The accuracy of statistical confidence estimates in shotgun proteomics

Viktor Granholm
Abstract

High-throughput techniques are currently some of the most promising methods to study molecular biology, with the potential to improve medicine and enable new biological applications. In proteomics, the large scale study of proteins, the leading method is mass spectrometry. At present researchers can routinely identify and quantify thousands of proteins in a single experiment with the technique called shotgun proteomics.

A challenge of these experiments is the computational analysis and the interpretation of the mass spectra. A shotgun proteomics experiment easily generates tens of thousands of spectra, each thought to represent a peptide from a protein. Due to the immense biological and technical complexity, however, our computational tools often misinterpret these spectra and derive incorrect peptides. As a consequence, the biological interpretation of the experiment relies heavily on the statistical confidence that we estimate for the identifications.

In this thesis, I have included four articles from my research on the accuracy of the statistical confidence estimates in shotgun proteomics; how to accomplish it and evaluate it. In the first two papers a new method to use pre-characterized protein samples to evaluate this accuracy is presented. The third paper deals with how to avoid statistical inaccuracies when using machine learning techniques to analyze the data. In the fourth paper, we present a new tool for analyzing shotgun proteomics results, and evaluate the accuracy of its statistical estimates using the method from the first papers.

The work I have included here can facilitate the development of new and accurate computational tools in mass spectrometry-based proteomics. Such tools will help making the interpretation of the spectra and the subsequent biological conclusions more reliable.
List of publications

I have included the following articles in the thesis.

PAPER I:  **On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics.**
Viktor Granholm, William Stafford Noble & Lukas Käll
*Journal of Proteome Research, 10(5)*, 2671–2678 (2011).
DOI: 10.1021/pr1012619

PAPER II:  **Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics.**
Viktor Granholm, José Fernández Navarro, William Stafford Noble & Lukas Käll
*Journal of Proteomics, 80(0)*, 123–131 (2013).
DOI: 10.1016/j.jprot.2012.12.007

PAPER III:  **A cross-validation scheme for machine learning algorithms in shotgun proteomics.**
Viktor Granholm, William Stafford Noble & Lukas Käll
*BMC Bioinformatics, 13(Suppl 16)*, S3 (2012).
DOI: 10.1186/1471-2105-13-S16-S3

PAPER IV:  **Fast and accurate database searches with MS-GF+Percolator**
Viktor Granholm, Sangtae Kim, José Fernández Navarro, Erik Sjölund, Richard Smith & Lukas Käll
DOI: 10.1021/pr400937n

The articles are printed here with permission from the respective publishers.
Other publications that are not included in the thesis.

**Quality assessments of peptide-spectrum matches in shotgun proteomics**
Viktor Granholm & Lukas Käll
DOI: 10.1002/pmic.201000432

**Mass fingerprinting of complex mixtures: protein inference from high-resolution peptide masses and predicted retention times**
Luminita Moruz, Michael Hoopmann, Magnus Rosenlund, Viktor Granholm, Robert Moritz & Lukas Käll
DOI: 10.1021/pr400705q

**Membrane protein shaving with thermolysin can be used to evaluate topology predictors**
Maria Bendz, Marcin Skwark, Daniel Nilsson, Viktor Granholm, Susana Cristobal, Lukas Käll & Arne Elofsson
*Proteomics*, 13(9), 1467–1480 (2013).
DOI: 10.1002/pmic.201200517

**HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics**
Rui Branca, Lukas Orre, Henrik Johansson, Viktor Granholm, Mikael Huss, Åsa Pérez-Bercoff, Jenny Forshed, Lukas Käll & Janne Lehtiö
DOI: 10.1038/nmeth.2732
Contents

Abstract

List of publications

Abbreviations

1 Introduction
   1.1 The scientific method
   1.2 Questions of investigation
   1.3 Thesis overview

2 Biological overview
   2.1 Life
   2.2 DNA and RNA
   2.3 Proteins

3 Statistical overview
   3.1 Statistical tests
   3.2 Machine learning

4 Data collection in mass spectrometry-based proteomics
   4.1 Liquid chromatography
   4.2 Mass spectrometry
   4.3 Shotgun proteomics

5 Data analysis in mass spectrometry-based proteomics
   5.1 Identification of spectra
   5.2 Confidence estimation methods
   5.3 Identification of peptides and proteins

6 Present investigation
   6.1 Paper I overview
   6.2 Paper II overview
6.3 Paper III overview ........................................ 36
6.4 Paper IV overview ........................................ 37

7 Future perspectives ........................................ 39

Sammanfattning på svenska .................................. xli

Acknowledgements ............................................. xliii

References ...................................................... xlv
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>CID</td>
<td>Collision-induced dissociation</td>
</tr>
<tr>
<td>Da</td>
<td>Dalton</td>
</tr>
<tr>
<td>ECD</td>
<td>Electron capture dissociation</td>
</tr>
<tr>
<td>ESI</td>
<td>Electrospray ionization</td>
</tr>
<tr>
<td>ETD</td>
<td>Electron transfer dissociation</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>G</td>
<td>Guanine</td>
</tr>
<tr>
<td>HCD</td>
<td>Higher-energy collisional dissociation</td>
</tr>
<tr>
<td>LC</td>
<td>Liquid chromatography</td>
</tr>
<tr>
<td>MS-GF+</td>
<td><em>Software name:</em> The mass spectrometry generation function</td>
</tr>
<tr>
<td>MS/MS</td>
<td>Tandem mass spectrometry</td>
</tr>
<tr>
<td>m/z</td>
<td>mass-to-charge</td>
</tr>
<tr>
<td>Ph.D.</td>
<td>Doctor of philosophy</td>
</tr>
<tr>
<td>PSM</td>
<td>Peptide-spectrum match</td>
</tr>
<tr>
<td>SVM</td>
<td>Support vector machine</td>
</tr>
<tr>
<td>T</td>
<td>Tyrosine</td>
</tr>
<tr>
<td>Th</td>
<td>Thomson</td>
</tr>
</tbody>
</table>
1. Introduction

1.1 The scientific method

The basis of science is not knowledge. In fact, regarding knowledge as something fixed is the exact opposite of science. On the other hand, scientists regard the scientific method as rather fixed. Science is a method to change, and update, our knowledge of the world, and in many cases completely overturn what we previously believed. The method requires setting up a practically testable model, hypothesis or prediction, and to test it in a reproducible way. The result will help us understand the validity of the model.

From a scientific perspective, the actual testing procedure is obviously crucial. The test must be well designed, for example to avoid that the testing procedure itself influences the results [1]. Furthermore, the outcome of the test must be measurable. In this thesis I address the latter point, applied to a part of the field of mass spectrometry-based proteomics. For interpreting the tests, the measurements in this field are accompanied by confidence estimates, and here I deal with the accuracy of these estimates.

1.2 Questions of investigation

In mass spectrometry-based proteomics, one wants to characterize biological samples, such as cells or blood plasma, with respect to their protein content. Protein concentrations vary enormously, some proteins can be present in $10^{10}$ as many copies as others [2], so the technical requirements of the task are breathtaking. Although protein microarrays also permit large scale proteomics studies [3], there is currently no alternative for identifying and quantifying proteins to the extent that mass spectrometry enables. One consequence is that the results can not be directly validated by other technologies, and the quality of the results remain uncertain. From a scientific perspective, however, the quality issue is critical. We need to know the accuracy of the results in order to draw biological conclusions.

This thesis covers how to evaluate the quality of central intermediary results from the most common mass spectrometry-based proteomics technique, shotgun proteomics. More specifically, it deals with the accuracy of the statis-
tical confidence that is assigned to peptides identified from mass spectrometry. In the first two articles included in the thesis, we investigate how biological samples with known protein content can be used to evaluate the accuracy of the statistical confidence estimates. In the next article, we deal with the risks of inaccuracies when adjusting statistical models to each experiment at hand. In the last article, we combine two previously described computational tools used in mass spectrometry-based proteomics, and investigate how they perform together and whether their combination evaluates well using the methods in the first articles.

1.3 Thesis overview

The following chapters serve as an introduction to the research field of the four included publications. I have organized the chapters into a few topics that I believe will cover the most important background information needed to understand the articles on which the thesis is based. In the next chapter, I start off by introducing some basic biology. This is followed by a chapter with an overview of statistics. I then continue by describing how to collect proteomics data from mass spectrometry, before covering the computational procedures of how the data is analyzed. In the last two chapters I summarize the included articles and briefly outline my perspectives for future research in the field of mass spectrometry-based proteomics.
2. Biological overview

As with any technical topics, this thesis requires reading up on some background information. Fortunately, for both the reader and me, much of the background to this work concerns one of the most unbelievable and interesting phenomena that exist, Life. Here, I therefore briefly introduce some selected topics of life from a molecular perspective (one of certainly many perspectives).

2.1 Life

There is no clear definition of what life is, but some properties seem to be shared between virtually all living organisms. Growth, reproduction, metabolism and the organization into cells are typical such characteristics. Cells are often regarded as the smallest units of life. They are closed containers that can live individually, or in groups to form a larger organism like yourself. Among many other groups of molecules, they contain DNA, RNA and proteins, which are further explained below. Note that I have excluded viruses from this introduction as they can be argued to not be alive. Furthermore, it appears like all forms of life comply to the rule called the central dogma. Somewhat simplified, the central dogma states that the flow of genetic information in living cells is directed, as it goes from DNA, through RNA to proteins, and not the other way, see Figure 2.1. Out of these three groups of molecules, it is mainly the DNA that is inherited to new generations. Its main function is to store the information needed to run the cell, or organism [4].

![Figure 2.1: The central dogma.](image)

The flow of genetic information goes from the DNA through the RNA to proteins.
2.2 DNA and RNA

DNA (or deoxyribonucleic acid) is a molecule to store information about RNA and proteins. It consists of two anti-parallel strands of nucleotides, bound together by nucleobase pairs. The entire structure forms a characteristic helical shape. Whereas our electronic storage devices tend to store binary information using bits, DNA stores quaternary information using four different nucleobases, adenine (A), tyrosine (T), guanine (G) and cytosine (C). The nucleobases typically pair up as A with T and G with C between the two DNA strands, this helps to efficiently copy the strands during replication. Some organisms contain circular DNA molecules, whereas others contain multiple linear molecules. The entire set of all DNA in a cell, or organism, is called the genome. Furthermore, a gene is often defined as a distinct section of the genome that holds the information about an RNA or a protein [4].

RNA (ribonucleic acid) also contains nucleobases, but has a more versatile structure and a wide range of functions. The cells contain many types and shapes of RNA molecules, that tend to be single-stranded and hence very flexible. An important role of RNA is its part in the translation of genes in DNA into proteins, performed by a class of RNA that is called messenger RNA, or mRNA. As explained below, the protein alphabet consists of 20 letters (amino acids), whereas the nucleobase alphabet only has four. Thus, the nucleobases are translated in triplets, called codons. As the four nucleobases give $4^3 = 64$ different such codon combinations, the number of different codons is more than enough to code for the 20 amino acids. In fact, many codons code for the same amino acid. Using a process called alternative splicing, in which parts of a gene are skipped, a single gene can produce several different mRNA molecules. Alternative splicing offers an additional level of cell regulation, as these mRNA molecules also will code for different proteins [4].

2.3 Proteins

Although RNA molecules perform a range of functions in the cell, they are limited in versatility compared to proteins. Proteins are made up of sequences of amino acids. The amino acids are small molecules that share a common “back-bone”, but that each have characteristic side-chain structures. There are 20 different amino acids, some of which are denoted with letters that overlap with the nucleobase abbreviations. The chemical bonds that hold amino acids together in a protein are called peptide bonds, and are mentioned later in the thesis. Different combinations of typically a few hundreds of these amino acids in sequence give proteins with completely different shapes, sizes and functions. But the proteins can be even further amended with post-translational modifica-
tions; extra molecules that are added to the amino acids of a protein during its lifetime to enable more advanced and specialized functions. Figure 2.2 shows a simple illustration of a protein [4].

Figure 2.2: A schematic representation of a protein. A protein can be thought of as a sequence of amino acids (represented by circles with letters inside) that adopts a three-dimensional form. Some of the amino acids may be post-translationally modified (squares).

What makes up a distinct type of protein can be hard to define. With alternative splicing a single gene can produce proteins with different amino acid sequences. Furthermore, they can differ with respect to post-translational modifications and charges. While many differences have specific functions, others may not, and the distinctions between them may be irrelevant. Proteins can also group into stable clusters, called complexes. The lack of definition makes it hard to count the number of distinct proteins in a cell, estimates vary from tens of thousands to several millions [5].

Different proteins are responsible for a large variety of cellular functions, ranging from for example photosynthesis to immune defense. Frequently, more than one protein is involved in these functions, creating complex networks of co-dependent proteins. The abundance of the proteins also varies enormously, sometimes with over ten orders of magnitude between different proteins [2]. Their concentrations can also vary over time, to adjust the metabolism for different conditions during the day, for example. The types, states and quantities of all proteins in a cell at a given time point, is called the proteome. Variations in the proteome explain how cells in multicellular organisms can specialize, into for example muscle or skin cells, despite sharing the same genome. The field of studying a proteome at a large scale is called proteomics.

The motivation behind proteomics research is that monitoring the proteins, on top of the DNA and RNA, should give a more complete picture of the cell. One can use efficient sequencing techniques to relatively cheaply measure mRNA concentrations of cells [6], but these do not seem to be completely correlated to the protein content [7; 8]. Furthermore, the mRNAs do not give much information about protein lifetimes, cellular locations or post-
translational modifications. The idea of studying the proteome at a large scale, instead of a few proteins, is motivated by that it enables a holistic approach to biology [9]. It may be, that such a holistic system-level understanding is required to enable efficient development of medicines or diagnoses, and to make way for other applications of biological systems.

2.3.1 Enzymes

Proteins that act as catalysts to speed up chemical reactions are called enzymes. A group of enzymes that is of particular technical importance for mass spectrometry-based proteomics, is called proteolytic enzymes, or proteases. These enzymes cleave the peptide bonds of proteins into shorter amino acid sequences, called peptides. In many cases, a specific enzyme is specialized at cleaving proteins at certain sequence patterns. A relevant example of a protease is trypsin, which cleaves proteins next to the basic amino acids arginine and lysine [10].

2.3.2 Protein mass

The mass of a physical body is a property that describes its gravitational attraction to other bodies and its resistance against acceleration. As the mass depends on their elementary composition, we can use it to identify molecules. The elementary compositions and the masses of all amino acids are known. The amino acids are composed from the elements carbon, hydrogen, oxygen, nitrogen and sulfur. In addition, we also know the masses of many post-translational modifications. However, each element comes in a few different forms, called isotopes. The isotopes of an element are distinguished by their numbers of neutrons, hence each element can have a couple of different masses. Thus, proteins or peptides often exist with slightly different masses because of random variations in their isotope compositions.
3. Statistical overview

Here, I refer to statistics as a method for drawing conclusions from data. There are two main approaches to probability and statistics, denoted frequentist (or classical) and Bayesian statistics [11; 12]. As the work in this thesis can be completely understood from a frequentist perspective, I focus on this approach here. This chapter first covers some basic statistical concepts, and then proceeds to briefly introduce machine learning, and support vector machines in particular.

3.1 Statistical tests

In statistics, we call the entire set of entities we are interested in the population. Although the population could be abstract, with a frequentist perspective we assume it can be described by some fixed population parameters. The population is often very large or even infinite, however, so these parameters may be impossible to measure and generally remain unobserved. Instead, we draw smaller samples from the population and make our measurements on those. Estimating the population parameters from the measurements on the samples is called inference. Furthermore, when we make a statement about the population parameters, we call it a hypothesis. An example population is all past, present and future patients subject to new medical treatment, and a typical sample from this population would be a subset of the present patients. A little more abstract example is the population of all (infinite) potential rolls of a six-sided die. Typical population parameters are the expected frequencies of the six sides, for balanced dice these are all $1/6$. A sample from this population is any finite number of rolls [11].

We use statistical tests to choose between different hypotheses. The classical test contains a null hypothesis that is compared to an alternative hypothesis [11]. The null hypothesis commonly represents a default, or a normal situation. For example, it could represent that the die is balanced. The alternative hypothesis generally represents our new idea or interest; that the die is unbalanced. To decide between the two hypotheses, we make a test and observe the outcome. We summarize the observations in a single value that is a function of the sample, a test statistic. This value is then used when making
the decision. An example test statistic for testing the balance of the die is the chi-squared ($\chi^2$) test statistic. We obtain it with the equation

$$\chi^2 = \sum_{n=1}^{6} \frac{(O_n - E_n)^2}{E_n},$$

where $n$ is the side of the die, and $O_n$ and $E_n$ are the observed and expected frequencies of each side, respectively. The higher the chi-squared test statistic, the less probable are the results, given that the die is balanced. In many cases the researchers choose a test statistic threshold over which they say the test is significant, meaning that there is enough reason to believe in the alternative hypothesis. If we would choose to believe the alternative hypothesis, it is called to reject the null hypothesis. Note that with the decision, we risk making one of two mistakes. First, we could erroneously reject the null hypothesis when it is in fact true. Second, we could accept the null hypothesis when it is in fact false.

To help decide a threshold test statistic above which we say all tests are significant, we want to associate the threshold with some confidence measure. For example, how would we choose an appropriate chi-squared test statistic over which we say the die is unbalanced? In the following sections I will cover the meaning of two such measures, the $p$ value and the false discovery rate. In two different ways, they describe the expected probability of making a mistake when using a given threshold to make decisions. To estimate the confidence measures, we should take properties such as the variability of the test statistics into consideration. In chapter 5 we will discuss how these confidence measures are estimated in practice, in mass spectrometry-based proteomics.

### 3.1.1 $p$ values

The $p$ value has traditionally been one of the most frequently used confidence measures, although it is often misused or interpreted incorrectly [13]. For a given observation, it is the probability of observing an equal or a more extreme observation given that the null hypothesis is true [11]. Like above, the observation is formulated in the form of a test statistic. If we would use a $p$ value threshold to decide whether to believe in the null or the alternative hypothesis for an experiment, we hence control the proportion of true null tests that we erroneously call significant. In Figure 3.1 I have illustrated the test statistic distributions of a null (left) and an alternative (right) hypothesis. Given that the observations in the figure are independent and identically distributed, the $p$ value that corresponds to the threshold is the fraction of area $C$ in proportion to the entire null distribution, area $A+B+C$. In the case of the die example above, the $p$ value is the probability that a balanced die would get an equal or higher chi-squared test statistic than the threshold. If this probability is very low, we might consider rejecting the null hypothesis and say that the die is
unbalanced.

**Figure 3.1:** Example distributions of a null and an alternative hypothesis. The two curves each represent a hypothetic test statistic distribution from an infinite number of independent and identically distributed observations. The left side curve represents the observations where the null hypothesis is true, and the right side curve where the alternative hypothesis is true. A threshold can be used to decide between the two hypotheses. However, with the decision there is a risk for mistakes, as the test statistic distributions may overlap.

An interesting (and useful) consequence of how the \( p \) value is defined is that \( p \) values of multiple tests that are truly null will distribute uniformly between the probabilities 0 and 1. Note that the test statistics may distribute arbitrarily, but the \( p \) values corresponding to them will still be uniformly distributed. This can be understood by considering that a \( p \) value equal or lower than \( x \) will occur with a proportion of \( x \) among the null tests. For example, 100% of the null tests will have a \( p \) value lower than 1, and 5% will have \( p \) value lower than 0.05. In contrast, we hope that the tests that are not truly null will have \( p \) values shifted towards 0. The uniformity feature of the null \( p \) values make them attractive for testing the methods and assumptions we use to estimate our confidence measures. If we were given a set of tests where the null hypothesis is known to be true, any reasonable confidence estimation method would produce \( p \) values that lie close to the uniform distribution. In Paper I and Paper II we demonstrate how this feature of the \( p \) values can be used to test confidence estimation procedures in shotgun proteomics.

When trying to draw conclusions from datasets with many tests, the \( p \) values are unfortunately hard to interpret. To understand why, consider the following. Imagine that we have 10,000 dice, and we want to test to filter out some potentially biased ones. We roll each die 500 times, and with the same chi-squared test statistic as above, we assign a \( p \) value to each of the dice. (In this case, we can calculate the \( p \) values using a chi-squared distribution with 5 degrees of freedom.) We then use the seemingly stringent \( p \) value of 0.001 as a threshold for unbalanced dice. This means that only 0.1% of the balanced dice will get such a chi-squared test statistic or higher, by chance. Further assume
that we discovered that 12 dice got significant test statistics with this threshold. We might be tempted to conclude that these 12 dice are unbalanced. However, since we tested 10,000 dice, we in fact expect to observe \(0.001 \times 10,000 = 10\) such extreme dice, even when all of them are balanced. The significant test results are so few that all of them could easily have occurred by chance alone. The more tests that we do, the harder it becomes to interpret the \(p\) value. As proteomics experiments can include tests of tens of thousands of peptides, we discuss the concept of multiple hypotheses correction in the next section.

3.1.2 False discovery rates

When making decisions about a large number of statistical tests, the \(p\) values become harder to interpret. One solution is to simply make the \(p\) value threshold more stringent, or to control the family-wise error rate, the probability of having at least one error among the tests above the threshold [14]. In bioinformatics it is common to report the the \(p\) value multiplied by the number of tests, the \(E\) value, to get the expected number of null tests that are called significant [15][17]. It could be more informative, however, to estimate the proportion of erroneous tests above the threshold. The false discovery rate (FDR) is formulated in terms of the tests that are called significant, and is more easily interpreted when the number of tests is large. We define the FDR as the expected proportion of true null tests, among all significant tests [18]. In terms of the distributions in Figure 3.1 the FDR can be understood as area C divided by the areas C+C+D. Area C is included twice as it represents both significant null tests from the left-side distribution, and significant alternative tests from the right-side distribution. For the dice test discussed above, the FDR translates to the proportion of truly balanced dice that we would expect by random chance among the dice above the threshold.

The FDR is assigned to and describes a set of tests, in contrast to the \(p\) value that refers to an individual test. This distinction is important, as it means that we can not tell the confidence of an individual test in a group, when we only know the FDR of the entire group. To assign an FDR to individual tests, the \(q\) value has been formulated as the minimum FDR required to call a test significant, considering all possible thresholds [19] [20]. This \(q\) value definition enables judging the confidence of each potentially unbalanced die individually, while avoiding the multiple hypotheses issues of the \(p\) value. Hence, a researcher can set a threshold test statistic corresponding to a \(q\) value of 0.01 to assure an FDR of 1%. At the same time, this \(q\) value indicates the confidence of the individual die, in the sense that we would have to allow an FDR of at least 1% to call it significant.
3.2 Machine learning

The term machine learning usually encompasses a set of methods used to find patterns in large amounts of data. It is common to divide them into supervised and unsupervised algorithms, depending on whether they learn from labeled training data or not. (Note that due to the rather redundant terminology the words training, fitting and learning all refer to same operation.) With supervised learning the parameters of a model are fitted to the training data, in which we already know the patterns. If this is successful, we can then use the fitted models to interpret new data points. The focus of the supervised methods is on classification and regression. That is, to categorize data points into two or more groups, or to look for relationships between data points and continuous variables, respectively. We describe the data points themselves by a number of variables, that we usually refer to as features [12; 21].

The papers included in this thesis involve supervised machine learning, and especially two common issues with the learning process itself. First, when the models fit to the training data, there is a risk of over-fitting. This occurs if the model is too flexible, so that it learns from random variations in the data. The risk can be reduced by increasing the number of training data points in relation to the number of features, for example. A standard method to avoid over-fitting is to keep some part of the training data away from the learner, and subsequently test the resulting model on this isolated test set. The idea is that if the model has over-fitted, we will detect it as it will perform poorly on the test set. If the training data is scarce, one can do this repeatedly by dividing the data into smaller sets. We then train the model on all but one of the sets, and test the model on the set we left out. By leaving different sets out, we can do this several times which allows us to estimate the average performance even when the amount of training data is limited. The procedure is called cross-validation, and treated in Paper III [12; 21].

The second machine learning issue I treat in this thesis is biased learning. With biased learning I refer to training data that systematically differs from the actual unlabeled data we want to interpret. In many machine learning scenarios this is not a problem, because there is plenty of relevant and high quality labeled data to learn from. Processes that run for a long time without changed conditions are typical examples. However, when machine learning is used to repeatedly adjust models to processes with ever changing conditions, obtaining training data becomes harder. In such cases one could learn from some type of models, or consider unsupervised learning. Standard test sets and cross-validation will not help to detect biases when learning from models, as the systematic errors will be present also in the test set. Paper I and II, however, deals with how biased learning can be detected in mass spectrometry-based
For all four articles in this thesis, I have used Percolator, a so-called post-processor based on a support vector machine (SVM)\cite{22,23} classifier used to categorize mass spectrometry-based proteomics identifications as either correct or incorrect. The SVM is a supervised algorithm that can be used to classify points into one of two groups. For training, it takes a set of example data points for which the group label is known. Using rather sophisticated mathematical methods, it then finds the high-dimensional plane that separates the two labeled groups with the largest possible margin, see Figure 3.2 for a two-dimensional example. We can then use this plane to categorize new, unknown data points.

**Figure 3.2: A support vector machine illustration.** In this example, each data point is described by two features, and belonging to one of two groups, the squares or the circles. The support vector machine assumes that the best separation between the two groups is the straight line with the largest possible margin, the two dashed lines.
4. Data collection in mass spectrometry-based proteomics

Here, I cover some instrumental aspects of mass spectrometry-based proteomics. I begin by introducing two central technologies, liquid chromatography and mass spectrometry, before describing how these are used to perform the technique shotgun proteomics.

4.1 Liquid chromatography

To physically separate molecules such as peptides that are dissolved in a liquid sample from one another, one can use liquid chromatography (LC) [24]. It has the form of a thin pipe (column) designed so that when you push through a liquid sample, molecules of the sample require different times to run through it. Hence, the molecules will come out, or elute, at different time points and can thus be separated. The time a molecule is in the column is called retention time. The time depends much on how well the molecule dissolves in water, especially with the reverse-phase liquid chromatography technique relevant for this thesis. A detailed explanation of liquid chromatography and the prediction of retention times that lie close to the work in this thesis can be found in a recent Ph.D. thesis [25].

4.2 Mass spectrometry

Mass spectrometry is a technique to measure the ratio between the mass and the charge of ions. Since its first use in the very end of the 19th century [26], many different types of mass spectrometers have been developed, specialized in different molecules and samples. They are usually described as composed of three different parts [27]. First, an ion source, or ionizer, is used to electrically charge the atoms or molecules into ions. Second, a mass analyzer separates the ions based on their mass-to-charge \((m/z)\) ratio. Third, a detector measures a signal induced by the separated ions.

There are several ionization techniques, but I will not cover them here. Here I will only mention the technique most relevant for this thesis, the elec-
trospray ionization (ESI) [28,29]. Proteomics researchers often use ESI as it is readily connected to a liquid chromatography system, and produces peptide ions with a range of different charges [30].

Depending on the type, mass analyzers utilize a variety of physical properties to separate ions of different $m/z$ ratios. Analyzers can for example measure the time accelerated ions take to travel a distance [31,32], the stability of flight trajectories in alternating electric fields [33], oscillation frequencies along an electrode [34] or the angular frequencies of ions circulating in a magnetic field [35]. In general, however, the forces that act on ions by electric and magnetic fields are the central physical principles behind the separation. An important instrument for this thesis is the Orbitrap mass analyzer in which ions rotate around, and oscillate along, a central electrode. The device is designed so that an axial oscillation frequency depends on the $m/z$ ratio of the ions that then can be detected by currents they induce [36]. This detection is that last step of the mass spectrometry analysis, where the detector utilizes that charged particles can induce or produce currents.

The signals from mass spectrometers are called mass spectra and consist of pairs of $m/z$ ratios and intensity values that relate to the detected current. The mass is counted in atomic mass units (u), or the equivalent daltons (Da). These are defined as a twelfth of the mass of the most common isotope of free carbon and are close to the mass of a single proton. Next, the charge is counted in units of a single elementary charge [37]. In mass spectrometry, we can count the $m/z$ ratios using Thomsons (Th), daltons per elementary charge [38]. The intensities are typically counted in an arbitrary scale, but are proportional to the number of detected ions. We typically display the $m/z$ ratios on the horizontal axis and the intensities as peaks on the vertical axis, see Figure 4.1 for an example. In the mass spectrometry data used in this thesis, the resolution of the $m/z$ separation is generally high enough to distinguish peaks of different isotopes from each other. We can see this in the mass spectra as a distribution of peaks for each ion, corresponding to its different isotopes [27].

![Figure 4.1: An example of a mass spectrum. The mass to charge ratio ($m/z$) is showed on the horizontal axis, and the observed intensity on the vertical axis.](image-url)
4.2.1 Tandem mass spectrometry

Tandem mass spectrometry (or MS/MS) is the name for running two sequential mass spectrometry analyses, called scans, to get additional information about a molecule. While the first scan measures the $m/z$ ratios of intact, full size ions, the second scan usually measures their fragments \[39, 40\]. With the mass of the fragments, one can more easily deduce the elementary composition and structure of a molecule than from the complete mass alone. This becomes even more useful as the machine often has the ability to select and isolate molecules within some $m/z$ interval, and to fragment and measure only the molecular species with those particular masses. In this thesis I will denote the first scan as the precursor scan, and the second scan as the fragmentation scan, which generates a fragmentation spectrum. For each fragmentation scan, the machine also reports the mass of the precursor. This precursor mass is valuable when identifying the fragmented molecule.

There are a couple of widespread techniques to fragment ions in MS/MS-based proteomics. Collision-induced dissociation (CID) is the primary technique in the data used for this thesis \[41\]. With CID, the ions are collided with neutral gas molecules, often nitrogen or argon gas, which break the ions into fragments. An alternative to CID, higher-energy collisional dissociation (HCD), uses a higher voltage to collide the ions with the gas \[42\]. A different approach is used with electron capture dissociation (ECD) \[43\] and electron transfer dissociation (ETD) \[44\] where a chemical reaction transfers an electron onto the ion to fragment it. Different methods are suitable for ions of different sizes and charge states, and could often complement each other \[45\].

4.3 Shotgun proteomics

Researchers can use LC and MS/MS in a variety of ways to study the proteome. During my Ph.D. I have looked at perhaps the most widely used method, namely shotgun proteomics \[46\]-\[48\]. However, most conclusions that I present in the included articles are applicable to any methods that use the fragmentation of peptides as a means for identifying proteins, for example data independent acquisition \[49\]-\[51\]. The name shotgun proteomics refers to the similarities with shotgun sequencing; the genomics technique that helped to sequence the first complete human genome \[52\]-\[53\]. Somewhat analogous to a shotgun, both techniques begin by splitting up the molecules of interest (proteins and DNA, respectively) into pieces that are more easily handled in a high throughput manner. Figure \[4.2\] shows a schematic overview of shotgun proteomics, which is described in more detail below.

In shotgun proteomics we cut the proteins of a sample into peptides using
Figure 4.2: An illustration of a shotgun proteomics experiment. Starting from a sample of proteins, we digest the proteins to peptides using a protease before an LC-MS/MS analysis that produces fragmentation spectra. In the computational part of the analysis, the fragmentation spectra are matched to peptides in a protein database. Ultimately, the peptide-spectrum matches are used to identify the peptides and the proteins present in the sample.

proteases, (see section 2.3.1). Although this digestion complicates some computational steps (section 5.3) it greatly facilitates the rest of the experimental procedures [10]. The different proteins can have very distinct properties, but the peptides dissolve more easily under the same conditions. Moreover, shorter sequences help the mass spectrometer to obtain the sequence information. After digesting the proteins into peptides, the researcher injects the sample into a liquid chromatography system that is directly coupled to a tandem mass spectrometer (a setup called LC-MS/MS). First, the liquid chromatography step separates the peptides and elutes them directly into an ESI ionizer of the mass spectrometer, often over about an hour. Due to the high number of peptides they are still generally not completely isolated from each other, hence a number of different peptides are inserted into the mass spectrometer at every given time point. Still, this step enables the mass spectrometer to make measurements over a considerable time, which increases the number of peptides that can be identified. One can also design additional separation steps that further fractionate the sample to allow even more time for mass spectrometry mea-
measurements [48][54][55]. At the time of writing, fast shotgun proteomics set ups can measure about 20 fragment scans per second [56].

As the peptides are injected into the tandem mass spectrometer over time, the machine repeatedly performs precursor and fragmentation scans. Although it definitely varies, a typical set up could run one scan in the precursor mode, followed by five scans in the fragmentation mode. The preceding precursor scan measures the masses and intensities of all peptides eluting from the liquid chromatography step and that were successfully ionized at a given time point. In the fragmentation scans, on the other hand, the machine attempts to fragment one peptide species at the time. For each scan, it selects and isolates a peptide species that corresponds to one of the five highest intensity precursor peaks. Hence, the five following fragmentation scans shows the fragments derived from the five most intense peptides of the precursor scan. To avoid fragmenting the same peptide species over and over again, however, the machine stores previously isolated masses in an exclusion list for some period of time during the analysis.

To understand and predict how the peptides fragment is important when interpreting the fragmentation spectra. Although each individual peptide molecule produces only one or a few detectable fragments, many peptides of the same sequence will generate a range of different fragments in the same fragmentation spectra. The fragments depend on the amino acid sequence and the fragmentation technique. CID and HCD, for example, often split peptides right at the peptide bond, while ETD splits them between atoms within the amino acids. As the fragmentation predominantly occurs along their backbones, the fragment masses represent subsets of the entire peptide sequence. See Figure 4.3 for an illustration of the fragments. As explained in the next chapter, this information can be used to calculate the peptide sequence. [10][27]
Figure 4.3: An example peptide and some of its fragments. Among other places, the example peptide AQVEMR can fragment at its peptide bonds, *i.e.* between its amino acids. Each such split can give two possible fragment ions. For example, fragmentation between amino acids V and E would generate fragments AQV with mass 316.17 Da or EMR with mass 416.18 Da. In the resulting fragmentation spectrum, we observe these fragments with different intensities, depending on their chances to be charged and detected.
5. Data analysis in mass spectrometry-based proteomics

Shotgun proteomics experiments generate thousands of fragment spectra, each corresponding to fragments of one or more peptides. Here, I introduce some computational means to process the spectra to produce lists of the peptides and proteins that were present in the sample. I also describe some methods to estimate confidence measures for the identifications. More comprehensive descriptions of the field can be found in one of the several good reviews [10, 57-60].

5.1 Identification of spectra

Given a fragmentation spectrum, we would like to figure out what peptide or peptides that generated it. Since we know the exact masses of all amino acids and the most common post-translational modifications, one approach could be to test combinations of amino acids and see whether their masses explain the fragment peaks or differences between them, such as in de novo sequencing [61-63]. One can also limit the de novo sequencing to short sequences, called peptide sequence tags, and use these tags to find entire peptides in a database [64-66]. Another approach is to match the spectra against previously identified spectra, in a spectrum library search [67-69]. In this thesis, however, I will only cover another method, the database search approach. At present, database searching is likely the most common method for identifying spectra in shotgun proteomics.

5.1.1 Database searching

DNA sequencing techniques advance rapidly, and genome sequences of a high number of organisms are currently available. To some extent, the protein coding parts of the genomes have been identified and can be translated into protein sequences. In human for instance, there are readily available databases that contain amino acid sequences of all currently known human proteins. Sequences can be downloaded from sources such as the European project Ensembl [70], the National Center for Biotechnology Information (NCBI) [71] in
the US or the collaborative Uniprot \cite{72} database of protein sequences. With these protein databases, algorithms called database search engines can match spectra to peptides to produce peptide-spectrum matches (PSMs) \cite{73}. Compared to de novo sequencing and sequence tag-based approaches, database searching is suitable for the large number of spectra in shotgun proteomics. Limiting the search space to database peptides has some disadvantages, but speeds up the search considerably and avoids statistical issues with large search spaces \cite{74}. Although spectrum library searches are also fast, they require extensive spectrum libraries and are not yet as widespread as database searches.

There is a large variety of different database search algorithms available, but below I try to generalize the steps they take to identify the spectra. To begin with, the search engine theoretically digests the protein sequences of the database into peptides according to the cleavage rules of the protease used in the experiment. It then uses the precursor mass of each spectrum as a clue to select plausible candidate peptides from the database. Typical candidate peptides have masses that lie close to the precursor mass of a given spectrum. For each of these candidate peptide sequences, the search engine generates a theoretical fragment spectrum based on its sequence. To find the best matching candidate peptide, the search engine then tests and evaluates the similarity between the experimental spectrum and the (many) theoretical spectra. See Figure 5.1 for an illustration of this.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{peptide_spectrum_matching.png}
\caption{An illustration of peptide-spectrum matching. Search engines, represented by the rectangle, typically take experimental spectra and a protein database as input. For each experimental spectrum, it identifies the peptide with the best matching theoretical spectrum. It then outputs the best match (or matches) for each experimental spectrum.}
\end{figure}

Naturally, most of the candidates were not actually present or fragmented
in the machine at the time the spectrum was generated. I will denote any matches to these peptides as incorrect PSMs, and matches to peptides that were fragmented as correct. The search engine often outputs only the PSM of the best matching peptide for each spectrum, however, there is still no guarantee that this top PSM is correct. In fact, a large proportion of the top PSMs are incorrect, in the next section we will discuss this in more detail. Nevertheless, to help ranking the PSMs according to how well they match, the search engine assigns scores to the PSMs. The scores are a measure of the similarity between the experimental and theoretical spectra, and a higher value usually means a better match. Note that the scores can be used twice. First, to identify the best match for each experimental spectrum. Second, to distinguish correct from incorrect PSMs among the top scoring PSMs of all experimental spectra. In the remainder of the thesis I will assume that each spectrum is assigned a single peptide, representing the best matching peptide in the database. I will refer to this match simply as the PSM of the spectrum.

The functions to calculate the scores are central parts of the search engines, and are often their main distinction. A good scoring scheme clearly discriminates between correct and incorrect PSMs, and thus helps to interpret the spectra more confidently. In this work I have used a number of different search engines and score functions. In paper I, II and IV, I used Crux [75], an open source implementation of the first published database search engine Sequest [73]. Crux’s and Sequest’s primary score is called XCorr, which requires summarizing each spectrum in a vector of intensities. The score is calculated by taking the cross-correlation between the experimental and theoretical vectors, minus the mean background cross-correlation [73]. We have also used the search engines X!Tandem [76] and MS-GF+ [45; 77], the latter of which is explained in more detail in Paper IV, where it plays a central role. In addition, there is a smorgasbord of other search engines that I did not use here, some popular examples are Mascot [78], OMSSA [79] and Andromeda [80].

5.1.2 Errors when identifying spectra

A significant proportion of the spectra, sometimes in the order of 50%, fails to be matched to the peptides that generated them. We denote these matches as incorrect PSMs, in contrast to correct PSMs. Considering database search engines, there are a number of possible explanations for why we see so many incorrect PSMs. One reason is that the peptide that generated a fragmentation spectrum might not be represented in the protein database. Either the database failed to include some expressed genes, splice variants or mutations [81; 82], or the peptide was modified in an unexpected way and became unrecognizable [65; 83]. A post-translational modification on an amino acid that was not
accounted for in the search engine would alter the mass of the peptide, so that a correct match may not be found.

Furthermore, some incorrect PSMs are likely explained by that the spectrum was generated by more than one peptide [84]. Such spectra will have a more complicated pattern of peaks, and depending on the matching algorithm of the search engine, only some or none of the peptides might be identified. Other errors could occur due to unexpected reactions during the peptide fragmentation, resulting in fragment peaks that are hard to interpret [85]. This problem is further confounded by peptides that share similar sequences [86]. In addition, one can think of errors stemming from low peptide concentrations or fragmentation of other molecules than peptides.

5.2 Confidence estimation methods

To help decide a threshold score to separate correct from incorrect PSMs, we use the confidence measures described in chapter 3. Here, we consider each PSM as a sample from the hypothetical population of all methods that can prove that the peptide was in the machine at the time of the fragmentation. Our null hypothesis is that there is no evidence that the peptide was there, we say that the PSM is incorrect, for short.

As the exact values of the confidence measures are unknown, below I introduce some common procedures to estimate them from the data. Many of the procedures come in the form of post-processors, a program run subsequent to a search engine to estimate, or to improve, the quality of the results. The post-processors are often useful as the original search engine scores can be quite general, and not optimized for the conditions of the experiment at hand. To facilitate their description, I have divided the estimation methods into parametric and empirical approaches.

5.2.1 Parametric methods

With parametric models, I refer to models that can be described by a fixed, limited number of parameters, regardless of the size of the data [12]. As a first example, the search engine X!Tandem uses a typical parametric method as it fits a hyper-geometric distribution to the scores of all but the top scoring PSM for each spectrum. From this distribution it estimates the $E$ value (described in section 3.1.2) of the top PSM [17; 87]. Another search engine, OMSSA, uses a Poisson distribution to describe the number of fragment peaks that are matched by chance when comparing a spectrum to a database to estimate the $E$ value [79].
PeptideProphet \cite{88} is a post-processor that looks at all top scoring PSMs. Each PSM is first described by a number of variables, or features, such as the PSM score and the mass accuracy. Using the expectation maximization algorithm \cite{89} it then trains two parametric models to fit the correct and the incorrect feature distributions, respectively. However, in contrast to the frequency-based confidence measures described previously, PeptideProphet outputs posterior probabilities of the models directly, in accordance to Bayes Law.

5.2.2 Empirical methods

The second group of confidence estimation methods uses an empirical distribution described by some sort of sample, to model the identifications. In this group, there is no fixed number of parameters to describe the model when the amount of data varies. The search engine MS-GF+ is one example, it uses a dynamic programming approach to generate the score distribution of all theoretical peptides matched against a given spectrum \cite{77}. From this score distribution one can estimate confidence measures related to the \( p \) value.

The perhaps most widely used method for estimating confidence measures in shotgun proteomics is the target-decoy analysis \cite{90, 91}. It assumes that scores of top scoring matches to a nonsense decoy database distribute similarly as those of top scoring incorrect matches to the original target database. The method requires only a decoy database, which we often create by literally shuffling or reversing the sequences of the organism under study, the target database. Ideally, the decoy database should be of the same size, amino acid composition and homology level as the target database, but not contain any sequences that may be found in the sample.

To estimate confidence measures, we search the spectra against both the target and the decoy database. Although there are two fundamentally different ways to do this, for this work I have most often searched the databases separately. The so called separate target-decoy analysis will produce one top target PSM and one top decoy PSM for each spectrum \cite{92}. The target PSM might be correct or incorrect, but the decoy PSM is assumed to always be incorrect. The other approach uses concatenated target and decoy databases and assigns only one PSM to each spectrum, the highest scoring target or decoy PSM \cite{91}. In both methods, the decoy score distribution acts as a null model from which error rates can be estimated. See Figure 5.2 A for an example of target and decoy score distributions.

To estimate the FDR of a certain score threshold using the target-decoy analysis, one can do as follows. Count the number of target and decoy PSMs with scores equal to or above the threshold, and then divide the decoy count by the target count. In Figure 5.2 B I demonstrate an example of how FDRs and
Figure 5.2: The target-decoy analysis. (A) Example score distributions from a target and a decoy search. (B) An example of how FDRs and \( q \) values can be estimated from their ranks, based on their scores. A score threshold to include the group of the four top ranking target PSMs, for example, have an estimated FDR of 1/4. The reason is that there are 4 target PSMs (circles) and 1 decoy PSM (square) including in the group. The \( q \) value of a PSM is the lowest FDR among all groups that includes it. In this example the FDRs are not adjusted with their prior probability of being incorrect, \( \pi_0 \).

\( q \) values can be estimated with this approach. However, the above calculation assumes that there are as many decoy PSMs as incorrect target PSMs. For a separate target-decoy analysis this assumption is often false, since hopefully some target PSMs are correct. To correct for this imbalance we multiply the FDR estimate with the prior probability of target PSMs to be incorrect, \( \pi_0 \). The value of \( \pi_0 \) can be estimated from the proportion of target and decoy PSMs among the low scoring PSMs \([19, 20, 92]\). We can also use the target-decoy analysis to easily estimate \( p \) values. For the same score threshold as above, just divide the decoy count by the total number of decoy PSMs. However, to assure the correct type 1 error rate for the \( p \) values, first add one to the numerator and the denominator \([93]\).

One benefit of the target-decoy analysis is that it is applicable to almost any type of PSM scores. As an example, the post-processor Percolator uses the score from its SVM classifier (see section \[3.2\]) to estimate the confidence of the PSMs \([94]\). In Percolator, each PSMs is described by a number of features that make up the dimensional space. As examples of incorrect PSMs, Percolator uses the decoy PSMs. Examples of correct PSMs are trickier, as we do not
know what target PSMs are correct. Thus Percolator initially makes a qualified
guess on a set of likely correct PSMs, and gradually improves this guess by
running multiple iterations of the SVM. Scores can then be assigned to PSMs
based on their location relative to the separating hyperplane. With the target-
decoy analysis, this score is then used to estimate confidence measures. Note,
however, that there are search engines that inherently break the fundamental
assumption of the analysis, that decoy PSMs are good models of incorrect
target PSMs [95, 96]. Paper I and Paper II deals with a method to validate that
the scores produce accurate statistics, for example when using the target-decoy
analysis.

5.3 Identification of peptides and proteins

So far in this chapter, I have discussed how we generate and estimate the confi-
dence of PSMs. The PSMs are rarely the final goal of proteomics experiments,
however, as the peptides and primarily the proteins are biologically relevant.
The PSMs are important as an intermediary step that the peptide and protein
identifications depend on. One should never confuse PSMs with unique pep-
tides, as these are two different entities. A single peptide sequence could be
matched by several PSMs, so the statistical significance of a match to a pep-
tide is different to the significance of the peptide itself, see Figure 5.3 for a
graphical illustration. Note that the definition of a unique peptide can vary de-
dpending on whether you regard a post-translational modification on a peptide
as a distinguishing factor, or not, for example.

Regardless of the definition, I will refer to unique peptides plainly as pep-
tides here. In some studies, the focus is on the peptides themselves [97], and
in many other cases the peptides, and not the PSMs, are used to infer the pro-
teins in the sample [98, 101]. A post-processor that is designed to identify
peptides is iProphet [102], which uses a similar algorithm as PeptideProphet
described in section 5.2.1. One can also score peptides by removing all but
the highest scoring PSMs matching each peptide, and then run a target-decoy
analysis on the remaining peptides. Paper II describes the issues of estimating
the confidence measures for peptides in more detail.

Inferring the proteins of the sample is computationally difficult [103]. As
illustrated in Figure 5.3, one reason is that the peptides are often shared be-
tween many proteins, due to similarities by random chance, homology be-
tween genes or due to alternative splicing of a single gene. In early shotgun
proteomics experiments, the PSMs that were identified were simply mapped
to their respective proteins so that these could be considered identified [47,
48, 104]. However, this procedure does not provide statistical confidence esti-
mates for the proteins, nor does it handle the varying confidence of the PSMs
Figure 5.3: Different levels of information in shotgun proteomics. Here, the information retrieved from shotgun proteomics experiments has been divided into four different levels: spectra, peptides, proteins and genes. In addition, one can imagine more levels, such as post-translationally modified peptides. The figure illustrates an example of how multiple spectra can match a single peptide, and how this peptide is shared between several proteins. In turn, the proteins can come from different genes, or the same gene due to alternative splicing.

themselves or the issues of peptides that are shared between multiple PSMs and proteins. To meet these (difficult) challenges, proteomics researchers developed more sophisticated models [98; 101]. However, to solve problems of shared peptides, many algorithms use heuristics such as reporting confidence estimates for groups of proteins, or reporting the most parsimonious set of proteins that explains the peptides [99;105;106]. Unfortunately, this has been showed to not be very reproducible, as small variations among the peptides could generate completely different protein sets [107].
6. Present investigation

The publications I have included in the thesis are centered around the topic of quality control for shotgun proteomics. In a sort of meta-study, the focus is on the quality of the quality estimates reported for the results. To give an example, researchers often use the target-decoy analysis to estimate the statistical significance of PSMs or peptides, but how do we check that the estimates are accurate?

6.1 Paper I overview

Paper I deals with how we can use purified samples that contain only known proteins to evaluate the statistical estimates that are assigned to the PSMs. We say that we test the statistical calibration of a confidence estimation method. The idea of using purified samples is not new, and intuitively one could easily evaluate any statistical method with such a sample, since we know what proteins are present [108; 109]. Many researchers have evaluated statistical estimates with PSMs that they assumed were either correct or incorrect [88; 95; 110–114]. However, we start off Paper I by showing that even with these samples it is hard to tell what PSMs that are correct.

As a solution to the problem, we continue by proposing a new method to utilize these characterized samples. We call the method the semi-labeled calibration test, because we propose only labeling the PSMs that we know are incorrect. In an asymmetric fashion, identifying incorrect PSMs from these samples turns out to be easier than identifying correct PSMs. In the next step, we hence let a statistical method estimate the confidence of the incorrect PSMs, and express the confidence using \( p \) values. Recall from section 3.1.1 that \( p \) values of incorrect (true null) PSMs should distribute uniformly between 0 and 1, according to their definition. We can thus make a sanity check of the statistical method, by looking how uniformly these null \( p \) values distribute.

Our semi-labeled calibration test is not limited to \( p \) values, but can be formulated for any statistical metric that has an expected distribution under the null hypothesis. However, some issues still remain with our approach. These purified samples often contain very few proteins, and are very different from complex real world samples. Compared to real-world samples, a known
mixture does not over-load the machine with tens of thousands of different peptides, so we can expect the spectra to be slightly different. Nevertheless, the calibration test serves as a minimum requirement for any statistical procedure, and should be highly valuable for developers of computational methods in shotgun proteomics.

6.2 Paper II overview

In Paper I we demonstrated a test to evaluate the calibration of statistical confidence measures assigned to PSMs. In Paper II, we extend this test to deal with unique peptides. Like others have done before [100; 102], an important part of the paper is to highlight the difference between PSMs and peptides. There may be many redundant PSMs mapping to the same peptide, and as a consequence the statistical confidence of a PSM is not the same as the confidence of the peptide it maps to.

There are many potential ways of estimating statistical confidence measures for peptides. Ideally, we would like to combine the statistical estimates of all PSMs mapping to the same peptide into a single value. If the PSMs were statistically independent, one could for example use Fisher’s method to combine p values [115] or the product rule to combine probabilities [12]. However, multiple PSMs mapping to a single peptide can not be considered as independent evidence for the peptide [98]. As a consequence, researchers have resorted to choosing either the best score or the best confidence measure assigned to the redundant PSMs to represent their common peptide. Although many statistical tools use this method, its statistical calibration had not been tested. Furthermore, it has been unclear whether to let the best score, or the best confidence estimate, represent the unique peptide. We demonstrate that these two alternatives produce different results, and argue for choosing the best score and running the error rate estimation procedure again for the unique peptides.

6.3 Paper III overview

Paper III describes the method we employ in Percolator to avoid poor calibration. The method is a cross-validation procedure [12, 116], a standard method in machine learning to avoid problems like over-fitting, discussed in section 3.2. As Percolator trains a model with often more than 20 parameters to discriminate correct from incorrect PSMs, there is a risk of over-fitting. So, in a typical cross-validation procedure Percolator first separates the PSMs into three groups. Percolator then fits three models, each trained on two of
the groups and applied on the remaining third. Thus, we classify all PSMs with a model trained on a separate dataset. A proteomics researcher naturally wants statistics reported for all three groups together, so we proceed by normalizing the score from each model, and perform a target-decoy analysis on the normalized scores.

Due to the difficulties with known protein samples, we use synthetic data to show that the cross-validation scheme is necessary for producing well calibrated confidence estimates. This type of internal cross-validation procedure, combined with score normalization and the target-decoy analysis is applicable to virtually any machine learning post-processor of shotgun proteomics data.

6.4 Paper IV overview

With this work, we describe a new database searching pipeline for analyzing proteomics data to match spectra and assign significance to PSMs, peptides and proteins. It is based on the database search engine MS-GF+, which uses a dynamic programming approach to obtain scores from all possible peptides to a given spectrum [45][77]. This exhaustive approach enables highly informative scores with a strong discriminative power. In this study, we made a tool to convert the output from MS-GF+ to the input format of Percolator, which post-processes the PSMs to optimize the discrimination. Percolator also directly reports error rates for PSMs, peptides and proteins, which greatly facilitates the interpretation of the MS-GF+ output.

The resulting tool, denoted MS-GF+Percolator, provides a typical example of when to use the calibration test described in the Paper I and II of this thesis. Here, Percolator used more than 20 new and old features to describe the PSMs. With the semi-labeled calibration test we show that the statistical estimates from MS-GF+Percolator are well calibrated.
7. Future perspectives

In this thesis I have presented work to promote the quality of shotgun proteomics results. I believe the findings can play a constructive part among the plethora of instruments and algorithms we currently employ to do mass spectrometry-based proteomics. The work is not strictly limited to shotgun proteomics, but can make a difference also for alternative workflows. Anyhow, enough about my own research, I have summarized some of my thoughts about the future of the field below.

To me, the most critical challenge that many mass spectrometry-based proteomics workflows face today is the protein inference problem. We are often more or less forced to digest the proteins into peptides to enable high-throughput experiments, but computationally deducing the proteins from the peptides is not easy. First of all, we often identify only a couple of the perhaps 50 tryptic peptides of a protein. Second, these peptides can be shared between a number of proteins. Instrument developments that allow for longer peptides (or no digestion at all) and higher sequence coverage would definitely facilitate these inferences.

On the computational side of the protein inference problem, I think our current representation of the proteins is a weakness. I have predominantly used the Ensembl protein databases, where each known splice variant is represented as a unique protein. The sequences of many of these splice variants are highly similar and can differ by a single peptide. A different approach is taken by the Uniprot databases, as they attempt to report a single canonical sequence for each gene. I think none of the approaches are satisfactory. With the current instrumentation it may be almost impossible to tell the splice variants apart. Our protein inference algorithms might assign low (or high) probabilities to all of them. In contrast, with a single sequence for each gene we may miss out on biologically relevant splice variants. Instead of representing proteins as discrete entities, I think a model of the splice variants as some kind of probabilistic path over different exons in the gene could be beneficial.

After the protein inference problem, an interesting computational issue to address is that of the protein quantification. Most current inference algorithms assume that proteins are either present or absent, while the quantities of the present proteins are estimated more or less separately. It could perhaps be useful to model only the quantities of the proteins, so that an absent protein
has a quantity of zero. Imagine a Bayesian posterior distribution over different protein quantities. As we find evidence for a protein, the posterior distribution may shift towards higher protein quantities. The prior probability of different quantities could be estimated from previous studies.
Flera av de mest lovande molekylärobio logiska teknikerna analyserar prover i stor skala, och kan potentiellt revolutionera sjukvården och öppna för nya biologiska användningsområden. Den ledande metoden för storskaliga proteinstudier är masspektrometri. I dagsläget kan man rutinmässigt identifiera och kvantifiera tusental proteiner i ett enda experiment, med hjälp av tekniken shotgun proteomics.

En utmaning hos de här experimenten är beräkningarna, och hur masspektrometern ska tolkas. Ett shotgun proteomics-experiment kan enkelt generera tio tusental spektra, som alla kan representera en peptid från ett protein. Dock gör både biologiska och tekniska svårigheter att beräkningsmetoderna vi använder ofta tolka spektra felaktigt, så att fel peptid identifieras. Detta leder till att vi måste förlita oss fullständigt på statistiska uppskattningar om resultatens kvalité för att kunna dra biologiska slutsatser.

Den här avhandlingen innehåller fyra artiklar från min forskning om riktheten hos de statistiska säkerhetsuppskattningar som beräknas för shotgun proteomics-resultat, och hur denna riktighet kan mätas. I de första två artiklarna presenterar vi en ny metod för att använda redan karaktäriserade proteinpov för att mäta riktheten. Den tredje artikeln visar hur man kan undvika feluppskattningar när man använder maskininlärning för att analysera datat. I den fjärde artikeln presenterar vi ett nytt verktyg för att analysera shotgun proteomics-resultat, och mäter riktheten i de statistiska uppskattningar det rapporterar.

Resultaten som presenteras i den här avhandlingen kan förenkla utvecklingen av nya och exakta beräkningsverktyg inom masspektrometri-baserad proteomik. Sådana verktyg gör tolkningen av spektra och de efterföljande biologiska slutsatserna mer tillförlitliga.
I must thank my supervisor Lukas Käll for considering me and choosing me over the (presumably) hard competition for the Ph.D. position in his group. I am very grateful for all the ideas, knowledge, honest feedback and general encouragement that I got which made this project so much smoother. Many thanks to my fellow Ph.D. student Luminita Moruz for all the guidance and the great company during work, conferences and breaks. And thanks to Arne Elofsson for being my optimistic co-supervisor.

Erik Sjölund, thank you for teaching me what I should have already known about computers. I am still recovering from your win on 5 km though. I am also grateful to José Fernández Navarro for all your skilled programming advice, and for all the hilarious stories. I am very thankful to my knowledgeable co-authors abroad, William Stafford Noble, Sangtae Kim and Richard Smith, for great ideas and collaborations. And I owe many thanks to the other members of the research group, like Mattia Tomasoni, Magnus Rosenlund, Amin Saffari, Xiao Liang, David Menéndez Hurtado and Matthew The, among others, for help with the projects and the thesis. To Jorrit Boekel, for the quick and fun perspective on science and many other things, and to the other members of Janne Lehtio’s group that I have had the luck to work with.

Almost needless to say, I thank the many friends and colleagues at the Arrenius Laboratory for giving me such an inspiring start of my Ph.D.! You are too many to mention here, but I am especially grateful to Johannes Björnerås, although your research risks overturning the foundations of our society. And to Axel Abelein for your kindness when Munich trashed Barcelona. To my Ph.D. twin Ye Weihua for your contagious enthusiasm, Sofia Unnerståle for your playfulness and to Patrik Björklom for all the witty and true advice. Thanks to members of Arne Elofsson’s group that I’ve been discussing, teaching and planning with during these years. And thanks to Torbjörn Astlind and Haidi Astlind for the help with everything from appliances to administration, like when I needed a saw(!) at work.

The last years of my Ph.D. I spent at the Science for Life Laboratory surrounded by an army of talented, friendly and helpful people. Special thanks to Joel Sjöstrand for never rejecting any new ideas or perspectives, regardless of their quality. To Hossein Farahani for the passion and the perfected cappuccinos, you taught me the scientific importance of Beethoven. To Mattias
Frånberg and Kristoffer Sahlin for keeping it freq, and keeping me in competitive shape! Many thanks to Ikram Ullah, Muhammad Owais Mahmudi, Pekka Parviainen, Mehmood Khan, Auwn Muhammad Sayyed and Raja Hashim Ali for the riddles, discussions, cookies, cakes and the general education about the world. To the brilliant Rezin Dilshad of course, for countless of reasons. And to all other friends and colleagues at Scilifelab who really made me enjoy being there.

I must also thank Jonas Bengtsson, Tomas Taus, Jürgen Klevert, Edward Young and Leslei Nogueira for your great company and help during summer schools and conferences. And Johan Seijsing and Shahin Aeinehband for rewarding discussions and football games. At last, many greetings and thanks to friends and family!
References


