{th} exists in cell ij is given by:

$$\Pr_3(\psi|P_{\text{ex}}) = 1 - \Pr_2^{|\Psi(P_{\text{th}})|} \tag{2.5}$$

where $|.|$ is the set cardinality, and $|\Psi(P_{\text{th}})|$ denotes the set of peptides in the cell ij for P_{th} in the frequency table. Taking into account the assumption that the peptides from the theoretical proteins are independent, the joint probability $\Pr(P_{\text{th}}|P_{\text{ex}})$ is given by:

$$\Pr(P_{\text{th}}|P_{\text{ex}}) = \prod_{i=R(l),l\in|\hat{\Psi}|} [1 - (1 - \frac{m_{ij}}{M_j})^{|\Psi(P_{\text{th}})|}] \tag{2.6}$$

where the $\Pr(P_{\text{th}})$ is the probability of the theoretical protein P_{th} matching the experimental protein, $|\hat{\Psi}|$ is the number of theoretical protein masses that match to experimental masses, and $R(l)$ is the row number where the MS mass is found. This equation presents the probability for a match between experimental protein P_{ex} and theoretical protein P_{th}. The higher the value of $\Pr(P_{\text{th}}|P_{\text{ex}})$, the lower the score value of this protein P_{th}.

ProteinDecision can be accessed by the website: http://digbio.missouri.edu/ProteinDecision.

2.5.1 Limitations

ProteinDecision has the following drawbacks::

- There is no missed cleavages features.
- Post-translational modifications feature is not available .
- There is no contaminant removal mechanism.

2.6 Statistical Assessment for Mass-spec Using PMF

This method drives a new statistical model for confidence assessment of the results using PMF. Like other tools, the method starts the preprocessing of the theoretical proteins by digesting them into peptides. For the experimental data, mass/intensity values are read from MS spectra. This method uses the ratio $(|\psi_j - \psi_i|)/\psi_i$, to decide whether the theoretical peptide ψ_j is a candidate. The model uses the value 10^{-4} as mass tolerance so, if the result of $(|\psi_j - \psi_i|)/\psi_i$ $\leq 10^{-4}$, the theoretical peptide is ranked as a candidate peptide. The raw score S for protein P_{th} whose peptides are Ψ_j is calculated by:

$$S = (\sum_i I(\psi_i)/\left|\psi_i - \psi_j + d\right|) * [\text{TP}/(1+\text{FP})] \tag{2.7}$$

Where the $I(\psi_i)$ is the intensity of the experimental mass i, ψ_j is the mass of peptide j of the theoretical protein P_{th}, ψ_i is the mass for the experimental peptide i in the protein sample P_{ex}, and d a constant for the optimization. TP represents the number of true positive MS masses matches, and FP the false positives which is the number of MS masses not found in P_{th}. The higher the score S, the higher the chance of being the intended match.

This method also evaluates the statistical significance of the identified protein. All theoretical proteins are treated as statistical background to see whether the observed protein i is the query protein rather than just chance. A Q-score$_1$ is provided which is defined as a ratio between the number of residues for matched peptides N_i and the protein length L_i, or $\frac{\text{N}_i}{\text{L}i}$. The Q-score

is normalized to the ratio between the total number of residues of peptide matches N in the database and the total number of residues in the database L, or $\frac{\mathrm{N}}{\mathrm{L}}$.

The following is the transformation applied to Q-score:

$$\mathrm{Qscore}_2 = \mathrm{Qscore}_1 / \mathrm{mean}(\mathrm{Qscore}_1)$$
$$\mathrm{Qscore}_3 = \ln(\mathrm{Qscore}_2)$$
$$\mathrm{Qscore} = (\mathrm{Qscore}_3 - \min(\mathrm{Qscore}_3)) * 50$$

After calculating the Q-score for theoretical proteins, the method transforms Q-score for these proteins by applying a Gaussian distribution with observed mean μ and standard deviation σ. The protein is considered as a significant protein hit if its Q-score is larger than $(\mu + 2\sigma)$ [15].

2.6.1 Limitations

Limitations of these algorithm are:

- Post-translational modification is not available.
- Computation time for confidence assessment is long.
- There is no contaminant removal mechanism.

Chapter 3

Data Preprocessing

To achieve a successful database search, many steps should be performed beforehand. When these steps are accomplished in an efficient way, better results will be obtain by the PMF tool. However, when at least one prior step in the protein identification pipeline (Figure 1.4) is not performed properly, it will certainly affect the results. Some of these steps are already performed by biological instruments, so the quality of processing for these steps can not be handled by computational PMF tools. Nevertheless, these tools can filter the data generated by these instruments before doing the database search.

In this work, two main input files are handled: peak list and protein database. The peak list contains the peptide masses which are related to the unknown protein. The database contains the known proteins and is used as the database to query. This chapter explains the necessary preprocessing to have these files ready for matching and scoring. In addition, it addresses the problems caused by ambiguous amino acids in protein databases, as well as the proposed solutions.

3.1 Peak List Preprocessing

The peak list generated by MS contain both peptide masses, and contaminant masses. These contaminants hamper the protein identification process. Therefore, identifying and removing these masses are essential steps in obtaining good results in matching [28]. To achieve this, a

contaminant database containing all known contaminant sequences is used. These are initially converted to masses and then searched against the input peak list (query mass list) to find and remove them.

3.2 Protein Database Preprocessing

In many cases, some database sequences include symbols that do not encode amino acids, resulting in ambiguity in the protein sequence. In fact, there are three main cases in which these codes need preprocessing to make the sequences valid for digestion: the unknown symbol (X), ambiguous symbols (B, Z, and J).

3.2.1 Unknown Symbol Processing

A major problem that is occasionally found is that protein databases contain sequences with one or more unknown amino acids, coded as 'X'. A solution to this problem is to replace this unknown symbol by the set of valid symbols (Table 1.1). This will yield a set of "new" sequences with 'X' replaced each time. Each one of these new sequences is treated as a theoretical protein. This will increase the time of processing depending on the number of Xs found in the sequence.

3.2.1.1 'X' Replacement Approach 1

The main problem with 'X' symbol is that when it appears more than one time in the sequence, this significantly increases the number of generated sequences. For example, if the protein sequence contains a single 'X', it will be replaced with all standard amino acid codes. Each replacement generates a new protein sequence. Similarly, if the sequence contains two Xs, the number of generated sequences in this case will be 22^2. This increase in the number of sequences can be represented by the exponential function: $f(n) = 22^n$ where n is the number of unknowns in the sequence.

Essentially, the 'X' replacement strategy used in this work involves:

- Reading a protein sequence from the database.
- Check the sequence by ensuring that all its symbols are valid amino acid codes, if so, process the sequence. Otherwise, calculate the number of Xs
- For any sequence that contains Xs do the following:

  1. Get the number of Xs in the sequence.
  2. Get the positions of Xs in the sequence.
  3. For each 'X' occurrence, replace the 'X' with each of the 22 amino acids, in which each replacement produces a new sequence.

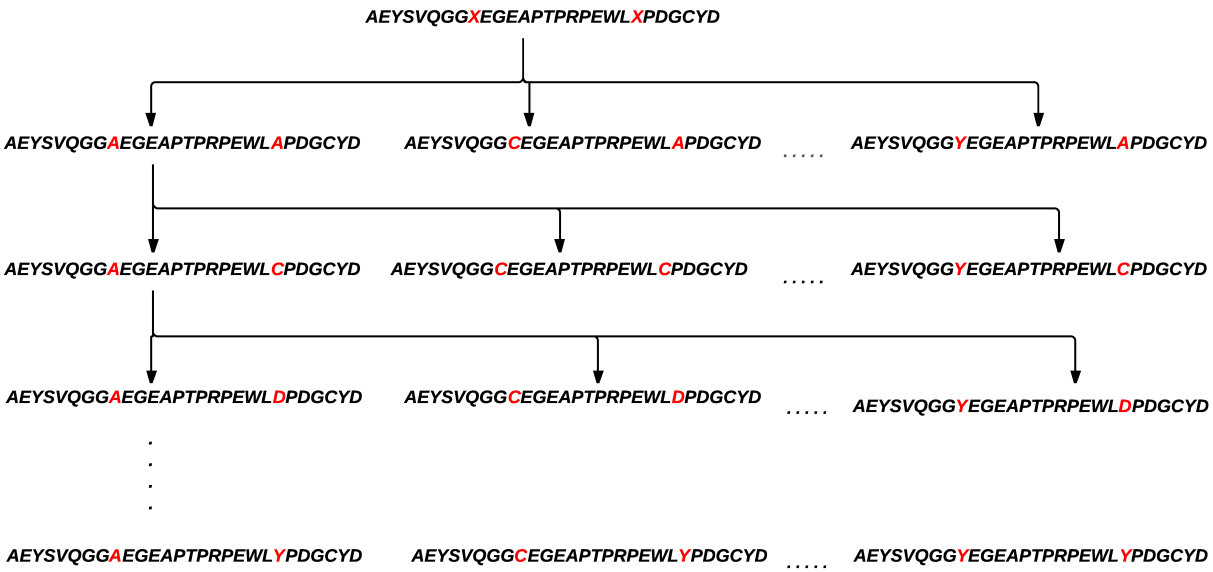


Figure 3.1: Replace Xs with the standard amino acids to get valid sequences.

3.2.1.2 'X' Replacement Approach-2

The approach presented above resulted in memory and speed problems. These problems arise from the high number of sequences produced by the X replacement. Another efficient and compact approach is described in Algorithm 3.2.1. The functions used in this algorithm are:

- GET_PROTEINS_SYMBOLS() that returns a list of valid amino acid symbols.

- GET_POSITIONS_OF_X(SEQUENCE) that returns a list of positions of Xs inside the sequence.
- DIGEST(SEQUENCE) that digests the sequence into peptides and returns these peptides.
- CALCULATEMASS(SEQUENCE) that calculates masses for peptides and returns these masses.

In summary, this algorithm performs the following main steps:

- Read the protein sequence and calculate how many Xs are in the sequence.
- If the sequence contains a single 'X' then, generate 22 sequences with X replaced with each amino acid.
- For sequences with more than one 'X', digest the sequence, and then assign the average mass of the amino acid for each 'X' occurrence, which is approximately 110Da [19].

Algorithm 3.2.1: REPLACEMENT_OF_X(*sequence*)

$.validSymbols[\,] \leftarrow$ GET_PROTEINS_SYMBOLS()
$xPositions[\,] \leftarrow$ GET_POSITIONS_OF_X(*sequence*)
if LENGTH(*xPositions*) $\neq 1$

then
- DIGEST(*sequence*)
- **for** $j \leftarrow 1$ **to** LENGTH(*sequence*)
  - **do**
    - **if** $sequence[j] =' X'$
      - **then** SET_MASS($'X', 110$)
- CALCULATEMASS(*sequence*)

else
- **for** $k \leftarrow 1$ **to** LENGTH($xPositions[\,]$)
  - **do**
    - **for** $i \leftarrow 0$ **to** LENGTH($validSymbols[\,]$) $- 1$
      - **do**
        - REPLACE(SYMBOLAT($xPositions[k]$), $validSymbol[i]$)
        - DIGEST(*sequence*)
        - CALCULATEMASS(*sequence*)

The generation of 22 sequences when 'X' is present one single time, causes another problem in the scoring step. MOWSE frequency table, as mentioned earlier, is created for all peptides in the database. This scoring algorithm depends on the occurrence of these peptides for each protein in the database. The 22 sequences generated, when replacing the 'X', are stored in the same column in the frequency table.

This happens because they are the same sequences with almost the same molecular weights except that the mass of the unknown symbol is replaced each time. When the same sequence with the same peptides (except the peptide containing 'X') is stored several times, it leads to a great increment of the number of occurrences in specific intervals. Consequently, this increase biases the scoring method, which mainly depends on the distribution of peptide entries in the database. In particular, when the number of occurrences of some peptides increases in specific intervals, these peptides will have more chance to be candidate peptides. As result, the score of their proteins will be high even if they are not the correct proteins.

To overcome this problem, a proposed solution to avoid the duplication of peptide masses for the same sequence is provided. Hence, peptides that remain unchanged are inserted into the table only once. In this case, only the mass of the peptide which contains 'X' will be inserted into the frequency table in each replacement.

3.2.2 Ambiguous Symbols Processing

Each of the ambiguous symbols 'B', 'Z', and 'J' has two standard codes related to it as shown in Table 3.1. These kind of symbols (like 'X') cause exponential growth (2^n) in the number of sequences when they appear in one sequence. Nevertheless, they differ from 'X' in the number of replacements. In 'X', there are 22 replacements for each occurrence of the 'X', while these symbol have 2 replacements each time they appear in the sequence.

Because of that, they can be processed in two ways depending on their number inside the sequence. If any or all of these symbols appear just one time, each one is replaced by its two valid codes. For example, if the sequence contains just one symbol 'Z', the application will generate two new sequences one with Glutamine (Q) and one with Glutamic acid (E) instead of 'Z'. Accordingly, If the sequence contains one 'B' and one 'Z', the number of generated sequences will be 4. Figure 3.2 shows an example of how the substitution process occurs for a sequence with 'X', 'Z', and 'B' respectively. If the sequence contains more than one ambiguous symbol, the digestion is applied first to the sequence as valid sequences, and then during the mass calculation the masses for the symbols are set to the average mass of their two corresponding possible masses [44].

Table 3.1: Ambiguous and unknown symbols and corresponding one and three letters codes.

Amnio Acid	**3- Letters**	**1- Letter**
Asparagine or Aspartic acid	Asx	B
Glutamine or Alutamic acid	Glx	Z
Leucine or Isoleucine	Xle	J
Unknown amino acid	Xaa	X

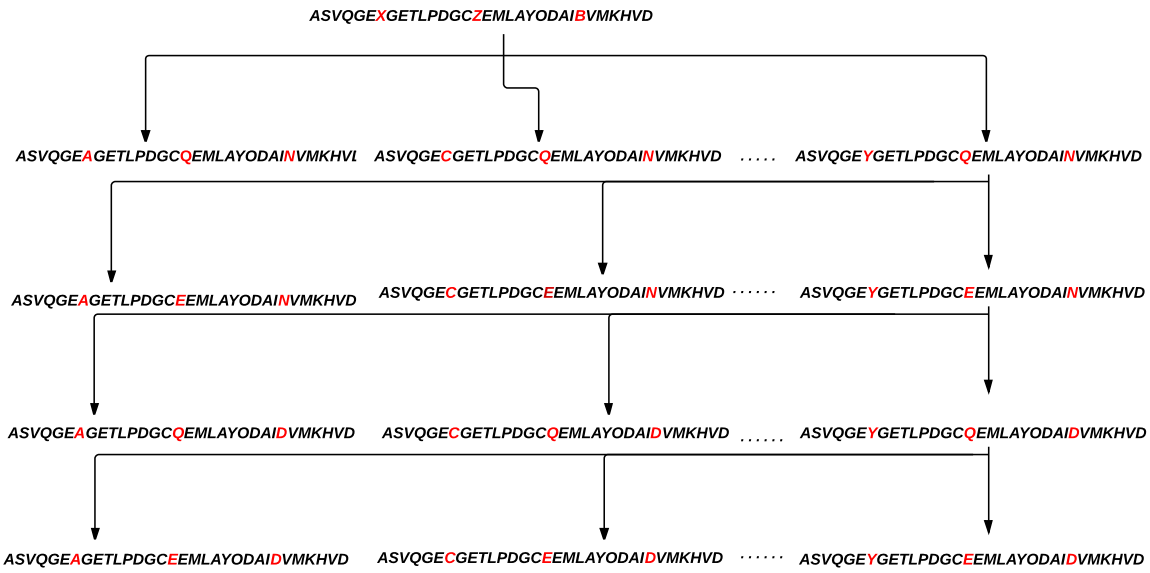


Figure 3.2: Replacement process for invalid codes 'X', 'Z', and 'B' with their corresponding standard amino acid codes to get valid sequences

3.2.3 Database Size Handling

Biological databases tend to be massive, which can be a major problem. In this work, the input database is processed, sequence by sequence, to handle this problem. Thus, instead of reading the entire database to the memory and processing it, one sequence is loaded and processed at a time in an in-process-out paradigm. This loading is optimized by buffered file reading. In this case, any database size can be handled efficiently. Figure 3.3 shows how one sequence is processed at time.

3.2.4 Digestion

In-silico digestion is a simulation of the experimental digestion performed by mass spectrometry. Therefore, simulated digestion should result in peptides that are as similar to the mass spectrometry produced ones as possible. To achieve that, PMF software use parameters such as the enzyme used for digestion, number of missed cleavages, which are the factors that affect experimental digestion. It is very important to choose these values very carefully to obtain the

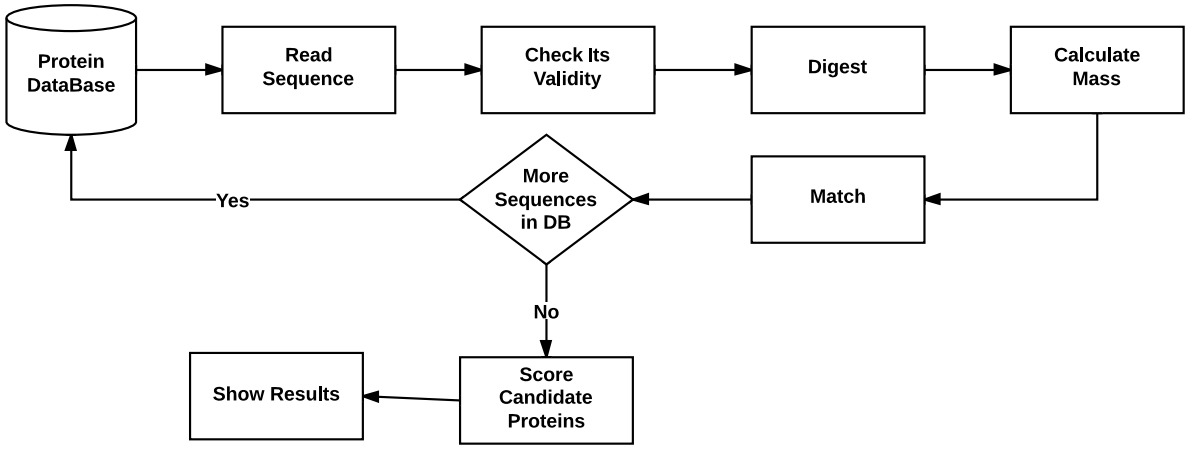


Figure 3.3: Processing the database proteins one-by-one.

right peptides. A brief explanation of these parameters is presented below.

Enzyme: It is set as the one used in the experimental digestion. An explanation of how digestion works is presented in Section 1.6.

Missed Cleavages: This parameter is used due to the imperfect nature of biological digestion. If the user is very confident that the experimental protein was perfectly digested, this parameter is set to zero which means no missed cleavages are expected. However, this parameter is commonly set to one due to the difficulty of getting a perfect digestion.

3.2.5 Mass Calculations

The masses of digested peptides are calculated in this step. Each peptide has a weight, depending on its amino acid composition. When computing the mass for a peptide (based on the digestion), the sum of weights for the amino acids contained in this peptide is calculated. So, by knowing the mass of each amino acid residue in a sequence (see Table 3.2), one can calculate the mass of any peptide sequence by the following formula:

$$\mathrm{MW} = \sum_{i=1}^{\mathrm{N}} \mathrm{m}_i + 18.01524 \tag{3.1}$$

Where the MW is the molecular weight of the sequence, m_i is the mass of the amino acid i inside the sequence, N is the number of amino acid residues in the sequence, and the constant 18.01524 is the average mass of the water molecule (H_2O) which is H from the amino (-NH2) group and -OH from the carboxyl (-COOH) group.

In fact, there are two ways to calculate amino acid masses for peptides: Average mass and monoisotopic mass. The difference between these two depends on the different compositions of isotopes in the amino acid chemical structure. Average mass is calculated using the average mass of the isotopes for each element weighed for natural abundance. On the other hand, monoisotopic mass is calculated using the mass of the most abundant isotope of each element present in the molecule [35]. The proposed application allows the user to choose the type of peptide mass, either average or monoisotopic mass. Table 3.2 lists these two types of masses for the 20 amino acid residues.

Table 3.2: Amino acid residues and their Monoisotopic and Average masses.

Amino Acid	**Monoisotopic Mass**	**Average Mass**
Glycine	57.02147	57.052
Alanine	71.03712	71.079
Serine	87.03203	87.078
Proline	97.05277	97.117
Valine	99.06842	99.133
Threonine	101.04768	101.105
Cysteine	103.00919	103.144
Isoleucine	113.08407	113.16
Leucine	113.08407	113.16
Asparagine	114.04293	114.104
Aspartic Acid	115.02695	115.089
Glutamine	128.05858	128.131
Lysine	128.09497	128.174
Glutamic Acid	129.0426	129.116
Methionine	131.04049	131.198
Histidine	137.05891	137.142
Phenylalanine	147.06842	147.177
Arginine	156.10112	156.188
Tyrosine	163.06333	163.17
Tryptophan	186.07932	186.213

There are additional user-selected parameters that affect the mass calculation: fixed and variable modifications.

Fixed Modifications: When an amino acid residue is subjected to a fixed modification, its mass is changed at every occurrence of this residue. This kind of modification does not need any additional computation and does not affect the database search speed.

In this work, carboxymethyl (C) fixed modification is considered, by which all cysteine (C) residue masses are changed to 161Da when it appears in the sequence (when it is selected). To apply the carboxymethyl (C), the application first checks if the user selected fixed modification. If so, the sequence will be digested and, then, for each residue 'C' occurrence, its mass is changed from 121.16 (unmodified cysteine) to 161Da (carboxymethyl-cysteine).

Variable Modifications: These chemical modifications, unlike the fixed modifications, can occur in an unpredictable pattern, which means they may or may not happen to each amino acid residue. They are also specified by the user. When a variable modification is applied, the original and modified amino acids masses are both calculated. For each occurrence, this calculation is applied with all possible combinations for each peptide containing the modified amino acid. All the resulting masses are compared to the experimental masses to find the best match. Taking variable modifications into account in a search, may help to identify the protein. On the other hand, specifying a large number of variable modifications at the same time leads to exponentially increasing the number of candidate peptides and a decrease in the search speed [14]. In this work, only one variable modification is considered, oxidation (M) [31].

Chapter 4

Proposed Method

This chapter presents the proposed scoring method for PMF matching. It is organized as follows: main steps achieved by this work are shown in Section 4.1; MOWSE is described in detail in Section 4.2.1; and, finally, Section 4.2.3 describes the new scoring method.

4.1 PMF Main Steps

In this work, an application for protein identification using PMF is developed. The application receives multiple files as input: peak list, contaminants, and query databases. The peak list file contains the experimental data that will be searched against the query database to find a match. The contaminants database contains the contaminants that will be searched against mass data to filter them out. Each input file needs preprocessing (discussed in Chapter 3) before performing the database search. When the peak list and query databases are preprocessed, they are matched to each other and, based on the matching, the theoretical proteins are scored and ranked. The top hits list and the details for each hit will be presented. Figure 4.1 shows the main steps of the performed PMF method.

In the matching phase, experimental masses are searched against theoretical proteins. For each theoretical protein, its peptide masses are compared against the experimental peptide masses. When any match is found, this protein is ranked as a candidate. Whenever the number of matches for a protein increases, its probability of being a "true" match increases as well. After

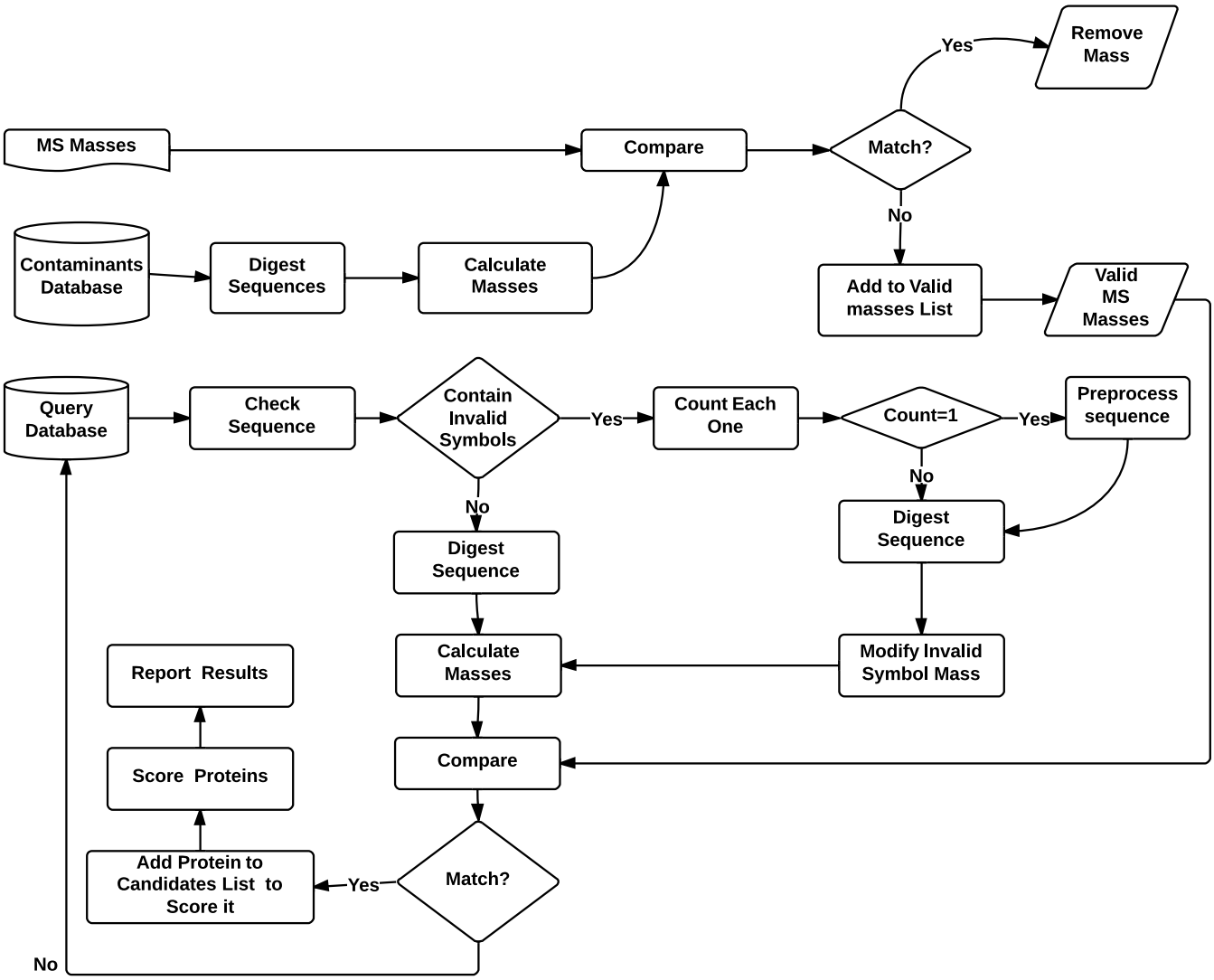


Figure 4.1: PMF flowchart which represents the steps performed by this work.

producing the list of candidate proteins, the respective theoretical proteins will proceed to the scoring phase to evaluate their matches.

4.2 Scoring

All candidate proteins are scored and, based on those scores, a ranking is determined. Ideally, the highest ranking entry should correspond to the experimental protein sample. MOWSE is used to produce the raw score, this raw score is improved to identify the proteins more accurately.

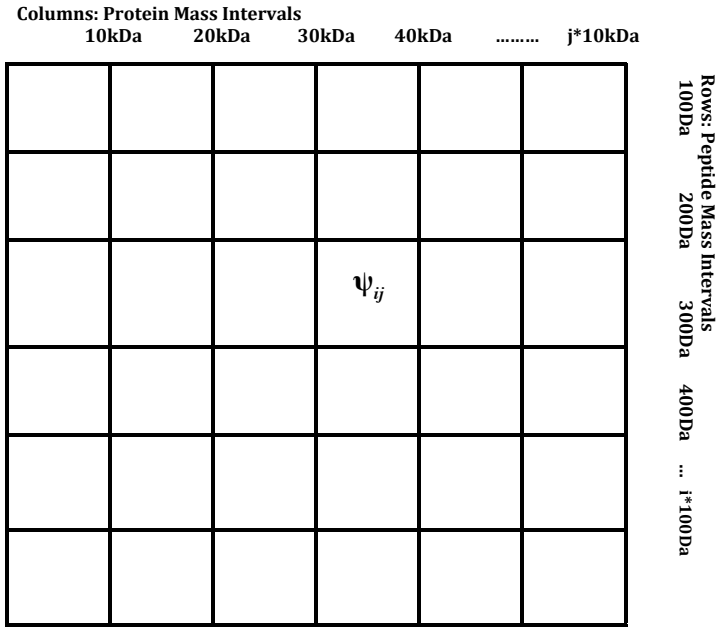


Figure 4.2: Columns represent database protein molecular weights, whereas the rows represent the database peptide molecular weights.

4.2.1 MOWSE Scoring Algorithm

MOWSE is one of the most important scoring methods of PMF [48, 40, 2]. It compares the calculated masses for each protein in the database with the masses of the experimental protein. Each value that falls within a given mass tolerance is considered a match. This method depends on the number of theoretical proteins that match to each protein sample, and the frequency of the molecular weight for each peptide. A frequency table is built where columns represent database proteins, and rows indicate the peptides for these proteins (Figure 4.2). Any peptide that belongs to the same protein will contribute in a single column which represents this protein [49, 36]. The following steps describe how this method works :

- Aggregate proteins to bins, so that each bin sums up to 10kDa.
- Arrange the peptides for each protein into 100Da bins. For example, Table 4.1 shows how these peptides are arranged for one protein.
- For each protein (10kDa), normalize the values in the bins by:

$$f_{ij} = \psi_{ij} / \max(\Psi_j) \tag{4.1}$$

Table 4.1: Peptides arrangement based on their mass value.

Peptide	Mass	Interval	Peptide in Interval
MASNTVSAQGGSNR	1732.91	400-500	MNR
DFSNIQDVAQFLLFDPIWNEQPGS	3405.84	600-700	LDIMQ
MNR	420.51	800-900	EQALAER
EQALAER	816.89	1700-1800	MASNTVSAQGGSNRPVR
YPELQTSEPSEDYSGPVESLE	3164.49	3100-3200	YPELQTSEPSEDYSGPVESLE
LDIMQ	619.76	3400-3500	DFSNIQDVAQFLLFDPIWNEQPGS

Where f_{ij} is the frequency of cell_{ij}, ψ_{ij} is the value in the cell, and $max(\Psi_j)$ is the largest value in column_j.

- Compare the peak list to each protein in the database.
- Score the theoretical proteins by retrieving the frequency of matched peptides and multiplying them by:

$$H(\hat{\Psi}) \propto \prod_{i=R(\hat{\psi})} f_{ij} \tag{4.2}$$

Where $H(\hat{\Psi})$ is the product of matched values in the frequency table, $R(\hat{\psi})$ is the row number of the table for the peptide whose mass matches the peak list, and $\hat{\Psi}$ is the set of fragments of the experimental protein matched with the theoretical protein.

- To get the final score τ_{mowse}, the product of matched peptides is normalized to an average protein of 50kDa to restrict the influence of random score growth for proteins >200kDa.

$$\tau_{\text{mowse}} = \frac{50000}{H(\hat{\Psi}) * W(P_{\text{th}})} \tag{4.3}$$

Where the $W(P_{\text{th}})$ is the molecular weight for the theoretical protein P_{th}.

4.2.2 Score Significance

One of the main challenges in PMF is to decide if the match between the experimental sample and the database protein is a correct match rather than just a random match or at least to which level the match is reliable (*i.e.* the confidence level of the match). The comparison between experimental masses and theoretical proteins is applied with a given mass tolerance. Within this

tolerance range, several theoretical masses can match one experimental mass. Many of these matches are not related to the correct protein and are considered random matches. In fact, a random match appears when the score of an unrelated protein is at least as good as the score of the correct protein [11, 12].

To overcome this problem, the scoring algorithm should keep the random match score to a minimum, to avoid identifying a wrong protein. A good scoring algorithm should also provide a significance score and not only the score of matching. A significance score can be defined as the score which is related to the correct protein and should be as distant as possible from the score which is related to a false positive.

4.2.3 Proposed Scoring Method

The scores produced by MOWSE are far from being a valid judgment criterion for matching. This method is very susceptible to random matches and large proteins are very likely to be ranked as potentially valid proteins. This is because they might have a lot of matched masses, and hence, their scores will be very high. In this work, a new scoring method is proposed based on the following assumptions:

1. MOWSE is good in arranging the proteins and the masses in a 2-D histogram.
2. MOWSE is efficient in scoring but does not involve any additional valuable features from the theoretical and experimental proteins.

In light of these facts, the occurrences product was first normalized by the number of matched peptides, then, the scoring function is defined as:

$$\tau = \tau_{\text{mowse}} * |\hat{\Psi}(P_{\text{th}})| * \left(\frac{|\hat{\Psi}(P_{\text{th}})|}{|\Psi(P_{\text{ex}})|}\right) * \left(\frac{|\hat{\Psi}(P_{\text{th}})|}{|\Psi(P_{\text{th}})|}\right) \tag{4.4}$$

Where $\hat{\Psi}(P_{\text{th}})$ is the number of matched peptides in the theoretical protein P_{th}, $\Psi(P_{\text{ex}})$ is the number of experimental protein peptides, and $\Psi(P_{\text{th}})$ is the number of theoretical protein peptides.

In Equation 4.4, the MOWSE score is scaled by the number of matched masses. Introducing the matched masses is two fold. Firstly, a scaled MOWSE score is the natural way of looking at the problem. It answers the question: how many masses produce this score? Secondly, the ratio between the matched masses and total masses in the protein, tells us about the coverage of the matching.

To represent the score in an understandable way, the logarithm with base ten is applied for this score by the formula:

$$-\log_{10}\left(\frac{1}{\tau}\right) \tag{4.5}$$

In addition to this new scoring, and to ensure good results, a thresholding approach is proposed. It is unlikely to consider a protein as a potential match if it has a very low matching ratio. In other words, random matches are very likely to be large proteins. A solution to this problem is to restrict the candidate list to include only the set of proteins that have a good matching ratio. This ratio is defined as:

$$\frac{|\hat{\Psi}(P_{\text{th}})|}{|\Psi(P_{\text{th}})|} \geq \lambda \tag{4.6}$$

Where $\hat{\Psi}(P_{\text{th}})$ is the number of matched peptides, $\Psi(P_{\text{th}})$ is number of peptides in the theoretical protein, and $0 \leq \lambda \leq 1$ is a threshold. In this work, when $\lambda \geq 0.09$, the theoretical protein score is considered. The threshold 0.09 was selected based on the experimental results.

Chapter 5

Results and Discussion

There are two widely-used approaches to evaluate the accuracy of methods for protein identification. In the first approach, the gold standard is the hit-miss criteria, *i.e.* whether the method found the intended protein or not [43]. In the second approach, other matching features like coverage and number of matched peptides are included [50]. This work used the first approach with ground truthing by manipulating noise levels in the data (peptide mass lists), either by adding contamination as random non-matching peptides, or by removing masses, (*i.e.* simulating missing peptides in the data).

The material and the details of the random contamination/missing data approach are described in Section 5.1. Section 5.2 presents the tests performed in order to assess the proposed tool.

5.1 Simulated Peptide Data and Noise Manipulation

The well-known protein database *Swiss-Prot* [17, 4] was used to test the proposed tool. In this work, three test sets were prepared to carry out three different tests. The sets were obtained by simulating the MS task.

To simulate the MS masses preparation, random sets of proteins were selected. These sets were *in-silico* digested and their resulting peptide masses calculated. A simulated contamination process was applied to the resulting masses prior to testing. One can describe the random "contamination" as a random noise with a given strength. Noise can take two forms; additional

masses not originating from the protein of interest (*i.e.* contamination), or missing peptides from the mass list for the protein of interest (*i.e.* incomplete spectra). Both types of noise are important and are often present in real biological data. The higher the noise the lower the chance of retaining the original protein. However, the classical and important question is to know to which level of noise a method can still be considered as valid. Therefore, a range of 40% to 90% was used to remove masses randomly. This was followed by a random addition of masses within the range of the original protein masses. The added noise falls between 5% to 40%. Table 5.1 shows the digested masses of one protein before and after the simulated contamination.

In fact, having a range of noise does not only give a stable way to test a given method but also can identify its breakpoint. In other words, with low noise, most of the methods may produce good results, but when the level of the noise increases, it is vital to know when a method under test fails.

5.2 Tests

For the evaluation, the simulated proteins with varying levels of contamination referred to in Section 5.1 were used as input samples in the comparison between the proposed tool and two common software packages: Mascot (commercial) and Ms-Fit (academic). Besides being two of the most popular tools in PMF, these software packages use the same raw scoring method (MOWSE) as the proposed method. In particular, Mascot is very robust and produces very reliable results [5]. Additionally, because the proposed method can be seen as an improvement to MOWSE, 18 samples were used to compare MOWSE and the proposed tool to highlight the enhancement of the new approach.

5.2.1 Parameters

The parameters used in all tests are shown in Table 5.2, column 3. These parameters are very important because of their influence in the performance of the proposed tool as well as all PMF software packages. The table also shows some parameter values that are commonly used in

Table 5.1: Protein masses before and after a simulated contamination process that removed 70% of the original peptide masses and added random masses corresponding to 30% of the original mass values. Cells in pink refer to the masses that will be removed, while cells in blue refer to the masses that have been added to the protein.

Protein Name	**Protein Accession Number**
ACTP_YERPB	B2K137
Original Masses	**Contaminated Masses**
278.1538396	278.1538396
288.2035596	2716.456916
2716.456916	175.1194996
3086.678729	262.1515286
175.1194996	459.2679536
276.1671796	914.4620726
262.1515286	375.1992036
1030.459402	6043.213981
5829.069439	1982.266356
288.2035596	3363.698101
459.2679536	7091.651684
1192.563587	432.2206696
910.5586536	848.4477696
2364.302386	341.1573416
4458.581324	6885.590332
1813.944695	1196.655396
1716.859152	522.4448242
480.2931636	5743.879883
3692.994273	1091.497803
914.4620726	5769.700684
375.1992036	4227.228516
147.1133496	5868.097168
3778.033451	3756.529053
6043.213981	2061.907959
147.1133496	3400.99585
204.1348136	
489.2421336	
532.2730896	
333.2135186	
1982.266356	
3363.698101	
490.2989196	
7091.651684	
432.2206696	
520.3498796	
636.3466586	
848.4477696	
341.1573416	

database searches.

Table 5.2: Parameters used in the software comparisons and respective values. The first column lists the parameter name, the second lists commonly used values, and the third lists the values used in the comparisons.

Parameter	**Possible Value**	**Used Value**
Peak List	Real Value List	Real Value List
Missed Cleavages	0, ..., 9	0
Fixed Modification	Carboxymethyl (C)	None Selected
Variable Modification	Oxidation (M)	None Selected
Database Mass Tolerance	Real-Value (in Dalton)	1Da
Mass Type	Average or Monoisotopic	Monoisotopic
Enzyme	Trypsin	Trypsin
Contaminants Mass Tolerance	Real-Value (in Dalton)	1Da
Taxonomy	All species	All species
Database(s)	All Databases	Swiss-Prot
Top Hits	1, ..., 50	1, ..., 40

5.2.2 Test Criteria

In order to understand how these software packages respond, several test factors were used. These are:

- Noise level: Amount of added and subtracted peptide masses, as '%' of initial number of peptides.
- Rank of the result: Rank order of the source protein found by the software.
- Number of matched peptides: Number of the peptides matched with peak list found by the software.
- Total number of peptides: Total number of protein peptides.

It is worth pointing out that whenever the software presents the correct protein at the top of its hit list, this indicates the strength of this software.

5.2.3 Software Comparison

Before presenting the tests, it is important to point out the general pros and cons of each tool. These are shown in Table 5.3.

5.2.4 Test No. 1

As it was not clear which range of noise should be used. The robustness of each software was tested over a broad range of contaminant addition (from 5% to 40%), and missing peptides (from 40% to 90%). A set of 28 different proteins was used. The levels of adding and removing random masses were changed gradually in order to identify the point where finding a top-ranked protein becomes difficult or impossible . The results of this test are shown in Table 5.4. Nevertheless, this test can be used to assess the quality of results produced by the tested tools, it is used in the first place to identify the critical noise ranges before applying a more rigorous test.

Table 5.4, indicates that target protein identification is more affected by missing data in the MS peak list than by contamination (addition of non-matching noise). For instance, 70% mass removal had little effect on correct identification, in the presence of 20% to 25% contamination. However, protein *ACP_CHLFF* with 60% and 30% removal and additive levels, respectively, showed poor results because of its original small size, *i.e.* 7 peptides: 4 were removed (57% of data loss) and 2 were added (28% of random noise), resulting in only 15% of correct data. In practice, it is more likely to deal with loss of data rather than added noise. Additionally techniques exist for removing contaminants (which this tool offers). In terms of effectiveness, the results showed that Mascot and the proposed tool were able to identify 20 correct proteins out of 28 samples as first rank. On the other hand, the number of correct proteins identified by Ms-Fit was only 6 out of 28 samples.

5.2.5 Test No. 2

Since Mascot and the proposed tool were robust to high proportions of missing data (*e.g.* 80%). A range of 70% to 90% was chosen to perform more tests. As additive noise can be preprocessed by contamination suppression techniques, a range between 5% to 25% was considered.

Table 5.3: Software packages comparison. Advantages and drawbacks for each software.

Software	**Databases Supported**	**Type**	**Local Version**	**Contamination Removal**	**Modification Supported**	**Missed Cleavages**
Mascot	Public	Commercial	Paid	Available	Yes	Available
Ms-Fit	Public	Academic	Not available	Available	Yes	Available
ProFound	NCBI	Academic	Not available	Not available	Yes	Available
Proposed	Public+local	Free	Available	Available	Yes	Available

Table 5.4: Results of PMF using Mascot, Ms-Fit, and the proposed software packages. Column *Peptides Number* indicates the total number of peptides of the protein sample. Column *Rank* indicates the rank order of protein, and column *Matched Peptides* indicates the number of peptides of the protein hit that match to the experimental peptides.

Protein Sample	**Peptides Number**	**Noise%**		**Mascot**		**Ms-Fit**		**Proposed**	
		Add	Remove	Rank	Matched Peptides	Rank	Matched Peptides	Rank	Matched Peptides
AAE13_ARATH	57	40%	40%	1	30	2	26	1	31
ACSA_AGRVS	72	40%	40%	1	42	1	42	1	41
1B02_GORGO	36	30%	50%	1	26	2	26	1	24
ACOD_CYPCA	33	30%	50%	1	22	1	22	1	20
ACDD_METKA	37	40%	50%	1	13	2	13	1	14
ACCD_STRPQ	35	30%	60%	3	10	3	10	1	10
ACP_CHLFF	7	30%	60%	23	3	N/A	0	6	3
AKP8L_MOUSE	80	30%	60%	1	31	3	31	1	34
ACE4_CAEBR	63	20%	70%	1	13	3	13	1	13
ACH1_MANSE	44	20%	70%	1	14	1	14	1	14
AMP11_ENCCU	103	20%	70%	1	25	2	25	1	25
ADN2_SCHPO	60	25%	70%	1	17	2	17	1	17
AMOL2_BOVIN	98	25%	70%	1	27	4	27	1	31
ACAC_SCHPO	254	15%	80%	1	56	3	56	1	69
ACTB_XENLA	37	15%	80%	33	7	N/A	0	30	8
AMI_XENTR	23	15%	80%	1	6	1	6	1	6
AN13A_MOUSE	62	15%	80%	5	9	26	9	2	9
A2MG_RAT	130	20%	80%	1	97	3	97	1	95
ARGB_CALS8	33	20%	80%	1	28	3	28	1	27
ACDH_CHLAD	29	20%	80%	1	6	2	6	1	6
ADE_PSEF5	29	20%	80%	1	5	N/A	0	1	5
ANR46_HUMAN	25	20%	80%	5	5	25	5	14	5
ARLY_SERP5	45	20%	80%	1	8	1	8	1	7
5NTD_TREPA	64	5%	90%	1	5	N/A	0	N/A	0
AMPM_SALTI	27	5%	90%	N/A	0	17	3	N/A	0
ADRB1_MOUSE	46	10%	90%	N/A	0	N/A	0	N/A	0
APL_ARATH	37	10%	90%	3	5	3	5	6	5
APAH_PSEPG	34	10%	90%	1	5	1	5	1	5

In the previous test, different proteins were tested with gradual change in the noise. This helped to design a protein-oriented test. This test is designed to study the effect of the gradual increase of noise (as missing MS peptide peaks in the data) on the same protein. As the peak list quality varies with equipment, sample, and the personnel experience, it is important to test the tools under different levels of noise.

Table 5.5 shows the results of Mascot, Ms-Fit, and the proposed tool using 7 different proteins with data loss ranging from 70% to 90% and a fixed contamination level is of 10%.

The results of Mascot and the proposed tool were similar, except on samples FADJ_ECOHS and MDTB_ECOLI. Mascot presented better results in sample FADJ_ECOHS by reporting it as first rank with 70% and 80% removal rate, and as third when the proportion of removed peptides was of 90%. The proposed tool reported it as fourth rank at all levels of missing data (70%, 80%, and 90%). On the other hand, the proposed tool gave better results with MDTB_ECOLI when removal rates were 70% and 80%. In some cases, when 90% of the peptides were removed, none of the tools could identify the protein. As in the previous test, Ms-Fit was less accurate than the other two software packages.

A concise summary that provides some statistical information based on the results presented in Table 5.5 is shown in Table 5.6. As shown by this table, Mascot and the proposed tool are highly comparable and in most of the cases identified the correct protein sample as a top hit.

5.2.6 Test No. 3

This test was designed to compare the new scoring method with the MOWSE approach from which was developed. It is important to highlight the improvement by a direct comparison between MOWSE and the proposed improved method. The results of this comparison are presented in Table 5.7.

In this table, it is clear that the proposed scoring method has a reasonable amount of enhancement reflected in its matching accuracy. The main problem of MOWSE is that it does not implement an outliers rejection mechanism. No protein was found as the first hit, and many proteins were not identified at all.

Table 5.5: Results of Mascot, Ms-Fit, and the proposed tool when contaminating each protein with 10% additive noise and three different data removal rates. Column *Peptides Number* indicates the number of peptides for the protein sample, column *Rank* indicates the rank order of the source protein, and column *Matched Peptides* indicates the number of matched peptides for that protein.

Protein Sample	**Peptides Number**	**Noise%**		**Mascot**		**Ms-Fit**		**Proposed**	
		Add%	Remove%	Rank	Matched Peptides	Rank	Matched Peptides	Rank	Matched Peptides
ASA1_CLAL4	32	10%	70%	1	10	1	10	1	10
			80%	1	6	1	6	1	6
			90%	1	4	1	4	3	4
DXS_WOLSU	65	10%	70%	1	19	1	19	1	20
			80%	1	11	1	11	1	12
			90%	N/A	0	N/A	0	N/A	0
EXOC2_DICDI	121	10%	70%	1	25	1	25	1	29
			80%	1	26	1	26	1	27
			90%	1	12	9	12	1	13
FADJ_ECOHS	81	10%	70%	1	20	4	20	4	20
			80%	1	13	6	13	4	13
			90%	3	8	7	8	4	8
MDTB_ECOLI	67	10%	70%	14	19	10	19	7	19
			80%	14	15	11	15	5	15
			90%	N/A	0	N/A	0	N/A	0
NDHF_BACSU	30	10%	70%	1	9	14	9	1	9
			80%	1	8	1	8	1	8
			90%	1	5	5	5	1	5
RL15_RHIME	30	10%	70%	1	7	1	7	1	7
			80%	1	6	1	6	1	7
			90%	N/A	0	N/A	0	N/A	0

Table 5.6: Mascot, Ms-Fit, and the proposed tool statistical information that represent the hits and miss for protein samples. *Number of Finding* indicates the number of proteins that the software found from the samples set listed in 5.5. *Number of First Ranks* indicates the number of correct proteins reported as a first rank. *Number of Missing* indicates the number of unidentified proteins by the software. *Top 5 Ranks* indicates the number of proteins that the software reported on top 5 ranks.

	Mascot	**Ms-Fit**	**Proposed**
Number of Finding	18	18	18
Average of Finding	86%	86%	86%
Number of Missing	3	3	3
Average of Missing	14%	14%	14%
Number of First Ranks	15	10	12
Average of First Ranks	71%	48%	57%
Top 5 Ranks	16	12	17
Average of Top 5 Ranks	76%	57%	81%

Table 5.7: Results comparison between MOWSE and the proposed method. Column *Peptides Number* indicates the number of peptides for the protein sample, column *Rank* indicates the rank order of the source protein, and column *Matched Peptides* indicates the number of matched peptides for that protein.

Protein Sample	**Peptides Number**	**Noise%**		**Proposed**		**MOWSE**	
		Add	Remove	Rank	Matched Peptides	Rank	Matched peptides
ACE4_CAEBR	63	20%	70%	1	13	4	13
ACH1_MANSE	44	20%	70%	1	14	3	14
AMP11_ENCCU	103	20%	70%	1	25	5	25
ADN2_SCHPO	60	25%	70%	1	17	4	17
AMOL2_BOVIN	98	25%	70%	1	31	7	31
ACAC_SCHPO	254	15%	80%	1	69	6	69
ACTB_XENLA	37	15%	80%	29	8	31	8
AMI_XENTR	23	15%	80%	1	6	2	6
AN13A_MOUSE	62	15%	80%	2	9	N/A	0
A2MG_RAT	130	20%	80%	1	95	5	95
ARGB_CALS8	33	20%	80%	1	27	7	27
ACDH_CHLAD	29	20%	80%	1	6	5	6
ADE_PSEF5	29	20%	80%	1	5	8	5
ANR46_HUMAN	25	20%	80%	17	5	N/A	0
ARLY_SERP5	45	20%	80%	1	7	9	7
AMPM_SALTI	27	5%	90%	N/A	0	N/A	0
ADRB1_MOUSE	46	10%	90%	N/A	0	N/A	0
APL_ARATH	37	10%	90%	5	5	N/A	0

5.2.7 Discussion

The comparison results shown in this chapter revealed several strengths and weaknesses of the tested methods and tools.

As expected, MOWSE (Table 5.7) was clearly hampered by low accuracy. It failed to report any source protein ranked first. In addition, it could not identify several proteins that were identified by the proposed method, *e.g.* protein APL_ARTH as shown in this table, especially when the proportion of missing data was high (90%).

Ms-Fit results were reasonable. It found the correct protein in many cases but not ranked as first (Table 5.5). Additionally, Table 5.4 showed that Ms-Fit missed several proteins like protein ADE_PSEF5 that was ranked first by Mascot and the proposed tool.

Mascot showed the best results and was more accurate than the other tools (Table 5.5). However, as shown in Table 5.4 Mascot also failed to identify the protein in some cases when the missing data rate became 90%, *e.g.* protein AMPM_SALTI.

The proposed tool presented good results when comparing to the other tools. Its scoring method showed better results than MOWSE (Table 5.7) in which it reported 12 out of 18 proteins correctly as first hit while MOWSE failed to report any protein as first hit. The proposed tool also presented good results when compared to Ms-Fit and Mascot. It was very competitive with Mascot and better than Ms-Fit (Table 5.4, and Table 5.5).

In summary, most of the tools identified the correct protein when levels of missing data are moderate. In contrast, with increasing loss of data, they either failed to identify the protein as first rank or can not find a match.

The size of the protein is another important factor. Small proteins are more susceptible to noise, specially because scoring methods try to use the data in the query to reject outliers (random matches) and to find as many matched peptides in the database as possible. Missing data in the peak lists of small proteins will have greater impact on search efficiency than missing data from longer peak lists of large proteins.

Another important point is that, when the size of the database increases, these methods may

become error prone due to large similarity between database proteins which may led to poorer ranking (a higher ratio of random matches).

Chapter 6

Conclusion

This work presents and discusses protein identification using PMF, beginning with a description of the main steps that should be followed to identify a protein of interest. Then, it illustrates the complete protein identification pipeline using PMF. The focus here was mainly put on the matching and scoring PMF steps. Additionally, the problems faced during this work like ambiguous amino acids and their solutions are discussed, like dealing with ambiguous amino acids in protein databases.

A background on PMF identification is presented and some of the state-of-the-art tools such as Mascot, Ms-Fit, and ProFound are discussed. MOWSE and additional scoring methods of these tools (when available), and statistical assessment for MS scoring methods are also discussed.

The main purpose of the computational protein identification is to implement a scoring method that can accurately identify a target protein in the presence of varying levels of noise in MS spectra, while accounting for ambiguity in database protein sequences. A key motivation of this work was to provide an effective and freely available solution for biologists, particularly those working on non-model systems where specialized local databases are essential. Therefore, a PMF tool with a new scoring method that produced results comparable, and better than state-of-the-art software was developed and introduced. This proposed software is open-source for free and allows the use of local databases. It reduces the impact of random matches in massive (and growing) protein databases and can be used as a valid alternative to commercial or non

open-source PMF protein identification software packages.

The proposed tool performed well in comparisons with two of the most popular tools. It was very competitive with Mascot, one of the most popular and reliable software available and gave results that more accurate than Ms-Fit.

In summary, the software developed in this work provides a real help to biologists by offering a local tool that allows the use of local databases for free. Furthermore, this software can work with any protein database and provides a friendly graphical user interface (GUI). It removes the possible contaminants from the experimental samples leading to enhanced protein identification results.

6.1 Future Work

At this time, this work provides carboxymethyl (C) as fixed modification, and oxidation (M) as variable modification. Even though these parameters are the most commonly used, it would be simple to extend the functionality of the application to include more protein modifications

Additional work is being done to include a graphical presentation of the results that displays the protein sequence and highlights the matched peptides in the complete protein sequence.

Although performance was not the main goal of this work, it is observed that some the software packages that have been compared to the proposed software were, in some cases, faster. However, given the restrictive nature of the other software (web-based or commercial), comparative bench-marking on the same local machine was not possible. In PMF identification, different parameters require the calculation of different intermediate results. These take a considerable amount of time to calculate and must be recalculated for every new sample search. Therefore, a mechanism is being developed for saving parameter values and respective intermediate results. In this case, each time the user starts an analysis, the program will check if those parameters have been used (with the same database), and if so, the data related to these parameters will be loaded and used, instead of performing all the calculations again from the beginning. If they have never been used, the program will do all calculations for these new parameters and save the intermediate results to a database for future use.

Appendix A

Prototype

An application for PMF protein identification is developed. This appendix presents the prototype and its main parts. It also presents a sample of the results report provided by the application.

A.1 Prototype

In this work, a prototype was developed to provide a computational tool for biologists to perform PMF efficiently. The requirements of the application are presented in Table A.1. This section presents the developed software package and its features.

Table A.1: Application requirements and availability.

Project Name: Computational Tool for Peptide Mass Fingerprinting.
Operating System(s): Platform independent.
Programming Language: Java.
Other Requirements: JRE $\geq$ 1.6 to run the application. To compile the source code, the Netbeans Platform $\geq$ V.7.3 IDE.
Required Memory: 2GB.
License: Free.
Any restrictions to use by non-academics: None.

A.1.1 The Main Window

Figure A.1 displays the four main parts in the main GUI window. These are:

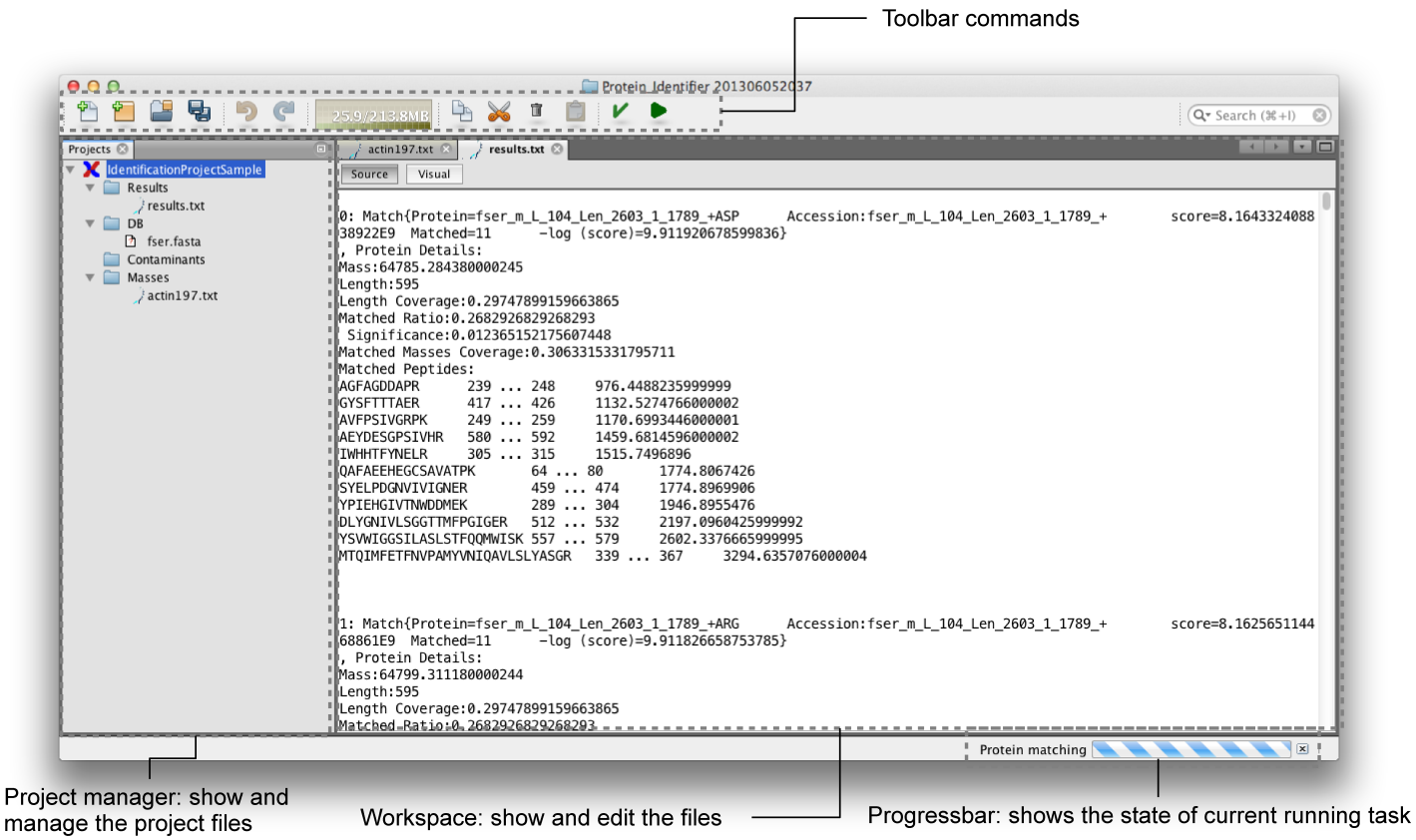


Figure A.1: Graphical user interface main window.

Toolbar: Contains the commands and their functionalities as explained in Figure A.2.

Progressbar: Shows the progress of the current project task.

Workspace: This area allows the users to display and edit the files listed in the project manager.

Project manager: In this area, all the folders and files used by the project are listed, (Figure A.3).

The folders displayed in this area are:

1. Database folder (mandatory): It contains all the required protein databases. It must contain at least one database. If this folder is empty, the application will display an error message and will stop running.

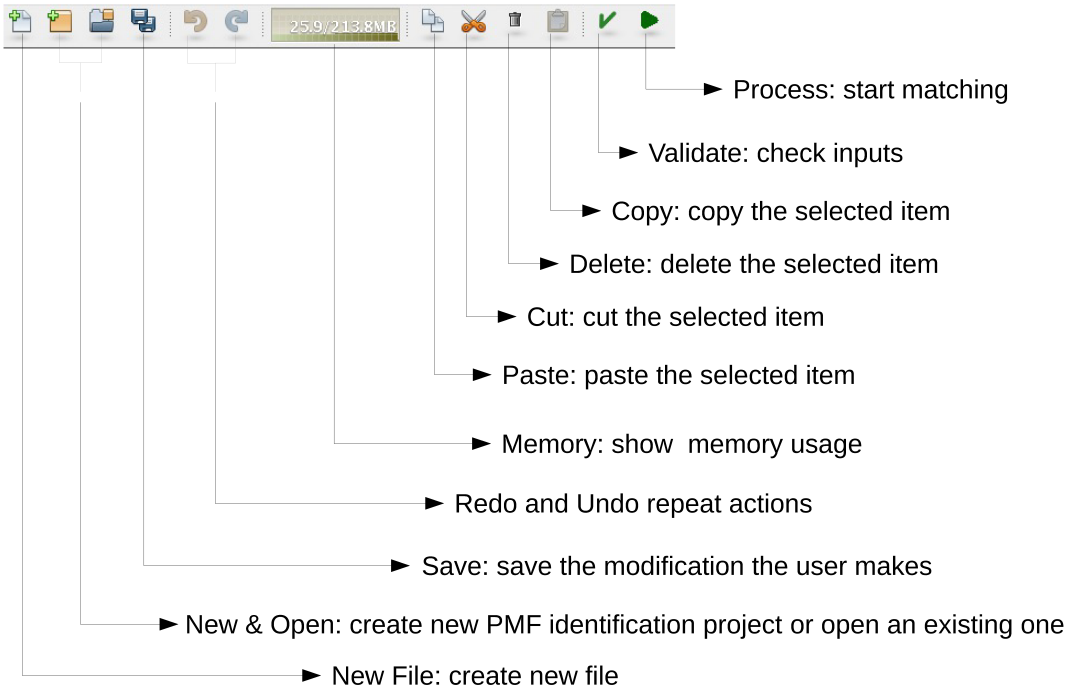


Figure A.2: Toolbar commands explanation.

2. Experimental data folder (mandatory): This folder contains the files of experimental samples (peak lists). Like the database folder, at least one file is required for the application to run.

3. Contaminants folder (optional): It contains the contaminants databases. It is optional to have contaminants files.

4. Results folder: After finalizing the analyses, the output files are stored in this folder which makes it easy for the user to display and analyze the results.

The user can delete, create, and edit the folders (listed in project manager) and their contents.

A.1.2 Search Parameters

When the required files are added to the application, the user can enter the parameters. Figure A.4 displays these parameters and their default values.

These are the following:

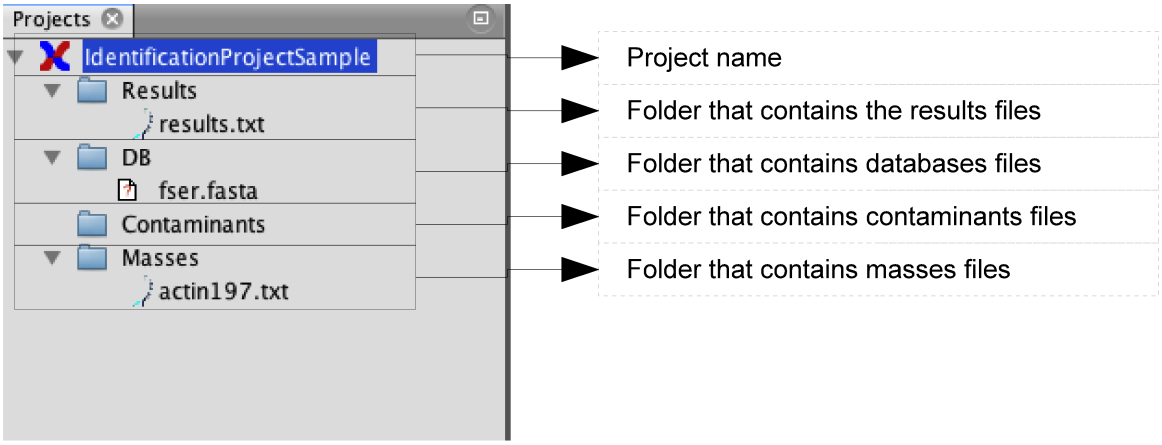


Figure A.3: Project panel contents.

- Enzyme: The enzyme used for digestion. The default value is Trypsin.
- Missed cleavages: The number of missed cleavages that the application should consider during the digestion. It ranges between 0 and 9. The default value is 0 (*i.e.* no missed cleavages).
- Fixed modification: sample-dependent. The default value is carboxymethyl (C).
- Variable modification: sample-dependent. The default value is oxidation (M).
- Mass tolerance: When comparing an experimental mass to a theoretical mass, this value represents the maximum error allowed to still consider this comparison as a match. The default value is 1Da-4Da.
- Contaminants tolerance: A threshold value used when comparing a contaminant mass to the experimental mass. It is entered in Dalton, and the default value is 1.0^{-4}Da.
- Mass type: Specifies whether the experimental mass values are average or monoisotopic. The default value is monoisotopic.
- Matching tolerance: A threshold used to validate the matching. If the matching ratio of a protein falls within this value, it is considered as a hit. The default value is (0.09).

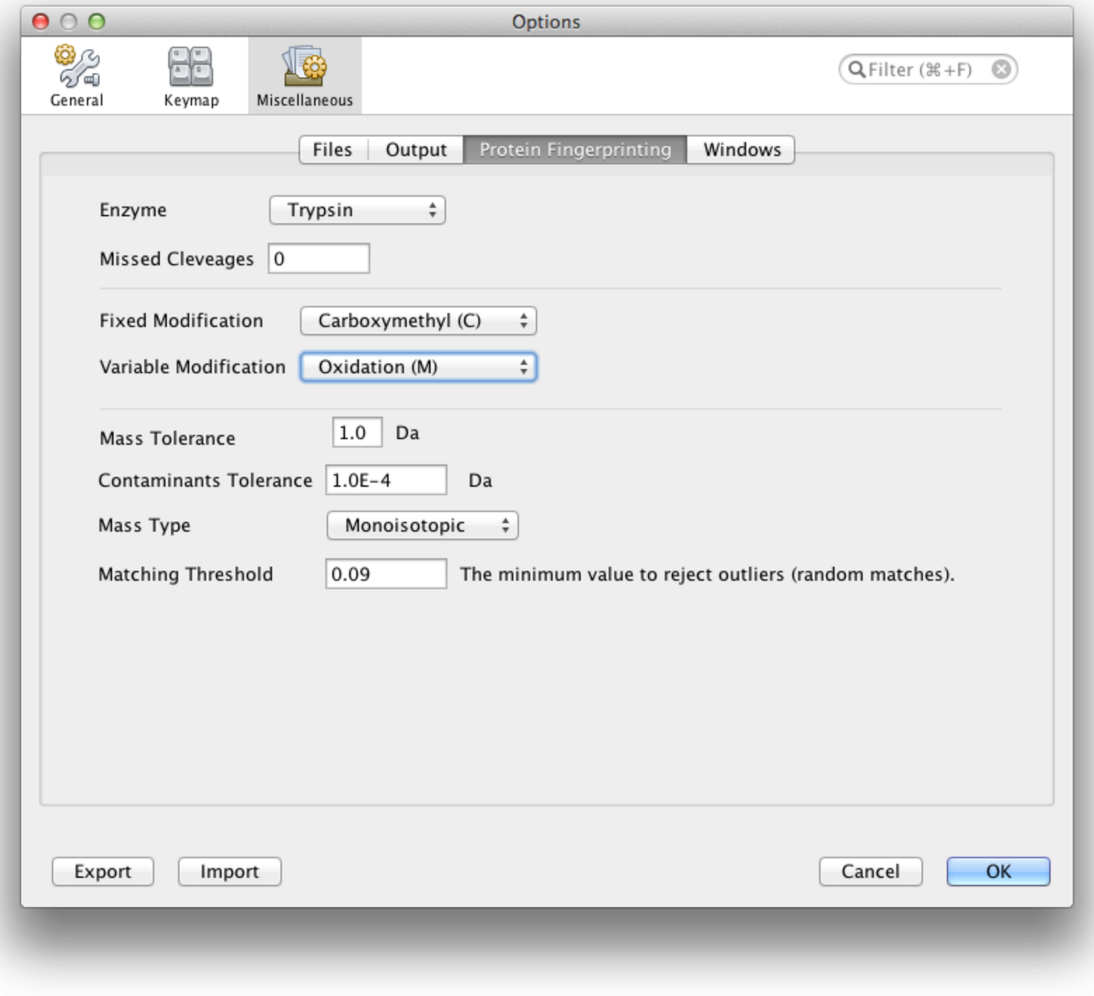


Figure A.4: Matching parameters.

A.1.3 Results

After running the application, the results are reported. Figure A.5 explains the format and details of the results.

Figure A.5 displays the top hits proteins and respective details. First, the name of the protein and its accession number, score, number of matched peptides, and the adjusted score are displayed. This information helps the user to determine if the hit is to a correct protein or just a random match.

Then, more detail is provided for each protein in the top ranks reported by the application to

```
1: Match{Protein=AB5A_ARATH  Accession:Q9STT7   Score=2.4556747552956332E16  Matched=23      Adjusted Score=16.390170845639283}

Protein Details:
Mass:105584.27698000005
Length:936
Length Coverage:0.29914529914529914
Matched Ratio:0.23958333333333334
Matched Masses Coverage:0.3026304180389718

Matched Peptides:
NIWSNVR 27 ... 33   888.4688985999998
DSCR    122 ... 125 480.18765859999996
ADYTNYLDPGILSDLPIFNVQPR 179 ... 201 2621.3248485999993
EVR 223 ... 225 403.2302296
GYR 247 ... 249 395.2042936
QGNLEEIINEVAAAYDLMDTDINNFNVTIWYNSTYK    250 ... 285 4181.964814600001
EMPK    327 ... 330 504.2491936
IIMK    373 ... 376 504.3219596
QQHLR   368 ... 372 681.3796295999999
FNDYSIQFIFYFLCINLQISIAFLVSSAFSK 418 ... 448 3638.8646126
GLYEFSQYAFK 499 ... 509 1352.6526666
NQNPFK  565 ... 570 747.3789636
QVSAIAIEMEK 581 ... 591 1218.6401166
VEQLMLETSTGHAIVCDNLK    601 ... 620 2201.0940565999995
GLSLAVPSGECFGMLGPNGAGK  637 ... 658 2062.0098775999995
TSFINMMTGLMKPTSGAAFVHGLDICK 659 ... 685 2870.4074316
NLK 721 ... 723 374.24033959999997
NHTAIILTTHSMEEAEFLCDR   800 ... 820 2431.1385866000005
DVEMLVQDVSPNAK  859 ... 872 1544.7622066
IYHIAGTQK   874 ... 882 1030.5685476
DNFR    904 ... 907 551.2577795999999
VAR 925 ... 927 345.2247536
TAQASNVFS   928 ... 936 924.4424155999998
```

Figure A.5: Matching report.

make the results analysis and validation easier and more accurate. This information is:

- Protein mass: The protein molecular weight.
- Length: The total number of amino acids.
- Length coverage: The ratio between the length of matched peptides and the length of the complete sequence.
- Matched ratio: The ratio between the number of matched peptides and the total number of peptides in the sequence.
- Matched masses coverage: The ratio between the matched peptide masses and the mass of the complete sequence.
- Matched peptides: Table of matched peptides in the following order: peptide sequence, start-position ... end-position, and peptide mass. Start and end positions show where the peptide starts and where it ends in the protein sequence.

Appendix B

UML Classes

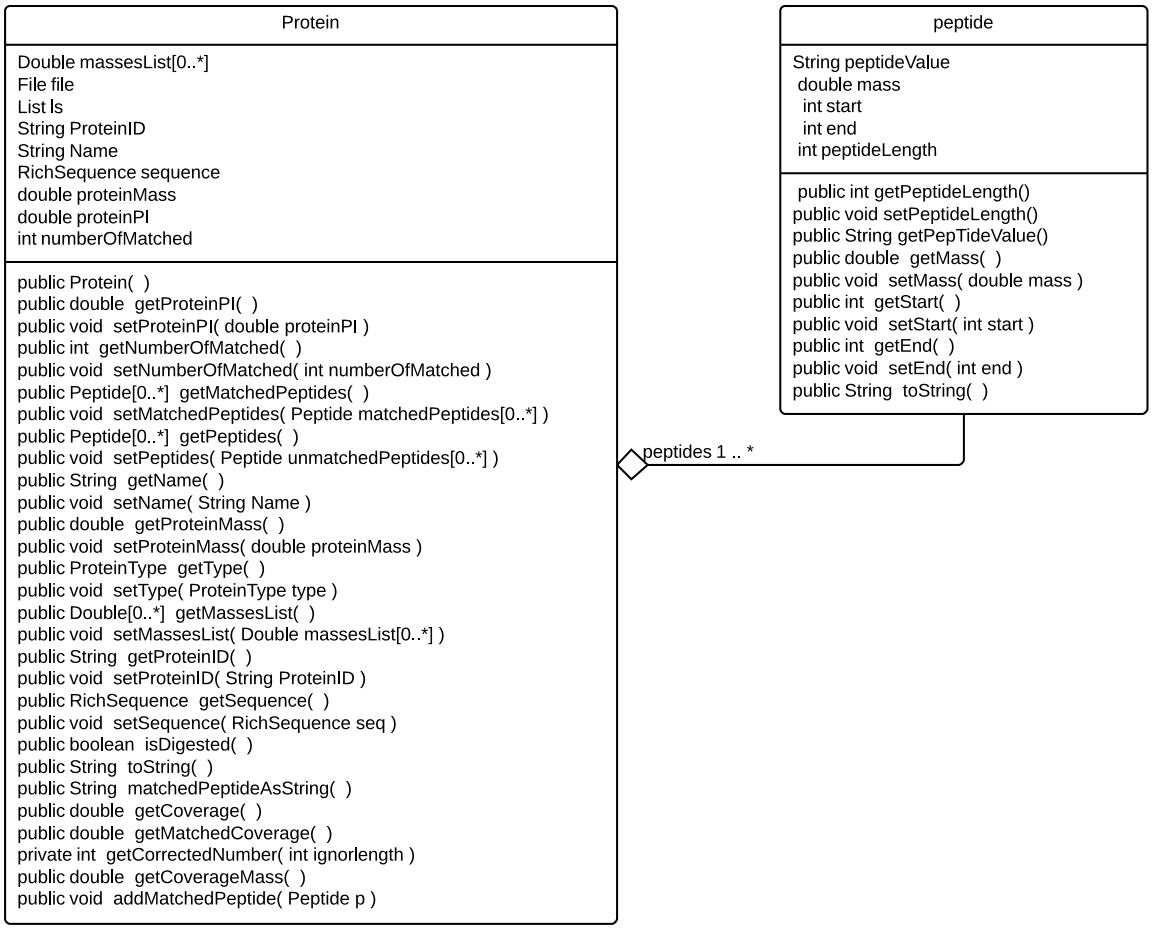


Figure B.1: Protein and Peptide classes.

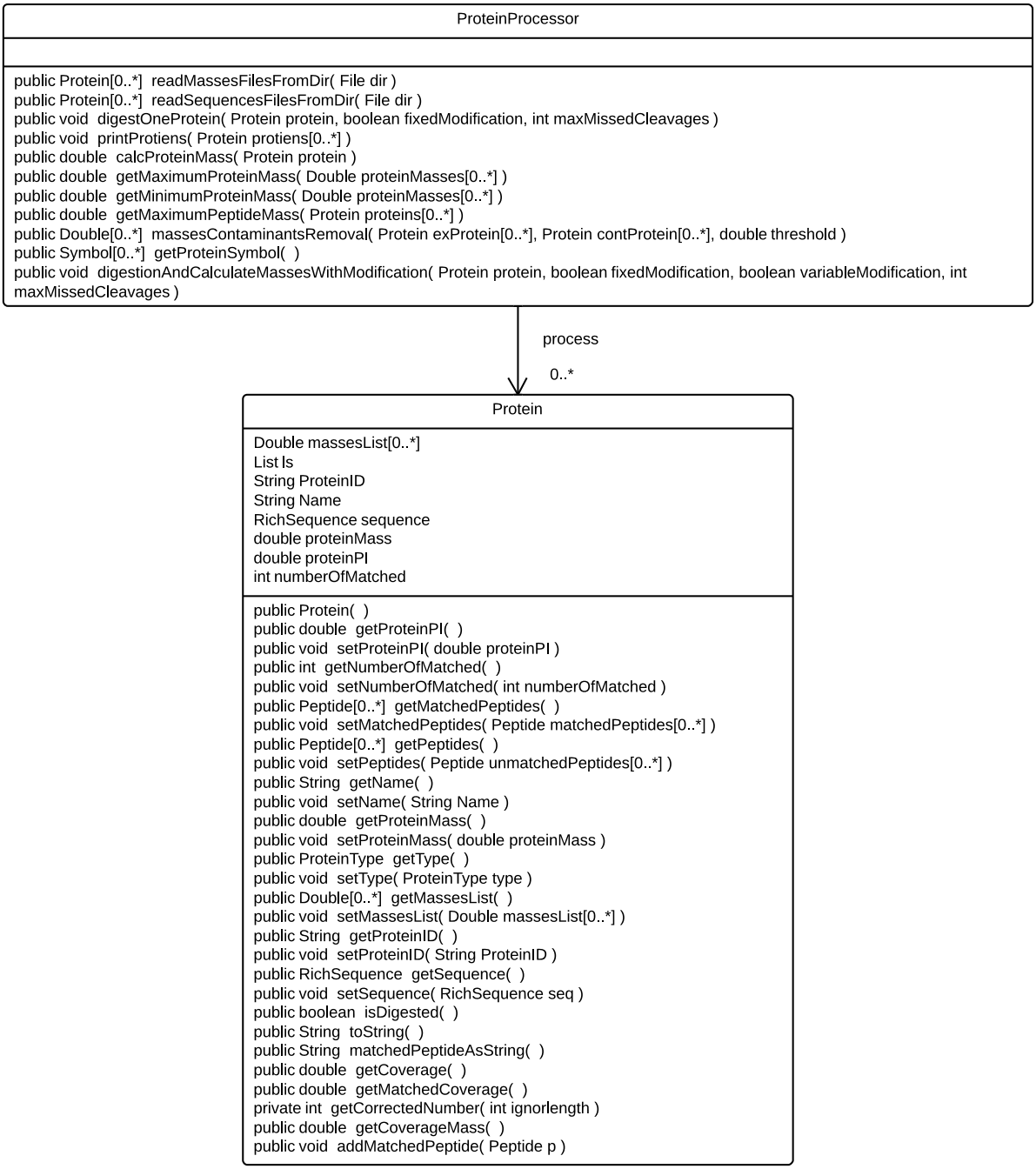


Figure B.2: Protein and ProteinProcessor classes.

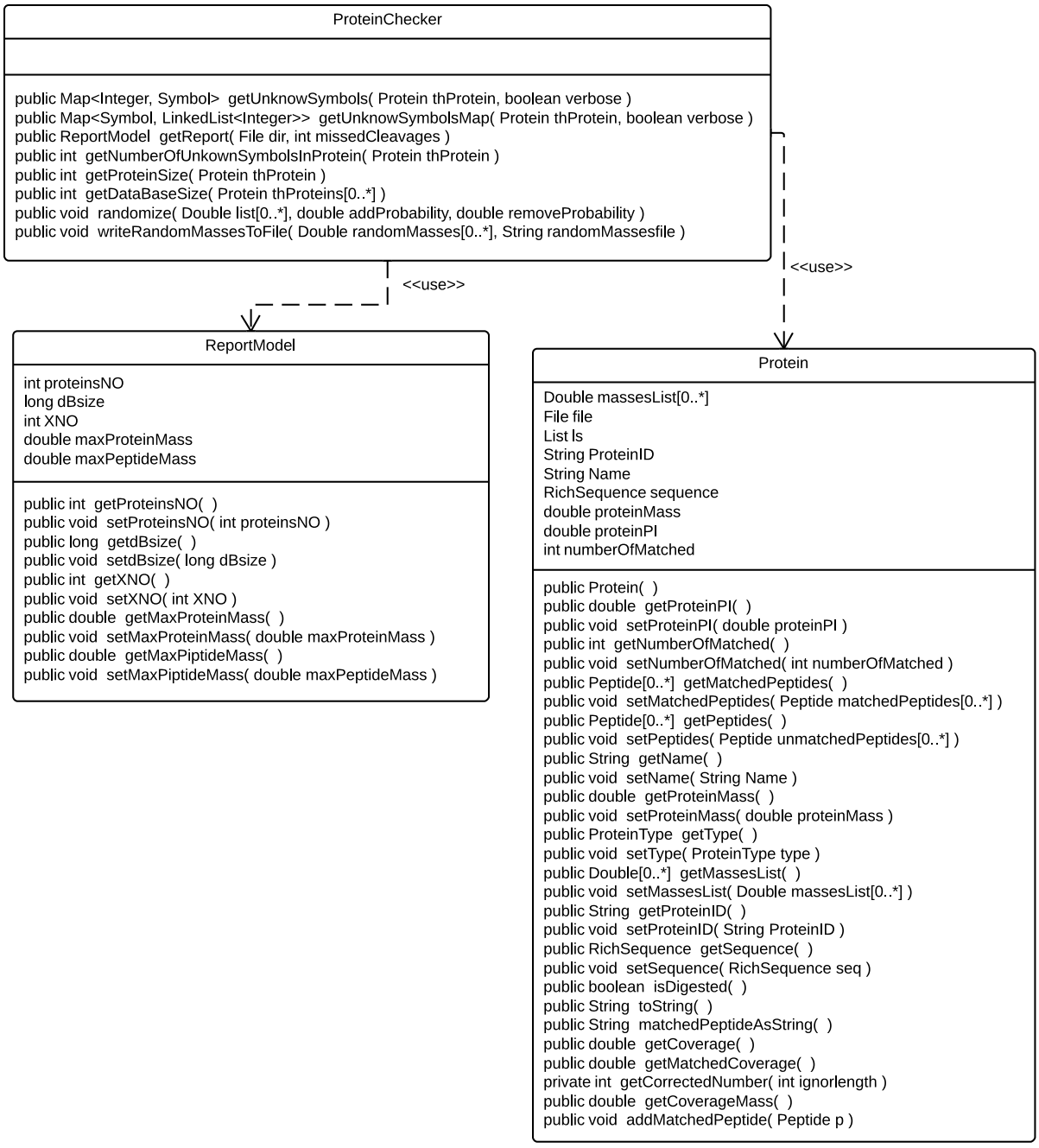


Figure B.3: Protein, ReportModel, and ProteinChecker classes.

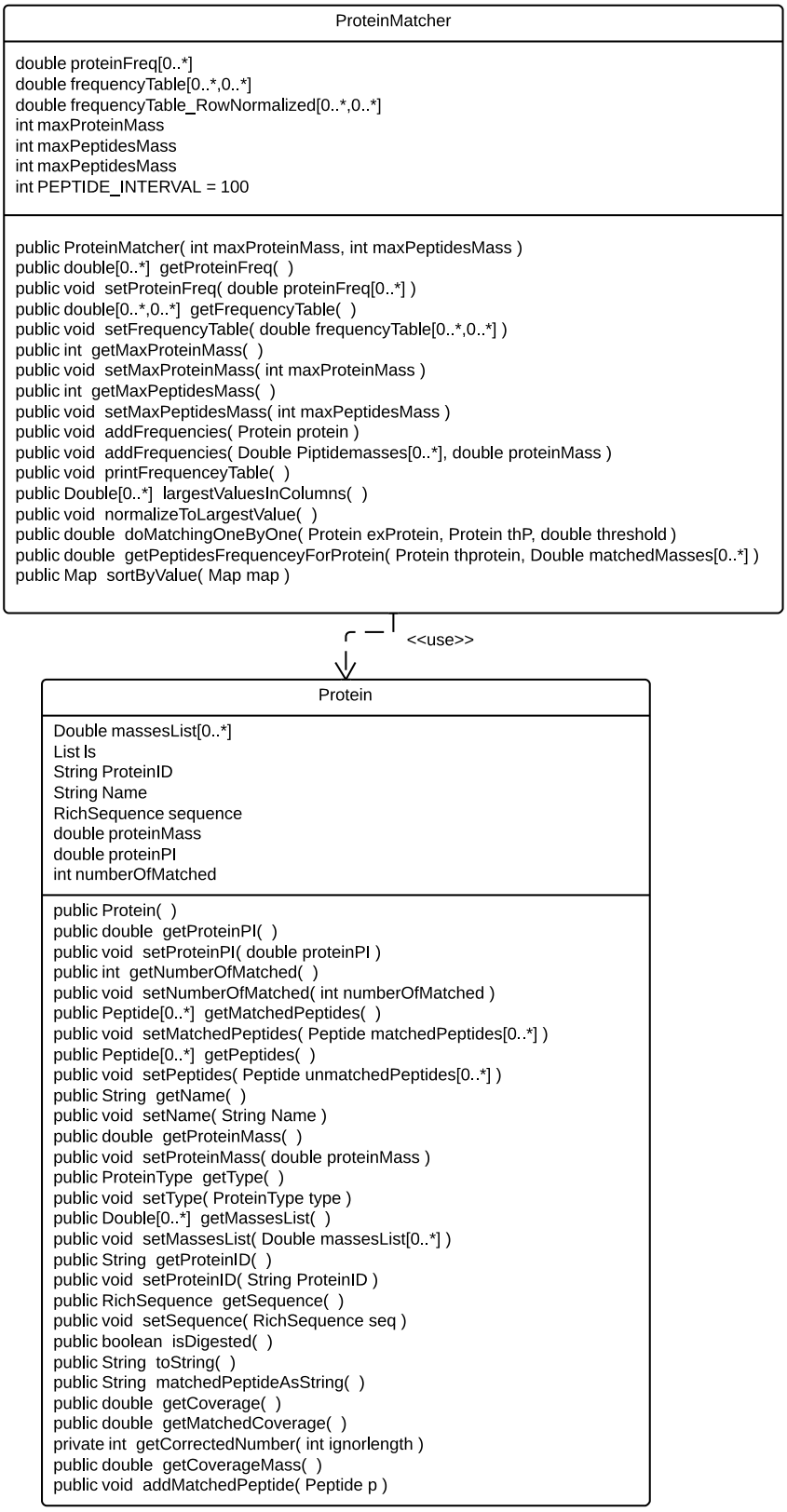


Figure B.4: Protein and ProteinMatcher classes.

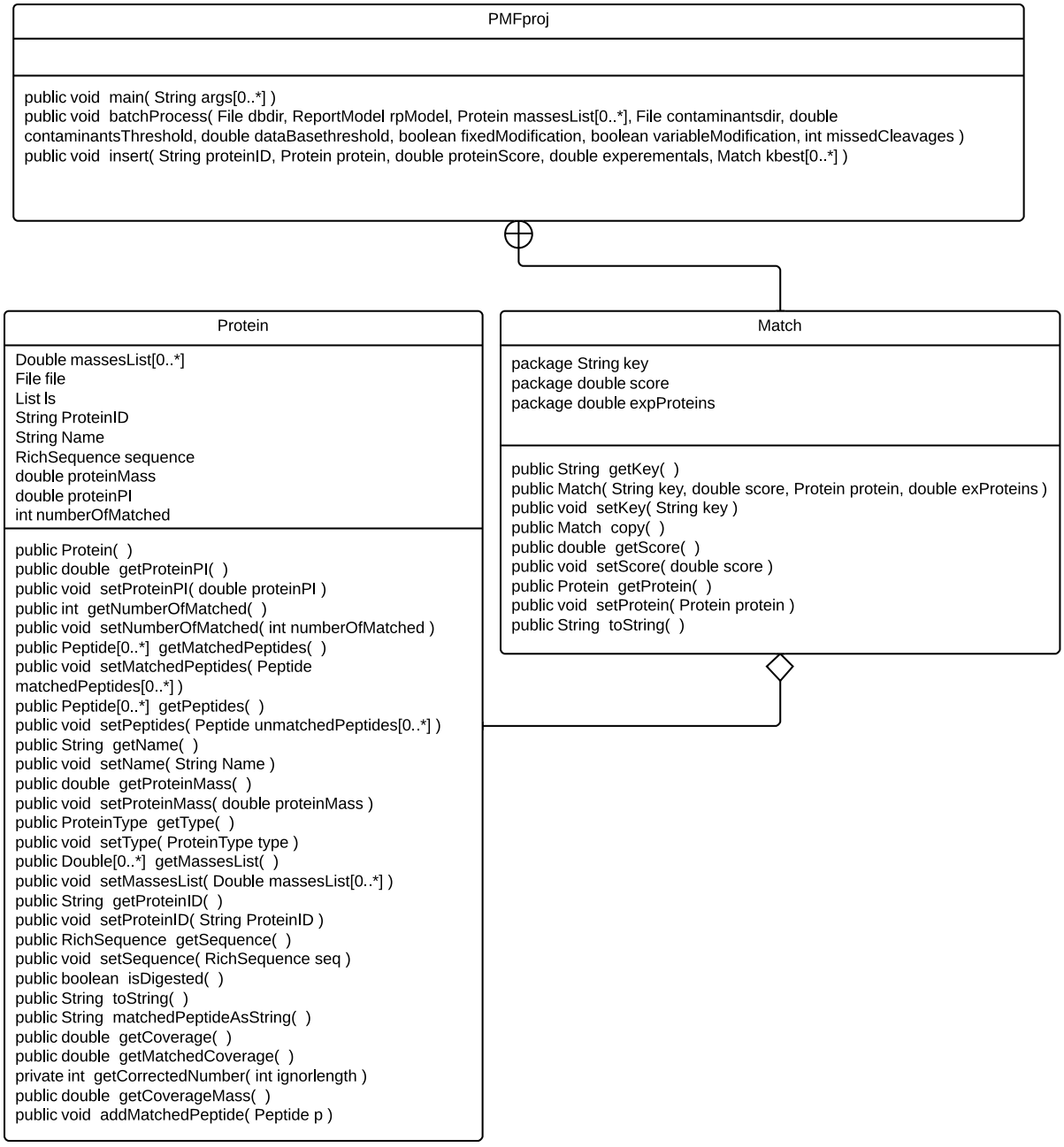


Figure B.5: PMFproj, Match, and Protein classes.

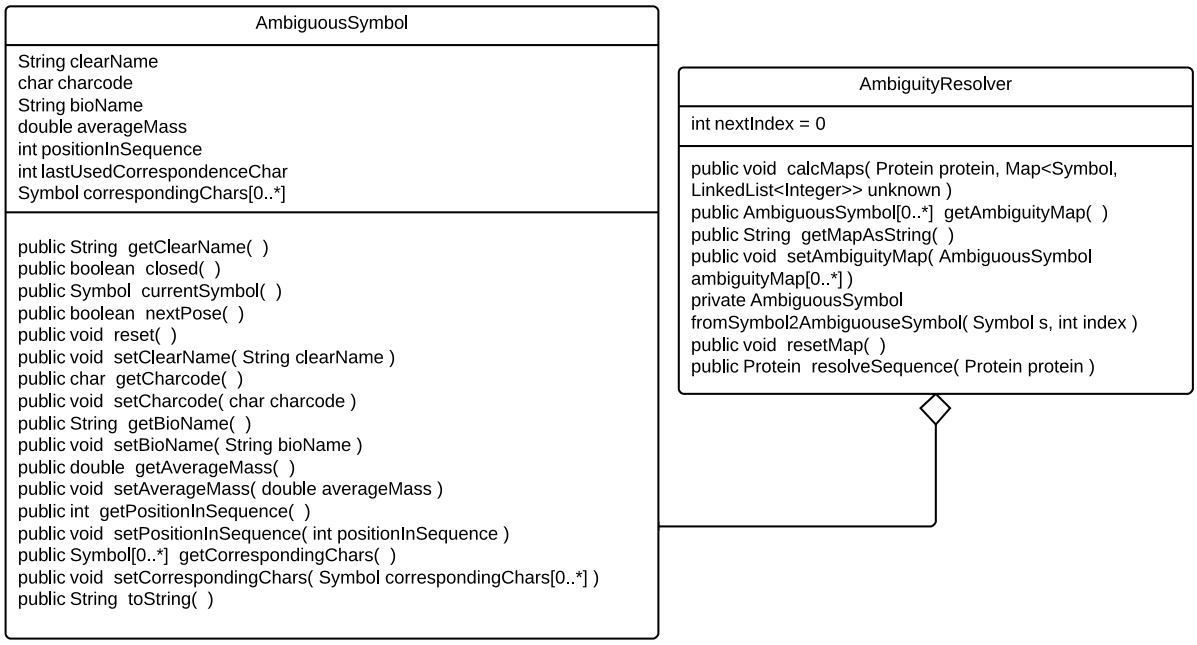


Figure B.6: Ambiguous Symbol and AmbiguityResolver classes.

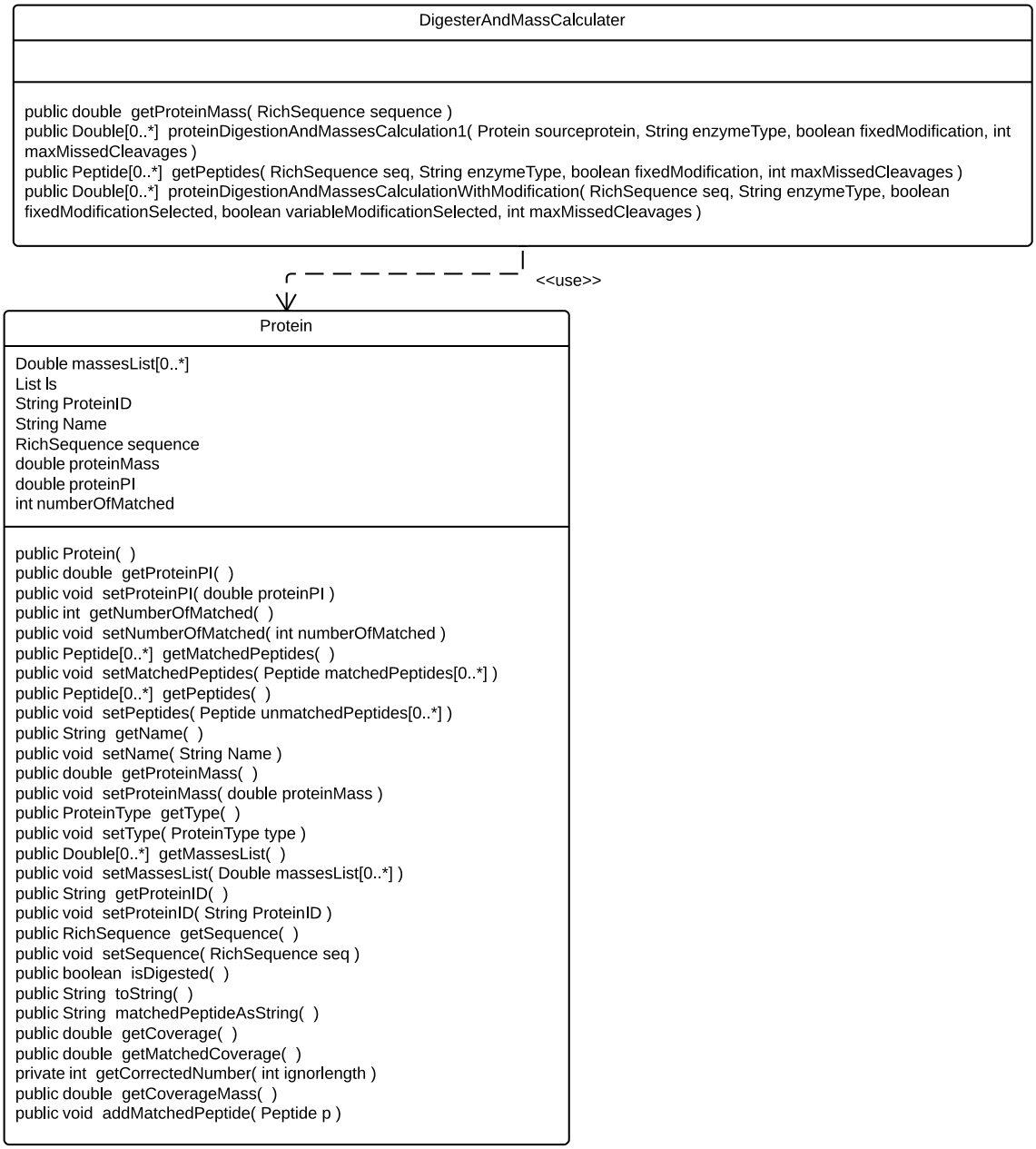


Figure B.7: DigesterAndMassCalculater and Protein classes.

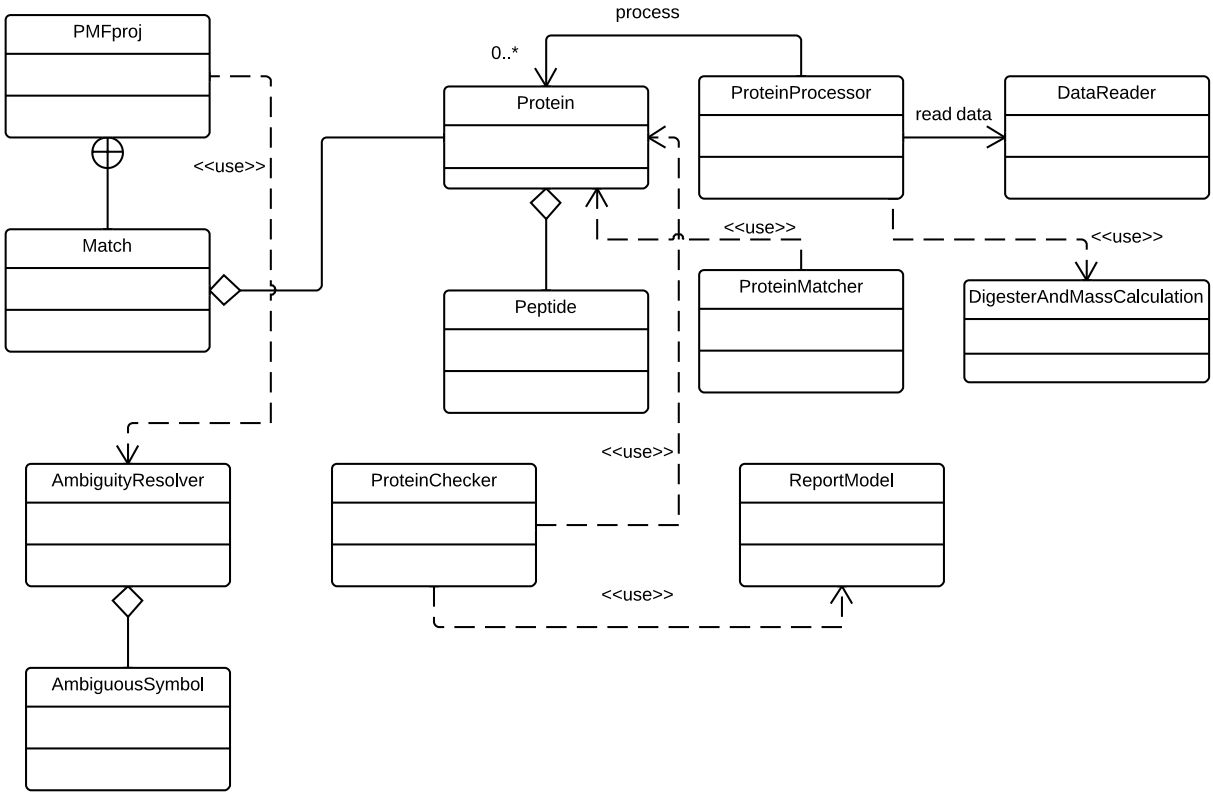


Figure B.8: UML classes relationships.

Bibliography

[1] H. A. Al Lawati. *Toward a microfluidic system for proteomics*. PhD thesis, University of Hull, 2007.

[2] R. D. Bagshaw, J. W. Callahan, and D. J. Mahuran. Desalting of in-gel-digested protein sample with mini-c18 columns for matrix-assisted laser desorption ionization time of flight peptide mass fingerprinting. *Analytical biochemistry*, 284(2):432–435, 2000.

[3] W. V. Bienvenut. *Acceleration and improvement of protein identification by mass spectrometry*. Springer, 2005.

[4] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365–370, 2003.

[5] D. C. Chamrad, G. Körting, K. Stühler, H. E. Meyer, J. Klose, and M. Blüggel. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, 4(3):619–628, 2004.

[6] A. Chernobrovkin, O. Trifonova, N. Petushkova, E. Ponomarenko, and A. Lisitsa. Selection of the peptide mass tolerance value for protein identification with peptide mass fingerprinting. *Russian Journal of Bioorganic Chemistry*, 37(1):119–122, 2011.

[7] K. R. Clauser, P. Baker, and A. L. Burlingame. Role of accurate mass measurement ($\pm$10 ppm) in protein identification strategies employing ms or ms/ms and database searching. *Analytical chemistry*, 71(14):2871–2882, 1999.

[8] J. Cottrell and U. London. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[9] M. Duncan, K. Fung, H. Wang, C. Yen, and K. Cios. Identification of contaminants in proteomics mass spectrometry data. In *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE*, pages 409–410. IEEE, 2003.

[10] F. Dycka. Identification and characterization of proteins regulated by plant hormones cytokinins by mass spectrometry.

[11] J. Eriksson and D. Fenyö. A model of random mass-matching and its use for automated significance testing in mass spectrometric proteome analysis. *Proteomics*, 2(3):262–270, 2002.

[12] J. Eriksson and D. Fenyö. Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results. *Journal of proteome research*, 3(1):32–36, 2004.

[13] D. Fenyö. Identifying the proteome: software tools. *Current Opinion in Biotechnology*, 11(4):391–395, 2000.

[14] Y. Fu, L.-Y. Xiu, W. Jia, D. Ye, R.-X. Sun, X.-H. Qian, and S.-M. He. Deltamt: a statistical algorithm for fast detection of protein modifications from lc-ms/ms data. *Molecular & Cellular Proteomics*, 10(5), 2011.

[15] A. Ganapathy, X.-F. Wan, J. Wan, J. Thelen, D. W. Emerich, G. Stacey, and D. Xu. Statistical assessment for mass-spec protein identification using peptide fingerprinting approach. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 2, pages 3051–3054. IEEE, 2004.

[16] E. Gasteiger, C. Hoogland, A. Gattiker, M. R. Wilkins, R. D. Appel, A. Bairoch, et al. Protein identification and analysis tools on the expasy server. In *The proteomics protocols handbook*, pages 571–607. Springer, 2005.

[17] E. Gasteiger, E. Jung, A. Bairoch, et al. Swiss-prot: connecting biomolecular knowledge via a protein database. *Current issues in molecular biology*, 3:47–56, 2001.

[18] W. R. Gilks, B. Audit, D. De Angelis, S. Tsoka, and C. A. Ouzounis. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, 18(12):1641–1649, 2002.

[19] R. Godavarti, M. Davis, G. Venkataraman, C. Cooney, R. Langer, R. Sasisekharan, et al. Heparinase iii from flavobacterium heparinum: cloning and recombinant expression in escherichia coli. *Biochemical and biophysical research communications*, 225(3):751–758, 1996.

[20] L. Hanley and R. Zimmermann. Light and molecular ions: the emergence of vacuum uv single-photon ionization in ms. *Analytical chemistry*, 81(11):4174–4182, 2009.

[21] Z. He, C. Yang, C. Yang, R. Z. Qi, J. Po-Ming Tam, and W. Yu. Optimization-based peptide mass fingerprinting for protein mixture identification. *Journal of Computational Biology*, 17(3):221–235, 2010.

[22] W. J. Henzel, C. Watanabe, and J. T. Stults. Protein identification: the origins of peptide mass fingerprinting. *Journal of the American Society for Mass Spectrometry*, 14(9):931–942, 2003.

[23] F. Hillenkamp, M. Karas, R. C. Beavis, and B. T. Chait. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical Chemistry*, 63(24):1193A–1203A, 1991.

[24] R. Jagtap and A. Ambre. Overview literature on matrix assisted laser desorption ionization mass spectroscopy (maldi ms): basics and its applications in characterizing polymeric materials. *Bulletin of Materials Science*, 28(6):515–528, 2005.

[25] O. N. Jensen, M. Wilm, A. Shevchenko, and M. Mann. Sample preparation methods for mass spectrometric peptide mapping directly from 2-de gels. In *2-D Proteome Analysis Protocols*, pages 513–530. Springer, 1999.

[26] R. S. Johnson, M. T. Davis, J. A. Taylor, and S. D. Patterson. Informatics for protein identification by mass spectrometry. *Methods*, 35(3):223–236, 2005.

[27] T. M. Karve and A. K. Cheema. Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease. *Journal of amino acids*, 2011, 2011.

[28] F. Levander, T. Rögnvaldsson, J. Samuelsson, and P. James. Automated methods for improved protein identification by peptide mass fingerprinting. *Proteomics*, 4(9):2594–2601, 2004.

[29] T. Li, K. Fan, J. Wang, and W. Wang. Reduction of protein sequence complexity by residue grouping. *Protein Engineering*, 16(5):323–330, 2003.

[30] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[31] X. Liang, A. Kaya, Y. Zhang, D. Le, D. Hua, and V. Gladyshev. Characterization of methionine oxidation and methionine sulfoxide reduction using methionine-rich cysteine-free proteins. *BMC biochemistry*, 13(1):21, 2012.

[32] J. A. Loo et al. Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrometry Reviews*, 16(1):1–23, 1997.

[33] G. Lubec, L. Afjehi-Sadat, et al. Limitations and pitfalls in protein identification by mass spectrometry. *Chemical Reviews-Columbus*, 107(8):3568–3584, 2007.

[34] S. Magdeldin, Y. Zhang, B. Xu, Y. Yoshida, and T. Yamamoto. Two-dimensional polyacrylamide gel electrophoresis–a practical perspective.

[35] S. D. Maleknia and R. Johnson. Mass spectrometry of amino acids and proteins. *Carbon*, 12(12C):12, 2012.

[36] L. McHugh and J. W. Arthur. Computational methods for protein identification from mass spectrometry data. *PLoS computational biology*, 4(2):e12, 2008.

[37] K. F. Medzihradszky, J. M. Campbell, M. A. Baldwin, A. M. Falick, P. Juhasz, M. L. Vestal, and A. L. Burlingame. The characteristics of peptide collision-induced dissociation using a high-performance maldi-tof/tof tandem mass spectrometer. *Analytical chemistry*, 72(3):552–558, 2000.

[38] K. Mizuguchi, C. M. Deane, T. L. Blundell, M. S. Johnson, and J. P. Overington. Joy: protein sequence-structure representation and analysis. *Bioinformatics*, 14(7):617–623, 1998.

[39] A. Pandey and M. Mann. Proteomics to study genes and genomes. *Nature*, 405(6788):837–846, 2000.

[40] D. Pappin, P. Hojrup, and A. Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. *Current biology*, 3(6):327–332, 1993.

[41] K. C. Parker. Scoring methods in maldi peptide mass fingerprinting: Chemscore, and the chemapplex program. *Journal of the American Society for Mass Spectrometry*, 13(1):22–39, 2002.

[42] U. Pieles, W. Zürcher, M. Schär, and H. Moser. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a powerful tool for the mass and sequence analysis of natural and modified oligonucleotides. *Nucleic acids research*, 21(14):3191–3196, 1993.

[43] J. Samuelsson, D. Dalevi, F. Levander, and T. Rögnvaldsson. Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics*, 20(18):3628–3635, 2004.

[44] M. Science. Amino acid reference data@ONLINE, 2013.

[45] I. Shadforth, D. Crowther, and C. Bessant. Protein and peptide identification algorithms using ms for use in high-throughput, automated pipelines. *Proteomics*, 5(16):4082–4095, 2005.

[46] J. A. Siepen, E.-J. Keevil, D. Knight, and S. J. Hubbard. Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *Journal of proteome research*, 6(1):399–408, 2007.

[47] Z. Song and D. Adviser-Xu. *Bioinformatics methods for protein identification using peptide mass fingerprinting data*. University of Missouri at Columbia, 2009.

[48] Z. Song, L. Chen, A. Ganapathy, X.-F. Wan, L. Brechenmacher, N. Tao, D. Emerich, G. Stacey, and D. Xu. Development and assessment of scoring functions for protein identification using pmf data. *Electrophoresis*, 28(5):864–870, 2007.

[49] Z. Song, L. Chen, C. Zhang, and D. Xu. Design and implementation of probability-based scoring function for peptide mass fingerprinting protein identification. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 4556–4559. IEEE, 2006.

[50] W. Thammasorn, K. Eadjongdee, A. Hongsthong, K. Porkaew, S. Cheevadhanarak, et al. Probability-based scoring function as a software tool used in the genome-based identification of proteins from spirulina platensis. *The Open Bioinformatics Journal*, 3:59–68, 2009.

[51] B. Thiedea, W. Höhenwarterb, A. Kraha, J. Mattowc, M. Schmidb, F. Schmidtb, and P. R. Jungblutb. Peptide mass fingerprinting. *Methods*, 35:237–247, 2005.

[52] A. Tiengo, N. Barbarini, S. Troiani, L. Rusconi, and P. Magni. A perl procedure for protein identification by peptide mass fingerprinting. *BMC bioinformatics*, 10(Suppl 12):S11, 2009.

[53] L. R. Yetukuri and G. V. Peddinti. Protein identification with mascot software.

[54] W. Zhang and B. T. Chait. Profound: an expert system for protein identification using mass spectrometric peptide mapping information. *Analytical chemistry*, 72(11):2482–2489, 2000.